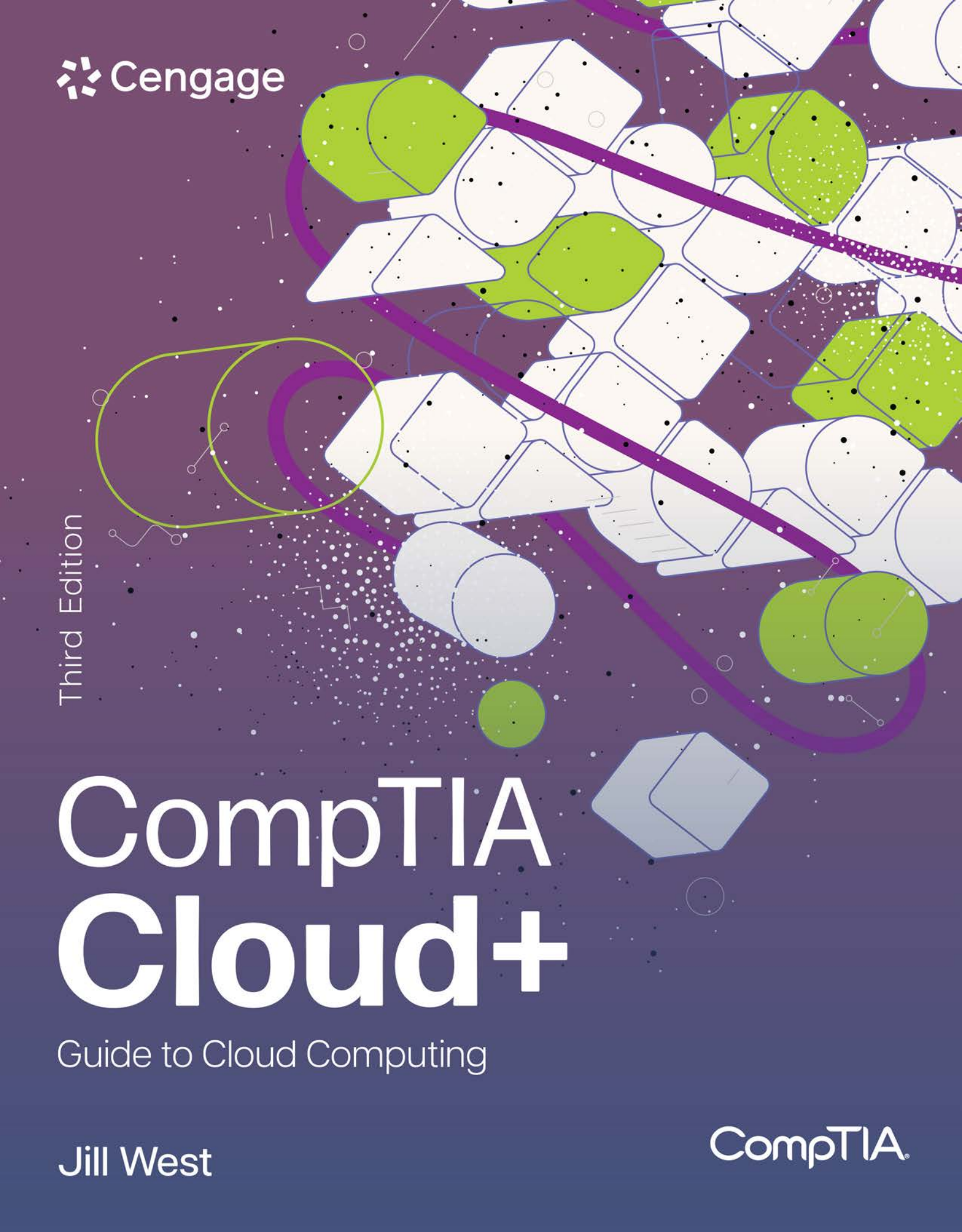
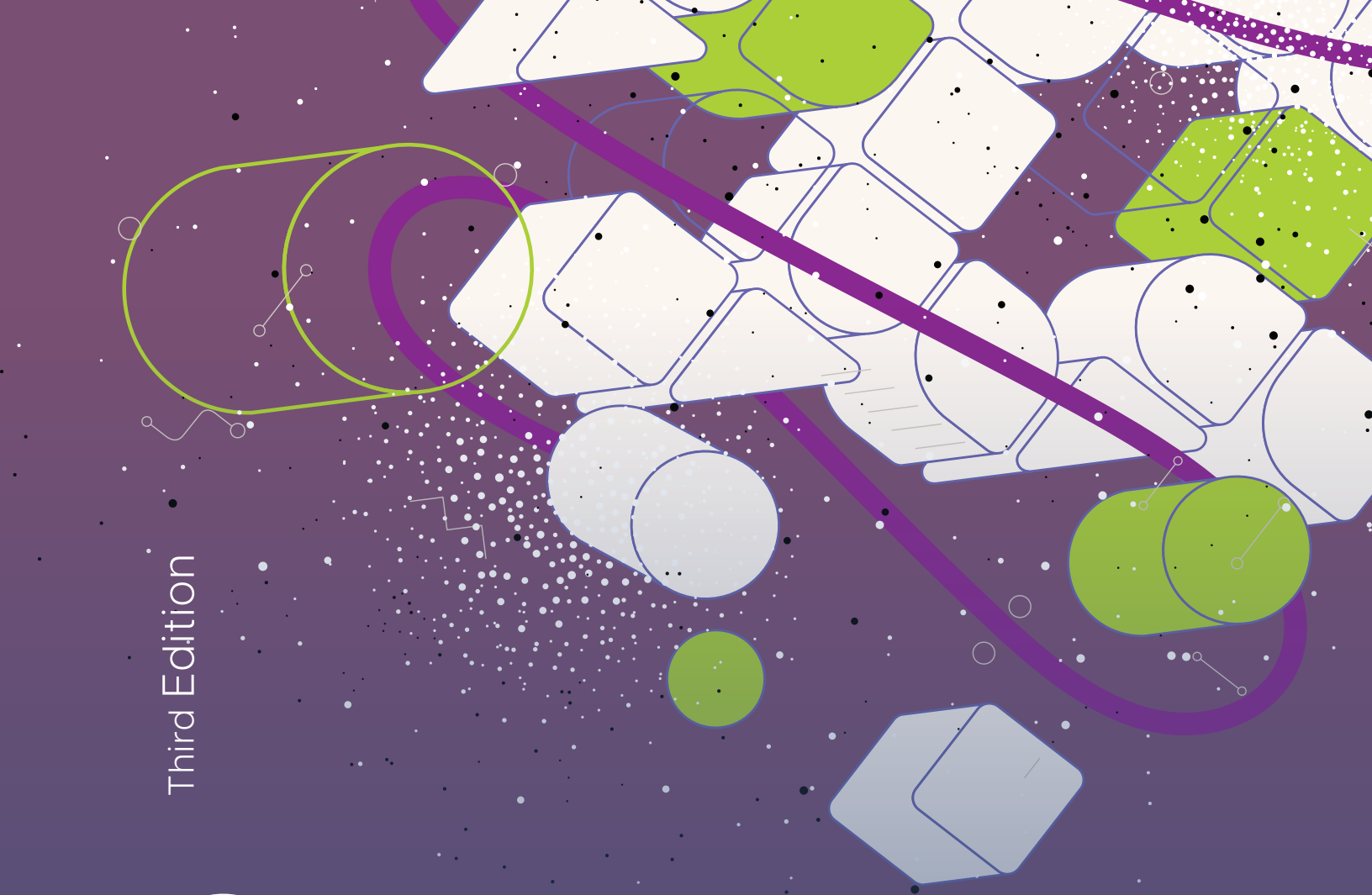

Cengage
Third Edition
CompTIA
Cloud+
Guide to Cloud Computing
Jill West
CompTIA



Third Edition

CompTIA Cloud+

Guide to Cloud Computing

Jill West

CompTIA®

Cengage

Australia • Brazil • Canada • Mexico • Singapore • United Kingdom • United States

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

Important Notice: Media content referenced within the product description or the product text may not be available in the eBook version.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CompTIA Cloud+ Guide to Cloud Computing
Third Edition
Jill West

EVP, General Manager, Cengage Academic: Nhaim Khoury

VP, Portfolio Management: Mark Santee

Portfolio Director: Rita Lombard

Portfolio Manager: Natalie Onderdonk, Kimberly Klotz

Portfolio Assistant: Anh Nguyen

Learning Designer: Zenya Molnar

Content Project Manager: Sivakumar Krishnamoorthy

Digital Project Manager: Jim Vaughey

Technical Editor: Danielle Shaw

Developmental Editor: Lisa Ruffolo

Senior Manager, Product Marketing: Neena Bali

Product Marketing Manager: Cybèle Beckham

Content Acquisition Analyst: Christy Frame

Production Service: Straive

Art Director: Erin Griffin

Cover Image Source: Sanjatosic/Shutterstock.com

Interior image Source: Sanjatosic/Shutterstock.com

Copyright © 2027 Cengage Learning, Inc. and its affiliates ALL RIGHTS RESERVED.

WCN: 02-300-526

No part of this work covered by the copyright herein may be reproduced or distributed in any form or by any means, except as permitted by U.S. copyright law, without the prior written permission of the copyright owner.

Unless otherwise noted, all content is Copyright © Cengage Learning, Inc.

Previous edition(s): © 2021, © 2023

For product information and technology assistance, contact us at
Cengage Customer & Sales Support, 1-800-354-9706
or support.cengage.com.

For permission to use material from this text or product, submit all requests online at **www.copyright.com**.

Library of Congress Control Number: 202594413
ISBN: 979-8-214-40982-5

Cengage
10650 Toebben Drive, Suite A
Independence, KY 41051
USA

Cengage is a leading provider of customized learning solutions. Our employees reside in nearly 40 different countries and serve digital learners in 165 countries around the world. Find your local representative at **www.cengage.com.**

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com.**

Notice to the Reader

Publisher does not warrant or guarantee any of the products described herein or perform any independent analysis in connection with any of the product information contained herein. Publisher does not assume, and expressly disclaims, any obligation to obtain and include information other than that provided to it by the manufacturer. The reader is expressly warned to consider and adopt all safety precautions that might be indicated by the activities described herein and to avoid all potential hazards. By following the instructions contained herein, the reader willingly assumes all risks in connection with such instructions. The publisher makes no representations or warranties of any kind, including but not limited to, the warranties of fitness for particular purpose or merchantability, nor are any such representations implied with respect to the material set forth herein, and the publisher takes no responsibility with respect to such material. The publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or part, from the readers' use of, or reliance upon, this material.

Printed in the United States of America
Print Number: 01 Print Year: 2026

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Brief Contents

Preface ix

Chapter 1
Introduction to Cloud Computing 1

Chapter 2
Cloud Compute 44

Chapter 3
Cloud System Design 83

Chapter 4
Cloud Networking 127

Chapter 5
Cloud Connectivity 169

Chapter 6
Securing Cloud Resources 222

Chapter 7
Identity and Access Management 276

Chapter 8
Cloud Storage and Databases 320

Chapter 9
Observability and Monitoring 372

Chapter 10
Cloud Automation 416

Appendix A
CompTIA Cloud+ Exam Objectives Mapped to Chapters 456

Appendix B
Chapters Mapped to CompTIA Cloud+ Exam Objectives 477

Appendix C
CompTIA Cloud+ Acronyms 500

Appendix D
Rubric for Hands-On Projects and Challenge Projects 504

Appendix E
Calculating Subnets 505

Glossary **516**
Index **526**

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Contents

Preface ix

Chapter 1

Introduction to Cloud Computing 1

Section 1-1: Characteristics of Cloud Computing 2
- CompTIA Cloud+ Certification 3
- Other Cloud Certifications 4
- What Is Cloud Computing? 5
- What Do I Need to Know? 7

Section 1-2: Cloud Deployment Models 9
- Public Cloud 9
- Private Cloud 10
- Hybrid Cloud 11
- Multi-Cloud 12
- Community Cloud 13

Section 1-3: Cloud Service Models 13
- Common Cloud Service Models 14
- Managed Services 16

Section 1-4: Cloud Services 17
- PaaS and IaaS Providers 17
- Common Cloud Services 18
- AI/ML (Artificial Intelligence/ Machine Learning) Services 21
- IoT (Internet of Things) 25

Section 1-5: Troubleshooting Methodology 26
- Common Cloud Computing Problems 26
- Troubleshooting Steps 27
- Preventive Measures 28

Summary 29

Key Terms 30

Acronyms Checklist 30

Review Questions 30

Scenario-Based Questions 31

Hands-On Projects 32

Challenge Projects 40

Chapter 2

Cloud Compute 44

Section 2-1: Virtualization Technologies 46
- VMs (Virtual Machines) 46
- VM Configuration 48
- Network Connection Types 51

Section 2-2: VMs in the Cloud 53
- VM Instance Types 54
- Cloning VM Instances 54
- Billing Models 56
- Resource Allocation Factors 59
- Affinity 60

Section 2-3: VM Alternatives 63
- Serverless Computing 63
- Containers 65

Summary 71

Key Terms 72

Acronyms Checklist 72

Review Questions 72

Scenario-Based Questions 73

Hands-On Projects 74

Challenge Projects 78

Chapter 3

Cloud System Design 83

Section 3-1: Cloud Migration 84
- Migration Types 85
- Migration Strategies 86
- Budgeting for Cloud 87
- Regulations and Compliance 89
- Project Management 92

Section 3-2: Cloud-Native Design 94
- Building with Microservices 94
- System Integration 96

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 3-3: System Development 98
DevOps 98
Version Control 99
Deployment Strategies 103
Testing Types 103
Section 3-4: Planning for Problems 105
Common Deployment Issues 105
Sizing Issues 107
Capacity Planning 108
Auto Scaling Types 109
Business Continuity Planning 110
Disaster Recovery 112
Summary 115
Key Terms 116
Acronyms Checklist 116
Review Questions 117
Scenario-Based Questions 118
Hands-On Projects 119
Challenge Projects 122

Chapter 4
Cloud Networking 127
Section 4-1: Networking Concepts in the Cloud 129
Networking Concepts 129
From OSI Model to Cloud Stack 130
Software-Defined Networking in the Cloud 133
Section 4-2: IP Address Spaces 135
IP Addressing 136
Subnetting 137
Cloud Network Interfaces 140
Section 4-3: Networking in AWS 142
Regions in AWS 143
Availability Zones in AWS 144
VPCs in AWS 145
Subnets in AWS 147
Gateways and Route Tables in AWS 148
Section 4-4: Networking in Azure 151
Regions and Availability Zones in Azure 152
VNets and Subnets in Azure 152
Route Tables in Azure 154
Section 4-5: Networking in GCP 155
Regions and Zones in GCP 155
VPCs and Subnets in GCP 156
Routes in GCP 157
Summary 158
Key Terms 158
Acronyms Checklist 159
Review Questions 159
Scenario-Based Questions 160
Hands-On Projects 160
Challenge Projects 165

Chapter 5
Cloud Connectivity 169
Section 5-1: Hybrid Cloud and Multi-Cloud Networking 171
Connecting Networks 172
Connecting Cloud Networks 177
Section 5-2: Extending Network Services 180
DHCP Services 180
DNS Services 184
Routing 185
Load Balancing 186
Section 5-3: Extending the Cloud 191
The Case for Edge Computing 191
CDN (Content Delivery Networks) 193
IoT Cloud Networking 193
Section 5-4: Troubleshooting Cloud Connectivity 197
Common Connectivity Problems 197
HTTP Status Codes 200
Misconfigured Routes 201
NTP (Network Time Protocol) 202
VLAN Issues 202
Summary 204
Key Terms 204
Acronyms Checklist 205

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Review Questions 205
Scenario-Based Questions 206
Hands-On Projects 207
Challenge Projects 217

Chapter 6

Securing Cloud Resources 222

Section 6-1: Cloud Security Configurations 224
- Threats to Cloud Security 224
- Vulnerability Management 229
- Shared Responsibility for Cloud Security 231

Section 6-2: Virtual Network Security 234
- Allow and Deny Rules 235
- Zero Trust Security 236
- AWS VPCs and Subnets 238
- Azure VNets 243
- GCP VPCs 246
- Securing Hybrid and Multi-Clouds 248

Section 6-3: Compute Security 250
- Endpoint Protection 250
- Application Security 252
- Pipeline Security 252

Section 6-4: Data Security 253
- Encrypting Data 254
- Key Management 257

Section 6-5: Troubleshooting Cloud Security 260
- Security Misconfigurations 260
- Key Management Issues 261

Summary 262
Key Terms 263
Acronyms Checklist 264
Review Questions 264
Scenario-Based Questions 265
Hands-On Projects 266
Challenge Projects 272

Chapter 7

Identity and Access Management 276

Section 7-1: Cloud Accounts 278
- Identity 278
- Account Types 279
- Account Management 283

Section 7-2: Authentication 284
- Authentication Processes 284
- Password Policies 286
- MFA (Multifactor Authentication) 288
- Certificate-Based Authentication 290
- Programmatic Access 292
- Securing Secrets 294

Section 7-3: Authorization 295
- Authorization Processes 295
- AWS IAM 297
- Azure IAM 299
- GCP IAM 300

Section 7-4: IAM for Hybrid Clouds 302
- SSO (Single Sign-On) 302
- AWS IAM in a Hybrid Cloud 304
- Azure IAM in a Hybrid Cloud 305
- GCP IAM in a Hybrid Cloud 306

Section 7-5: Troubleshooting Cloud IAM 307
- Common IAM Issues 307
- Troubleshooting Privilege Issues 308

Summary 309
Key Terms 310
Acronyms Checklist 310
Review Questions 311
Scenario-Based Questions 312
Hands-On Projects 312
Challenge Projects 316

Chapter 8

Cloud Storage and Databases 320

Section 8-1: Storage Types 322
- Data Types 322

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

On-Prem Storage Technologies 324
Cloud Storage Technologies 326
Types of Databases 330
Section 8-2: Storage Configuration 334
Tiering 334
Storage Performance 335
Storage Migration 336
Data Destruction 337
Section 8-3: Cloud Storage and Database Services 338
AWS Storage Services 338
Azure Storage Services 341
GCP Storage Services 343
Section 8-4: Creating and Storing Backups 345
Backup Strategies 346
Backup Retention 347
Backup Types 348
Backup Considerations 350
Section 8-5: Storage Security 351
Storage Security Techniques 351
Data Classification 353
Data Obfuscation 353
Regulations for Storage Security 354
Summary 356
Key Terms 357
Acronyms Checklist 357
Review Questions 357
Scenario-Based Questions 358
Hands-On Projects 359
Challenge Projects 366

Chapter 9

Observability and Monitoring 372

Section 9-1: Monitoring Resources 374
Targets to Monitor 374
Monitoring Tools 378
Dashboards 380
Section 9-2: Events, Logs, and Traces 383
Events 383
Logs 384
Traces 388
Section 9-3: Cloud Monitoring Services 390
Monitoring in AWS 390
Monitoring in Azure 395
Monitoring in GCP 396
Section 9-4: Troubleshooting Problems and Performance 399
Quotas and Throttling 399
Utilization Issues 400
Scaling and Sizing Issues 401
Restoration from Backup 402
Summary 403
Key Terms 404
Acronyms Checklist 404
Review Questions 404
Scenario-Based Questions 405
Hands-On Projects 405
Challenge Projects 411

Chapter 10

Cloud Automation 416

Section 10-1: Automation 418
Why Automate? 419
IaC (Infrastructure as Code) 419
CaC (Configuration as Code) 423
Comparing Tools 425
Cloud CLIs 425
Section 10-2: Automating Code Pipelines 430
CI/CD Pipelines 430
Scripting Logic 433
Section 10-3: Cloud Maintenance 435
Types of Updates 436
Patch and Update Methodologies 437
Decommissioning 440
Section 10-4: Troubleshooting Automation Issues 441
Common Automation Issues 441
Breakdowns in Workflow 442

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary 444
Key Terms 444
Acronyms Checklist 445
Review Questions 445
Scenario-Based Questions 446
Hands-On Projects 447
Challenge Projects 452

Appendix A
CompTIA Cloud+ Exam Objectives Mapped to Chapters 456

Appendix B
Chapters Mapped to CompTIA Cloud+ Exam Objectives 477

Appendix C
CompTIA Cloud+ Acronyms 500

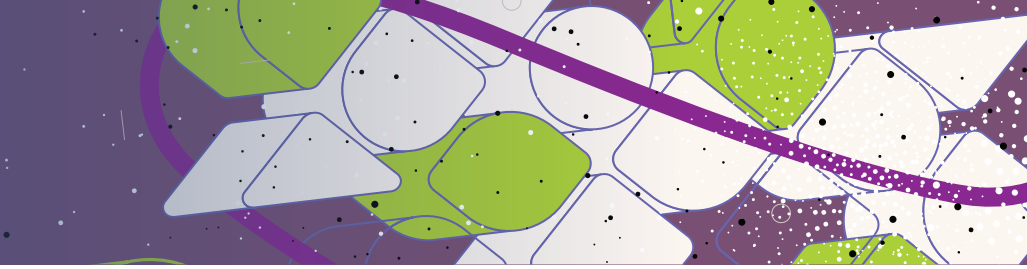
Appendix D
Rubric for Hands-On Projects and Challenge Projects 504

Appendix E
Calculating Subnets 505

Glossary **516**
Index **526**

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Preface



CompTIA Cloud+ Guide to Cloud Computing, third edition, serves students and professionals who are interested in mastering fundamental, vendor-independent cloud computing concepts. No previous cloud computing experience is necessary to begin learning from this course, although knowledge of basic computer, networking, and security principles is helpful. Those seeking to pass CompTIA's Cloud+ certification exam will find the course's content, approach, and numerous projects and study questions especially helpful. For more information on CompTIA Cloud+ certification, visit CompTIA's website at comptia.org.

Chapter Descriptions

The following list summarizes the topics covered in each chapter of this course:

Chapter 1, "Introduction to Cloud Computing," gives an initial overview of foundational cloud computing concepts, beginning with a survey of cloud certifications, a review of characteristics that define cloud computing, and a description of how an IT professional must adapt existing skills to succeed in a career focused on cloud technologies. The chapter then compares cloud deployment models (such as public cloud and hybrid cloud), cloud service models (such as SaaS and IaaS), and popular cloud services, including AI (artificial intelligence), ML (machine learning), and IoT (Internet of Things) service types. Finally, the chapter introduces popular cloud platforms, such as AWS (Amazon Web Services), Microsoft Azure, and GCP (Google Cloud Platform), and reviews sound troubleshooting methodology.

Chapter 2, "Cloud Compute," begins with a description of virtualization technologies, including a deep dive into VM (virtual machine) configuration parameters, especially virtualized processing and virtual memory. The chapter then applies these concepts specifically to the context of cloud computing, followed by a survey of VM allocation factors. The chapter concludes with coverage of serverless computing and containers, both popular VM-alternative technologies. The expanded content on containers discusses container configuration, support, networking, and troubleshooting.

Chapter 3, "Cloud System Design," focuses on how to design cloud-native systems that best take advantage of what the cloud has to offer. The chapter first compares migration strategies (such as lift-and-shift and lift-tinker-and-shift) where an existing system might need to be moved from an on-prem environment to the cloud. The chapter proceeds to discuss event-driven architecture powered by microservices and API (application programming interface) calls. It also covers business processes that serve application development, such as DevOps, version control, and testing types. The chapter concludes with a synopsis of common problems related to deployment, capacity, and auto scaling followed by a discussion of how to plan for business continuity and disaster recovery.

Chapter 4, "Cloud Networking," initiates the discussion of cloud-based networking concepts with a comparison of the OSI Model and the cloud stack followed by an examination of SDN (software-defined networking). The chapter includes a review of IP addressing and subnetting concepts and an exploration of networking services in AWS, Azure, and GCP. These sections also discuss cloud networking concepts such as regions, availability zones, VPCs (virtual private clouds), VNets (virtual networks), subnetting in the cloud, and the use of gateways and route tables to manage cloud traffic.

Chapter 5, "Cloud Connectivity," continues the cloud networking discussion with an exploration of technologies that connect the on-prem network with cloud-based resources. The chapter explores technologies available to extend networking services across a hybrid or multi-cloud, such as DHCP, DNS, routing, and load balancing. New content explores how the cloud can be extended using edge technologies, such as CDNs (content delivery networks) and IoT. The chapter concludes with an overview of common connectivity problems and how to troubleshoot these issues.

Chapter 6, "Securing Cloud Resources," introduces security-related threats specific to cloud computing. While security is addressed throughout most of the chapters, this chapter brings additional focus to cloud-based security strategies. The chapter highlights concerns and techniques specific to virtual network security, compute security (including containers and development pipelines), and data security, followed by an overview of common security vulnerabilities in the cloud.

Chapter 7, "Identity and Access Management," furthers the security discussion with thorough coverage of

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

IAM (identity and access management) techniques used to control access to cloud resources. The chapter covers account types that offer identity services to human users and to application or cloud services, along with a comparison of consumer accounts and workforce identities. A thorough discussion of authentication technologies and tools used in a cloud environment includes topics such as password policies, MFA (multifactor authentication), programmatic access, and secrets management. Continuing with the theme of the three-tiered AAA (authentication, authorization, and accounting) approach to network access control, the chapter then explores options for managing authorization and permissions in AWS, Azure, and GCP. The chapter concludes with a brief discussion of how to extend IAM across a hybrid cloud as well as common IAM troubleshooting issues.

Chapter 8, "Cloud Storage and Databases," explains common storage and database technologies both on-prem and in the cloud. The chapter then discusses storage configuration and optimization techniques, such as tiering, followed by highlights of popular storage services in AWS, Azure, and GCP. It continues with an exploration of common backup types and techniques and finishes with an emphasis on securing cloud-hosted data storage.

Chapter 9, "Observability and Monitoring," illustrates the need for effective visibility and monitoring techniques in the cloud along with an introduction to the benefits realized when automating cloud management tasks. Following a comparison of data collection tools—such as events, logs, and traces—the chapter continues with coverage of analysis and response tools available in AWS, Azure, and GCP. Finally, the chapter discusses common performance and capacity issues.

Chapter 10, "Cloud Automation," rounds out the foundations of cloud computing with further exploration into the possibilities and sheer necessity of using automation techniques in the cloud. Due to the fast-paced changes constantly occurring in cloud configurations and customer demand, automation through IaC (infrastructure as code) and CaC (configuration as code) technologies provide adaptable and efficient modifications to cloud infrastructure. The chapter then explores the role of CI/CD (continuous integration/continuous delivery) pipelines in the automation of infrastructure provisioning and maintenance. New content teaches scripting logic, such as variables, functions, and conditional statements. The chapter finishes with coverage of CLI (command line interface) and patch services in AWS, Azure, and GCP followed by a brief discussion of common obstacles to establishing successful automation workflows.

Features

To aid you in fully understanding cloud computing concepts, this course includes many features designed to enhance your learning experience.

- **Running scenario**—Each chapter begins with a running scenario giving real-world context for the technology and concepts presented. The ongoing story provides insight into a variety of cloud computing challenges from the perspective of an IT team preparing to migrate its data center to the cloud.
- **Chapter objectives and outline**—Each chapter lists the concepts to be mastered within that chapter. This list and the chapter outline serve as a quick reference to the chapter's contents and as a useful study aid.
- **Self-Checks**—Each major section within a chapter concludes with a few self-check questions so you can assess your own learning. This brief opportunity for quick recall of information just learned helps solidify new knowledge in your mind. Answers and thorough explanations for each question are provided at the end of the chapter so you can monitor your progress toward your learning goals.
- **Colorful illustrations, screenshots, tables, and bulleted lists**—Numerous full-color diagrams illustrate abstract ideas while screenshots of various cloud platform consoles help you visualize common cloud computing tools, structures, and concepts. In addition, the many tables and bulleted lists provide details and comparisons of both practical and theoretical information that can be easily reviewed and referenced in the future.
- **Notes and Exam Tips**—Each chapter's content is supplemented with Notes that provide additional insight and understanding, while Exam Tips guide you in your preparations for taking the CompTIA Cloud+ certification exam.
- **Chapter summaries**—Each chapter concludes with a brief summary highlighting concepts introduced in that chapter. These summaries help you revisit the ideas covered in the chapter.
- **Key terms and Glossary**—Key terms emphasize the core concepts of cloud computing and are defined in the convenient Glossary, which maps each key term to the chapter where it was introduced.
- **Review questions**—A set of 10 review questions at the end of each chapter provides exam practice to help you assess the effectiveness of your learning and identify weak areas for further study.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Scenario-based practice questions**—Each chapter includes scenario-based questions similar to what you might encounter on the CompTIA Cloud+ exam. These questions put chapter content in real-world context and provide on-time application of covered concepts.
- **Hands-On Projects**—Although it is important to understand the theory behind cloud computing technology, nothing beats real-world experience. To this end, each chapter provides several Hands-On Projects aimed at providing you with practical implementation experience as well as practice in applying critical thinking skills. Hands-On Projects use free trial or free student accounts in the three major public cloud platforms: Amazon Web Services, Microsoft Azure, and Google Cloud Platform.
- **Challenge Projects**—Each chapter concludes with an in-depth project where you implement the skills and knowledge gained in the chapter through real design and deployment scenarios. In each Challenge Project, you can use the cloud platform of your choice, whether that's AWS, Azure, GCP, or some other cloud platform. You can refer to guidance provided earlier in the chapter or course, and you can practice applying research skills to figure out how to complete new, related tasks.
- **Group activities**—One or more projects in each chapter offer optional group work activities to enhance the exploration of various concepts and skills.
- **Exam objectives mappings**—Each section of each chapter begins with a list of exam objectives addressed within that section. This information supports your study efforts if you plan to take the CompTIA Cloud+ certification exam. Every line item of every exam objective is addressed in this course.
- **Acronyms checklist**—As in all things IT, cloud computing relies on extensive use of acronyms. The CompTIA Cloud+ objectives document lists acronyms that are pertinent to the exam's content, and a list at the end of each chapter in this course indicates which of these acronyms are covered in that chapter.
- **User-friendly organization**—Logical arrangement of content consolidates similar concepts for efficient coverage. This organization gives you the opportunity for deeper investigation of particularly rich concepts and skills that are emphasized in the latest CompTIA Cloud+ CV0-004 exam. Topics include a strong emphasis on security, troubleshooting, and business context, with expanded coverage of IoT (Internet of Things), containers, automation, and cloud-native system design.

New to This Edition

Just as cloud technology continues to evolve, so does learning science and the insights available to course designers. In the interest of providing you with the most effective and durable learning experience, this latest edition is packed with improvements and enriched features:

- **Fully updated**—The content maps completely to CompTIA's Cloud+ CV0-004 exam for productive exam preparation. Extensive new content covers all the concepts listed in the new DevOps Fundamentals exam domain. Additionally, the focus of course content is shifted to emphasize cloud-native design and development, including automation, scripting, event-driven architecture, and code pipelines.
- **AI in IT boxes**—Each chapter contains an "AI in IT" box that explores ways AI technologies are used to manage cloud resources, are incorporated into cloud-hosted systems, or are used to augment the work of cloud professionals.
- **Applying Concepts activities**—Embedded in each chapter, these "just-in-time" activities help solidify understanding of concepts as you read, providing immediate practice of relevant skills with step-by-step instructions.
- **Flexible learning environment**—New and updated skills-based projects encourage hands-on exploration of chapter concepts. These projects include thought-provoking questions that encourage critical thinking and in-depth evaluation of the material. The software tools used in the projects are included in Windows or are freely available online, and hardware requirements are kept to a minimum, making these projects accessible to more students in a wide variety of learning environments.
- **Appendices**—Appendices provide exam-to-chapters mapping, chapters-to-exam mapping, a list of Cloud+ acronyms, and a grading rubric for projects. A new appendix gives an in-depth explanation of how to perform subnetting calculations on exams when you don't have access to a calculator.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Text and Graphic Conventions

Where appropriate, additional information and exercises have been added to this text to help you better understand the topic at hand. The following labels and icons are used throughout the text to alert you to these enhanced materials.

Certification

Each main section of a chapter begins with a list of all relevant CompTIA Cloud+ objectives covered in that section. This unique feature highlights the important information at a glance and helps you better anticipate how deeply you need to understand the concepts covered.

AI in IT | An "AI in IT" box in each chapter explores how AI and ML technologies are used within context of the chapter's content, such as cloud networking or cloud automation.

Note | Prolific notes draw your attention to helpful material related to the subject being described and offer expanded insights to enrich your understanding.

Caution | Occasional Caution boxes alert you to potential issues.

Exam Tip | The Exam Tip icon provides helpful pointers when studying for the CompTIA Cloud+ exam.

Applying Concepts

Embedded Applying Concepts activities give you brief, "just-in-time" practice with skills and concepts as you read about them.

Self-Check

To complete the learning cycle, these self-check questions help you practice recalling the information you've read. With answers and extensive explanations provided to readers at the end of each chapter, this low-stakes practice testing helps you assess how well you're learning and what material you might need to review before completing graded work.

Hands-On Projects

Each Hands-On Project in this course is preceded by a description of the project, required resources, and the relevant exam objective. Hands-On Projects help you understand the theory behind cloud design and technologies with activities using some of the most popular cloud services.

Challenge Projects

Challenge Projects are more in-depth assignments that require a higher level of concept application. By providing less detailed guidance, these projects help you explore the "big picture" context of what you're learning. They challenge you to demonstrate a solid understanding and application of skills required for the CompTIA Cloud+ exam and for a career in cloud computing.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Instructor's Materials

Instructors, please visit cengage.com and sign in to access instructor-specific resources, which include the Instructor's Manual, solution and answer guides, a test bank powered by Cognero®, and PowerPoint® presentations.

- **Instructor's Manual:** The Instructor's Manual that accompanies this course includes additional instructional material to assist in class preparation, including suggestions for classroom activities, discussion topics, and additional projects.
- **Solution and Answer Guides:** Guides provide answers to Review Questions, Scenario-Based Questions, Applying Concepts activities, Hands-On Projects, Challenge Projects, and Live Virtual Machine Labs, along with rubrics for the Discussion and "You're the Expert" Video Recording Scenario activities.
- **PowerPoint® Presentations:** This course comes with Microsoft PowerPoint slides for each chapter. These are included as a teaching aid for classroom presentation, to make available to students on the network for chapter review, or to be printed for classroom distribution. Instructors, please feel at liberty to add your own slides for additional topics you introduce to the class.
- **Test Bank:** Questions are aligned with each chapter's learning objectives.
- **Transition Guide:** The guide provides information on what has changed in each chapter in this edition so that instructors know what to expect.
- **Educator Guide:** The guide provides a detailed outline of the corresponding MindTap course.

MindTap for Cloud+ Guide to Cloud Computing

MindTap is an online learning solution designed to help you master the skills you need in today's workforce. Research shows employers need critical thinkers, troubleshooters, and creative problem-solvers to stay relevant in our fast-paced, technology-driven world. MindTap helps you achieve this with assignments and activities that provide hands-on practice, real-life relevance, and certification test prep. MindTap guides you through assignments that help you master basic knowledge and understanding before moving on to more challenging problems. MindTap activities and assignments are tied to CompTIA Cloud+ certification exam objectives. MindTap features include the following:

- **Videos** show concrete skills in two cloud platforms (Amazon Web Services and Microsoft Azure). These author-led videos demonstrate skills, tools, and concepts covered in the chapters, making abstract concepts and skills more concrete and preparing you to perform similar tasks in the Hands-On Projects. While videos assist in raising your comfort level with the platform(s) you're using in the projects, you'll benefit from watching videos for both platforms to develop your expertise across platforms and deepen your understanding of covered concepts.
- **Live Virtual Machine Labs** allow you to practice, explore, and try different solutions in a safe sandbox environment. Each chapter provides you with an opportunity to complete an in-depth project hosted in a live virtual machine environment. You implement the skills and knowledge gained in the chapter through real design and configuration scenarios.
- **Pre- and Post-Assessments** emulate the CompTIA Cloud+ certification exam.
- **Discussion** assignments encourage you to stay current with what's happening in the cloud industry.
- **"You're the Expert" Video Recording Scenario** activities encourage you to practice real-world and employability skills through creating a video in response to a scenario.
- **Lab Simulations** encourage hands-on application of chapter topics.

Instructors: MindTap is designed around learning objectives and provides analytics and reporting so you can easily assess where the class stands in terms of progress, engagement, and completion rates. Use the content and learning path as-is or pick and choose how your materials will integrate with the learning path. You control what the students can access and when. Learn more at cengage.com/mindtap/.

State of Cloud Computing in IT

Most organizations in nearly every industry rely on the cloud to some degree. The Flexera 2025 State of the Cloud Report, found at info.flexera.com/CM-REPORT-State-of-the-Cloud, polled 759 IT professionals from a wide range of industries in late 2024. All respondents report they currently use cloud services at their companies, with 96 percent in the public cloud

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and 74 percent using one or more private clouds. Organizations aren't limiting themselves to a single cloud platform, with 86 percent of respondents reporting their organizations employ a multi-cloud strategy. Furthermore, respondents reported that just over half of their workloads run in a public cloud, and over half their data resides in a public cloud. Interestingly, half also reported they already use generative AI public cloud services.

According to the same survey, the value of cloud computing is reflected in these organizations' budgets as well. About 60 percent of the respondents, representing both small and large businesses, report that their companies already spend more than $2.4 million annually for public cloud services, and a third are spending over $12 million per year. Over two-thirds of surveyed organizations report the operation of an existing cloud-specialist team, such as a CCOE (cloud center of excellence), with another fifth planning to create such a team in the next one to two years. Managing cloud expenses was listed as the top cloud challenge by 84 percent of respondents. Nearly 60 percent said their organization has a FinOps (financial operations) team that handles cloud cost optimization strategies—a nearly 15 percent increase over the previous year's report.

Certifications

According to CompTIA's Tech Jobs Report for 2025, found at comptia.org/en-us/resources/research/tech-jobs-report/, unemployment rates for technology occupations trend lower than the national rate for all job types. Rising salaries for jobs such as cloud architect, cloud infrastructure engineer, and cloud administrator reveal soaring demand specifically for cloud computing expertise. According to Skillsoft's 2024 IT Skills and Salary Report, found at insight.skillsoft.com/it-skills-and-salary-report/p/1, cloud computing job roles pay more and are in higher demand than nearly any other IT sector, exceeded only by IT architecture and design for pay and by cybersecurity for demand.

Traditional degrees and diplomas do not identify the skills that a job applicant possesses, especially in relation to fast-changing cloud technologies. Companies are relying increasingly on technical certifications to adequately identify skilled job applicants, and these certifications can offer job seekers a competitive edge in the job market. Certifications fall into one of two categories:

- Vendor-neutral certifications are those that test for the skills and knowledge required in industry job roles and do not subscribe to a specific vendor's own technology solutions. Some examples of vendor-neutral certifications include all the CompTIA certifications and certifications from CSA (Cloud Security Alliance) or ISC2 (International Information System Security Certification Consortium).
- Vendor-specific certifications validate the skills and knowledge necessary to be successful while utilizing a specific vendor's technology. Some examples of vendor-specific certifications include those offered by Amazon Web Services, Microsoft, Google, Red Hat, Salesforce, and Cisco.

As employers struggle to fill open IT positions with qualified candidates, certifications are a means of validating the skill sets necessary to be successful within organizations. Furthermore, pursuing certifications helps ensure continued job satisfaction and relevance. According to the Skillsoft research, most respondents reported pursuing training to improve their skill set, to be more competitive in the job market, and to improve job security. In most industries, salary and advancement are determined by experience and education, but in the IT field, the number and type of certifications an employee earns also contribute to salary and wage increases.

Certification provides job applicants with more than just a competitive edge over their noncertified counterparts competing for the same IT positions. Some institutions of higher education grant college credit to students who successfully pass certification exams, moving them further along in their degree programs. For those already employed, achieving a new certification increases job effectiveness and satisfaction, which opens doors for advancement and job security. Certification also gives individuals who are interested in careers in the military the ability to move into higher positions more quickly.

What's New with the CompTIA Cloud+ Certification

Now in its fourth iteration, the CompTIA Cloud+ exam (CV0-004) reflects the maturation of the cloud industry with updated technologies and concepts framed in the larger context of business priorities and existing IT systems. With less emphasis on physical host configuration and local virtualization, and much greater emphasis on cloud infrastructure, management, and security, the new exam requires proficiency with standard cloud services that demonstrates a student's understanding and knowledge in cloud networking, security, storage, maintenance, and automation. These concepts are not directly transferrable from the on-premises data center to the cloud but, rather, require an abstraction of functions to a software-defined environment where everything is virtualized and underlying hardware is essentially invisible to the cloud consumer. Common compute, network, security, and storage techniques must be reimagined to take full

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

advantage of the cloud's potential. The new CompTIA Cloud+ exam invites candidates to contemplate these ideas more deeply and in ways more relevant to the cloud ecosystem.

In service of demonstrating these hands-on skills, this exam contains many performance-based questions. These questions present simulations with complex interactions designed to test a candidate's ability to apply concepts and analyze problems. Mastering, rather than simply memorizing, the material in this course will help you succeed on the exam and on the job.

Following are the domains covered in the current CompTIA Cloud+ exam:

CV0-004 domain	Percentage of examination
Domain 1.0 Cloud Architecture	23%
Domain 2.0 Deployment	19%
Domain 3.0 Operations	17%
Domain 4.0 Security	19%
Domain 5.0 DevOps Fundamentals	10%
Domain 6.0 Troubleshooting	12%

About the Author

Jill West authors Cengage courses for *CompTIA Cloud+*, *CompTIA Network+*, *Data Communications*, the *Shelly Cashman Series*, and the *popular Technology for Success*. She has taught, mentored, or counseled kindergarten through college and currently teaches computer technology courses at Georgia Northwestern Technical College. With degrees in psychology, education, and IT, Jill innovates at the crossroads of IT and education, specializing in designing courses for popular IT certifications, presenting at conferences on IT education, and mentoring lifelong learners in IT. Jill and her husband, Mike, live in northwest Georgia with their four children.

Acknowledgments

One of my greatest gifts in life is to get to work with the people who contributed to the development of this book. I've worked with many of these folks through multiple projects, and I count them as friends. Lisa Ruffolo manages the numerous moving parts, keeping files flowing to their proper destinations. But more than that, she provides lively encouragement and constructive feedback with a keen understanding of the book both from a big-picture perspective and with an eye for detail. Zenya Molnar, a recent transplant from the Intro side where I've had the pleasure of working with her on other projects, brings brilliant insights as our learning designer. Cybèle Beckham energetically bridges our content creation work with instructor needs and requests. Sivakumar Krishnamoorthy faithfully keeps us on track with our schedule to ensure timely delivery of this course to our readers. Dani Shaw checks every chapter and project for technical accuracy, and I confidently rely on her eagle eyes to catch mistakes and contribute improvements. Rita Lombard stepped in as our fearless leader and works her magic when challenges arise—she embodies the essence of what all good leaders bring to the table in her balance of compassion and high standards.

I would like to give a personal "thank you" to all our reviewers who contributed expertise, ideas, insights, and passion:

- **Ahmad Al-Omari**, Our Lady of the Lake University
- **Shahpar Islam**, Northern Virginia Community College
- **Kathleen Kalata**, Ferris State University
- **Edward Sykes**, University of Guelph
- **Qingsong "Roger" Zhao**, Louisiana State University Shreveport

Although my name goes on the cover, what you don't see is the myriad of other people diligently contributing, brainstorming, checking, and rechecking to ensure you receive a cohesive and productive learning experience. It truly takes a village to develop an engaging, quality product such as this one.

Thank you to my kids—Winn, Sarah, Daniel, and Zack—for your patience as I work long hours. And thank you to my students for inspiring me to always do better.

A special thanks goes to my husband, Mike West, and all the ways he works tirelessly to keep our household running smoothly, covers the tasks I can't do when flooded with deadlines, and persistently encourages me through the struggles that come with intense schedules and high standards. Mike, I'm so grateful to be doing this life with you.

Read This Before You Begin

Getting into cloud computing is not as difficult as you might initially think. Many cloud providers offer generous free trials, especially for students. In this course, you'll have the opportunity to work in three public clouds: AWS (Amazon Web Services), Microsoft Azure, and GCP (Google Cloud Platform). You can choose any one, two, or all three of these platforms, depending on your needs, preferences, and available resources. They all offer some level of free trial credit and free

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

tier services during the free trial. Some of these options don't even require a credit card if your school collaborates to set up student access. Chapter 1 will walk you through the platform options. Regardless of which platform you decide to use, you can read through the material giving parallel information for the other platforms. You can also watch the author-produced videos for two of these platforms to compare cloud services across ecosystems.

Finally, the Live Virtual Machine Labs in each chapter will give you additional practice with cloud-based skills. These labs are accessed through MindTap and let you work in fully functioning virtual machines to build your own cloud resources. With detailed steps and helpful learning resources, you'll find the support you need to perform complex provisioning tasks. Or you can explore and experiment beyond the parameters of the lab because these virtual machines are real systems and not just simulations.

As for hardware and software, nearly all the projects in this course are completed through your browser. A decent computer with a good browser such as Chrome, Edge, Firefox, or Safari will work well. You use a generative AI chat tool, such as ChatGPT or Perplexity, for a few projects. In several projects, you need a drawing app, such as Visio, or a free, online drawing app, such as app.diagrams.net. In one project, you need access to a CLI on your local computer, and in another project, you create and use a free GitHub account.

These projects have been designed to make the cloud as accessible as possible for students and schools in all kinds of settings and circumstances. Nearly all the projects can be completed for free in their respective cloud platforms, especially if you're still within your free trial limits. Any steps that might accrue charges are identified, along with instructions on how to circumvent charges if you don't have free trial credits available.

Overall, the course offers many layers of options and resources that help you get into the cloud and build hands-on experience, making concepts concrete and memorable while helping you understand and apply the skills you're learning.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CompTIA

Your Next Move Starts Here!

Get CompTIA certified to help achieve your career goals and gain a powerful, vendor-neutral credential that is trusted by employers.

Save 10%
when you purchase your exam voucher from CompTIA.org.

Use code:
Cengage10



Why get CompTIA certified?

Increase your confidence
91% of certification earners show increased confidence.*

Earn more money
77% of IT pros got a raise within six months of earning their certification.*

Stand out to employers
64% of IT decision makers say certified employers add additional value.**

Join a global community
92% of IT professionals hold at least one certification.**



Get ready for exam day.

- **Download the exam objectives:** Visit CompTIA.org to find the exam objectives for your IT certification and print them out. This is your roadmap!
- **Create your study plan:** Decide how many hours each week you are going to dedicate to studying, choose your preferred study tools and get to work. Studying is a unique experience. Download a study plan worksheet on CompTIA.org.
- **Get certified:** If you haven't already, use the coupon on this page when you purchase your exam voucher and schedule your exam. CompTIA offers flexible testing options to fit your busy life.

Choose your testing option.

Online testing
Earn a CompTIA certification online, from your home – or any quiet, distraction-free, secure location – at a time that's convenient for you.

In-person testing
Test at any of the Pearson VUE test centers around the world, where you can use their equipment under the supervision of a proctor.

To purchase your exam voucher and learn how to prepare for exam day, visit CompTIA.org.

*Pearson VUE 2021 Value of IT Certifications
**2021 Global Knowledge IT Skills and Salary Report

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Chapter 1

Introduction to Cloud Computing

Chapter 1 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Describe the primary characteristics of cloud computing.
2. Compare cloud deployment models.
3. Compare cloud service models.
4. Identify popular cloud service providers.
5. Apply troubleshooting methodology.

Chapter 1 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 1-1: Characteristics of Cloud Computing
- CompTIA Cloud+ Certification
- Other Cloud Certifications
- What Is Cloud Computing?
- What Do I Need to Know?

Section 1-2: Cloud Deployment Models
- Public Cloud
- Private Cloud
- Hybrid Cloud
- Multi-Cloud
- Community Cloud

Section 1-3: Cloud Service Models
- Common Cloud Service Models
- Managed Services

Section 1-4: Cloud Services
- PaaS and IaaS Providers
- Common Cloud Services
- AI/ML (Artificial Intelligence/Machine Learning) Services
- IoT (Internet of Things)

Section 1-5: Troubleshooting Methodology
- Common Cloud Computing Problems
- Troubleshooting Steps
- Preventive Measures

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 1 Scenario The clock says 8:56 a.m. as you slide in through the basement door of your building, shaking the rain off your umbrella and wiping your feet on the doormat. You're right on time for once! You notice the empty boxes still sitting by the doorway after yesterday's delivery of new cabling supplies. When you glance up at the security camera in the lobby area, you wonder yet again who might be sitting on the other side of that camera watching you move through the office space toward your desk. You make a mental note to ask Dianne, the security guard at the front gate, where the camera feed goes and how often someone sits staring at the feeds … just out of curiosity, of course.

As you hang your jacket on the back of your chair and prop your umbrella in the corner near your desk, Kendra, your boss, calls out from her office, "Come on, everyone—we've got a meeting!"

Oh, that's right. Your IT team is discussing the upcoming cloud migration today. At the meeting, the CIO of the private school you work for summarizes an AWS (Amazon Web Services) conference he recently attended touting the benefits of cloud computing. He's envisioning dollar signs as he excitedly describes all the ways your school can save money by migrating large portions of your current network resources to the cloud. "We won't have to pay so much for IT hardware anymore, our expenses will flex with our changing IT needs throughout the year, and even our electric bill will go down. Plus, the cloud apps we can use in the classroom are good for our students. We'll be able to attract more students from tech-savvy families moving to the area because of the tech start-ups and other IT companies migrating here."

Sounds like a good plan. Now it's up to your team to figure out how to do it. Your cloud conversion team will be led by your boss, Kendra, the network administrator; she worked with existing cloud deployments at her previous job, and this will be her first cloud migration project. You've been recruited to help, along with a recently hired coworker, Malik, who passed his A+ certification exam last spring while taking night classes. He then asked to be transferred to IT from his former position as a security guard. Except for Kendra, none of you have much experience with the cloud yet, other than binge-watching movies on the weekends and using your personal Gmail and social media accounts. The technology seems a little futuristic and "out there" when you think about it. But you're also secretly excited to learn some new stuff. Some of your IT friends working across town at another private school migrated some systems to the cloud a year ago, and you're excited to jump on the cloud bandwagon.

Throughout the meeting, the entire team discusses what kinds of information you all need to know to have an intelligent conversation about the cloud. As you wrap up your action list in preparation for your next meeting, the questions you want to answer for yourself include the following:

- What is cloud computing?
- What new skills do I need to learn?
- Who owns the cloud?
- How does the cloud affect the security of my organization's data?
- What are common cloud problems we'll need to anticipate?

Section 1-1: Characteristics of Cloud Computing

Certification

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.2 Compare and contrast aspects of compliance and regulation.

4.3 Given a scenario, implement identity and access management.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CompTIA (Computing Technology Industry Association and pronounced "comp-TEE-uh") released its first cloud computing certification, Cloud Essentials, in 2011 (Table 1-1). Now called Cloud Essentials+, it is intended for non-IT professionals or for IT professionals needing to bridge the gap between technical concepts and business concerns. The more technical Cloud+ certification was first released in 2015. Its latest version is the exam this course is designed to help prepare you to pass.

Table 1-1 CompTIA cloud certifications

Cert version	Cert name	Year released
CLO-001	Cloud Essentials (Initial release)	2011
CLO-002	Cloud Essentials+ (Current version)	2019
CV0-001	Cloud+ (Initial release)	2015
CV0-004	Cloud+ (Current version)	2024

While there are no prerequisites for taking the exam, Cloud+ builds on the knowledge required for other technical certifications (Figure 1-1), including the following:

- **CompTIA A+**—Covers skills required for IT technical support specialists
- **CompTIA Network+**—Covers foundational networking concepts and skills
- **CompTIA Security+**—Surveys IT security technologies and strategies at a level necessary for any IT professional

Figure 1-1 CompTIA Cloud+ is an infrastructure-specialty certification



While these certifications ("certs" for short) are not required before taking the Cloud+ exam, the knowledge covered by their objectives is directly relevant to the skills required of a Cloud+ certified technician. Cloud+ takes the foundational concepts covered in these earlier exams and applies that information to a cloud environment. This course assumes you already have knowledge equivalent to A+, Network+, and possibly Security+ level skills and concepts.

CompTIA Cloud+ Certification

One aspect that sets Cloud+ apart from most other cloud computing certs is that Cloud+ is vendor neutral. This means that, throughout this course, you will learn cloud computing concepts that span the full range of cloud computing services rather than targeting specific skills for working with any one cloud provider's platform. You might pursue a Cloud+ certification for the following reasons:

- To prepare for a new job or a promotion that likely will include responsibility for interacting with an organization's existing cloud services.
- To build a foundational understanding of cloud computing in preparation for vendor-specific cloud certifications.
- To complement certifications in other specialty areas (such as infrastructure, security, database management, or programming).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- To develop a big-picture perspective of cloud computing technologies, major players, and industry expectations in preparation for choosing vendors and migrating on-premises ("on-prem" for short) services to the cloud.

You can learn more about Cloud+ and its exam objectives, target audience, and related job roles at comptia.org/certifications/cloud.

Other Cloud Certifications

Two other vendor-neutral cloud certifications are the CCSK (Certificate of Cloud Security Knowledge) from **CSA (Cloud Security Alliance)** and the more advanced CCSP (Certified Cloud Security Professional) from ISC2 (International Information System Security Certification Consortium).

The CCSK requires high-level knowledge of how to secure data in the cloud in the context of international standards. It's an open-book and online exam covering topics such as application security, automation, compliance, incident response, AI (artificial intelligence), and DevSecOps (development, security, and operations). You can learn more about the CCSK at cloudsecurityalliance.org/education/ccsk.

The CCSP also focuses on the security side of cloud computing at an expert level and requires a minimum of five years paid, full-time IT work experience, three years of which must be in information security and one year in work specific to one or more of the six exam domains. Alternatively, you must already have ISC2's CISSP (Certified Information Systems Security Professional) certification. You can learn more about the CCSP at isc2.org/certifications/ccsp.

Several highly respected, vendor-specific cloud certifications prove increasing levels of proficiency with specific cloud platforms and products. **CSPs (cloud service providers)**—such as Amazon, Microsoft, and Google—are companies that offer one or more cloud services to other businesses and individuals. They invest significant resources into providing training and certification programs for industry professionals. Some of the major vendor cloud certifications include the following:

- **AWS (Amazon Web Services)**, according to many statistics, holds the lion's share of the public cloud market. AWS role-based certifications follow one of several certification tracks or learning paths, such as Architect, Developer, or Operations. Within each path, roles progress from Foundational and Associate to Professional and, in some cases, Specialty levels. Popular AWS certs include the following:
  - **Cloud Practitioner**—This nontechnical, entry-level AWS certification is appropriate for professionals in technical, managerial, sales, purchasing, or financial roles.
  - **AWS Certified Solutions Architect – Associate**—This intermediate AWS cert focuses on designing applications and systems in AWS. The related, advanced cert is the AWS Certified Solutions Architect – Professional.
- A close second to AWS in the public cloud market, Microsoft supports the **Azure** cloud platform. Azure certifications are classified by four levels: Fundamentals, Associate, Expert, and Specialty. Microsoft offers several exam pathways to earn its Azure certifications, including the following popular certs:
  - **Microsoft Certified: Azure Fundamentals**—This entry-level Azure certification is appropriate for candidates with nontechnical backgrounds or for technical professionals validating foundational knowledge of cloud services.
  - **Microsoft Certified: Azure Administrator Associate**—This intermediate Microsoft cert focuses on implementing, monitoring, and maintaining Azure-based cloud solutions.
- **GCP (Google Cloud Platform)** is another major contender in the global cloud market and is quickly growing. While GCP offers fewer certifications, they are highly relevant if you will be working in a GCP ecosystem. Popular GCP certs include the following:
  - **Cloud Digital Leader**—This nontechnical GCP certification requires knowledge of Google Cloud products and services in addition to basic cloud concepts.
  - **Associate Cloud Engineer**—This intermediate-level GCP certification covers setting up, deploying, and securely operating a cloud solution. Google recommends at least six months of related work experience before taking the exam.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **VMware**, a well-known provider of virtualization services and now owned by Broadcom, offers solutions that are highly integrated into some public cloud platforms. VMware's cloud certs are classified by levels such as Technical Associate, Professional, Advanced Professional, and Design Expert. Two examples include:
  - **VMware Certified Technical Associate – Cloud Management and Automation**—This entry-level cert demonstrates a basic understanding of virtualization and cloud concepts.
  - **VMware Certified Professional – VMware Cloud**—This intermediate exam requires skills to install, configure, and administer the VMware cloud environment.

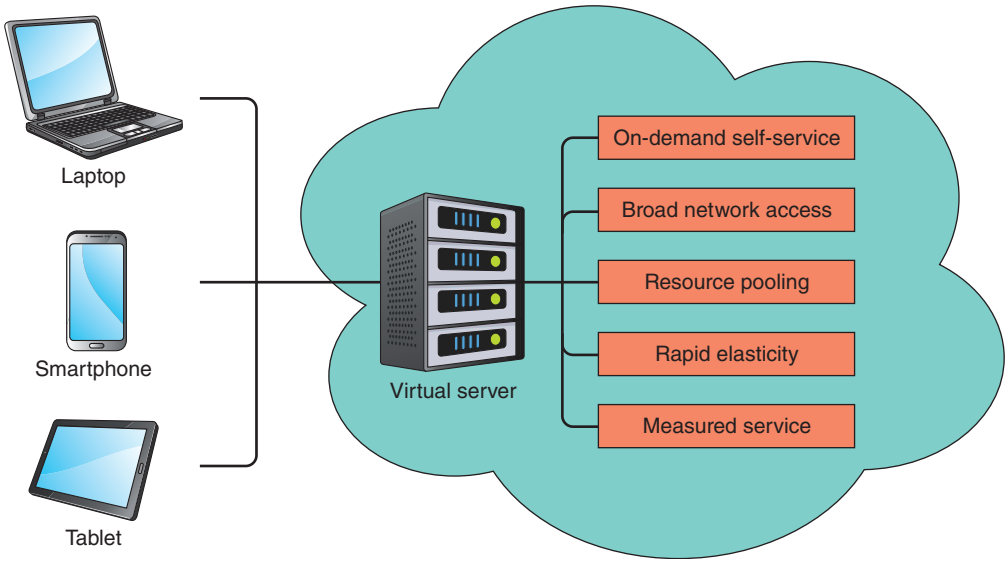
Note 1-1 | An entry-level cloud computing job might be called a junior cloud engineer, cloud implementation manager trainee, cloud technical architect, or junior cloud analyst, among many other possibilities. More advanced cloud computing positions might go by titles such as CSA (cloud systems administrator), senior cloud engineer, or cloud infrastructure engineer.

AI in IT | Some of these same certification organizations also offer certs in AI, even for nontechnical, business professionals. For example, AWS offers the AWS Certified AI Practitioner, and GCP recently released the Generative AI Leader cert. Both certifications require foundational knowledge in AI to assist business leaders in making decisions related to AI-driven initiatives.

What Is Cloud Computing?

What makes cloud computing so attractive to companies who want to develop cutting-edge innovations and maximize their bottom line? **NIST (National Institute of Standards and Technology)** is a nonregulatory agency of the U.S. Department of Commerce that promotes measurements, standards, and technology in support of innovation and industrial competition. According to NIST, a technology must exhibit five essential characteristics to fall under the classification of "cloud computing" (Figure 1-2).

Figure 1-2 Characteristics of cloud computing



- **On-demand self-service**—In a traditional network or data center setting, network resources must be carefully planned, purchased, configured, and implemented. It can take months after deciding to institute a new resource until that resource becomes available. Many people are typically involved in the decision-making and implementation processes. In contrast, cloud resources, such as VMs (virtual machines)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

or user accounts, can be created at any time by the cloud customer or other authorized users. The ability to add, adjust, or remove cloud resources at any time by the service subscriber or other approved users is called **on-demand self-service**. There's no built-in delay to order, install, or configure hardware because cloud resources are virtualized on top of existing hardware at the service provider's location. For example, suppose you need to add a web server to your network. Within a few minutes, you can "spin up," or create, a VM in the cloud; optimize its virtual hardware resources (such as memory and processing power) and network settings to support web services; and then install the web server software, website files, and content to produce a live website. This process, once established, can even be automated.

- **Broad network access**—Traditional network resources are mostly available only to users located within a specific geographic area. To access and configure these resources from outside the network, the user must "remote in" using remote access technology. Cloud services, however, can be accessed and configured from anywhere on the Internet and using any of several device types, such as a laptop, smartphone, or tablet, which works well for a BYOD (bring your own device) environment. This type of access is called **broad network access**, and it gives cloud computing the flexibility of access to a much larger user population in a much wider variety of circumstances.
- **Resource pooling**—In traditional networks, most physical and virtual resources reside at the organization's own location(s), and all those resources are dedicated to the organization's own use. If a server is functioning at 10 percent capacity to meet the organization's needs, the remaining capacity is simply unused. Cloud providers use **resource pooling** to maximize the potential capacity of physical and virtual resources, which simultaneously serve multiple customers—called **tenants**—at any given time. The cost benefits of this efficiency are passed along to customers to some degree. A single physical server might be performing compute operations for three or four different cloud customers, and a single customer might be running a virtual resource, such as a database, whose data is held by multiple servers at multiple locations owned and operated by the cloud provider. In most cases, cloud customers don't know exactly where their resources are hosted geographically; they only know how to access them through the Internet. Resource pooling is made possible by **multitenancy**, which means to support multiple customers' processes running on the same hardware. Multitenancy maximizes the potential usefulness of physical and virtual resources' capacity by simultaneously serving multiple customers.
- **Rapid elasticity**—**Rapid elasticity** refers to a cloud resource's ability to be added, released, increased, or decreased quickly—even automatically—according to demand. For example, suppose your company is preparing to run a Super Bowl ad and anticipates a tremendous increase in traffic to their website during the game. On a traditional network, you would need to order and configure additional servers to help support the increased traffic. This could take weeks of preparation and would be very expensive. After the ad traffic subsided, you would be left with several servers whose purpose had waned. If your web server is hosted in the cloud, however, you could provision additional web servers in minutes—or program them to auto-provision as needed—and then decommission those servers as the ad-generated traffic tapers off. You would only pay for the cloud servers while you used them, and once decommissioned, they would incur no ongoing costs.
- **Measured service**—The ability to track your usage of cloud resources at a granular level is what makes cloud computing a **measured service**. You are charged for the resources and services that you use as you use them. Many cloud providers allow for detailed tracking of these activities and charges. For example, AWS bills for VM instances by the second, which avoids costly roundups to full-minute or hourly pricing. This pay-as-you-go approach, if configured properly, can sometimes save companies a great deal of money even before other optimizations are made to a system's design that allow for the best advantages offered by cloud technology.

Cloud providers often describe additional characteristics and benefits to cloud computing, including the following:

- **Self-patching/self-healing infrastructure**—Because much of the cloud is automated, a cloud network can patch or repair itself when encountering certain types of problems.
- **Adaptive, intelligent security**—Cloud computing increasingly takes advantage of AI technology to improve built-in security defenses.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Cross-platform**—Cloud services can be accessed and used from devices running different OSs (operating systems). Typically, cloud resource configurations are made through a web portal accessed via a browser, through a command line tool, or through an application interface.

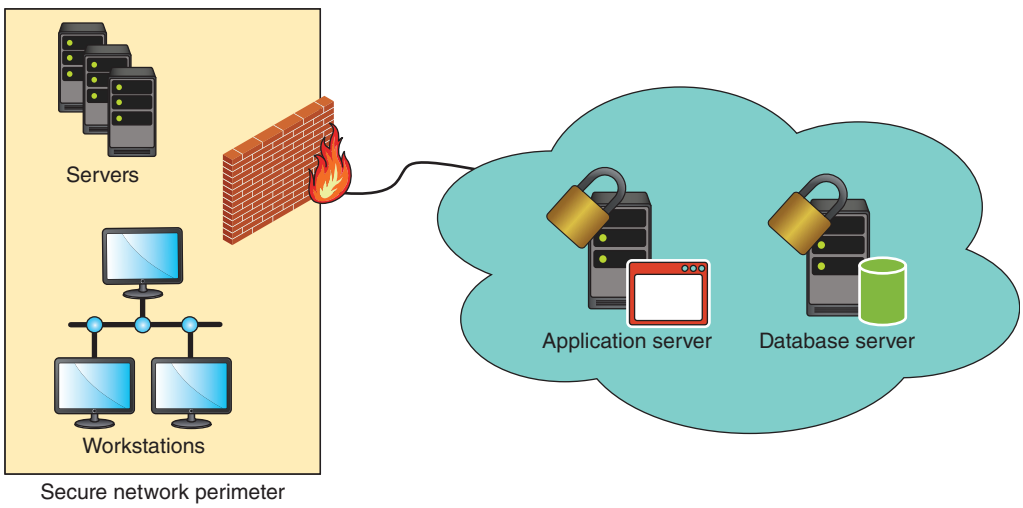
Organizations consider many reasons for transitioning to a cloud environment. They might be facing the expiration of a lease at their data center location; they might be searching for ways to optimize their services and become more competitive; or they might expect that cloud services will improve their bottom line. The cloud transition is not an overnight shift—organizations adopt cloud services in phases, and many will never become fully cloud-centric. Even after they've begun to make the transition to the cloud, business and user needs continue to change, and so do their cloud service requirements. Whether the changes are external, such as changes to regulations or laws, or internal, such as business mergers (two businesses blending into one), acquisitions (one business buying another), or divestitures (one business splitting off part of itself), the cloud services that support the business will need to be adapted, migrated, or replaced. All these changes require specialized skills from IT technicians to configure, deploy, secure, maintain, manage, and troubleshoot cloud computing services and resources.

What Do I Need to Know?

Cloud computing is a fairly recent—and rapidly changing—development in IT that is simultaneously revolutionizing the industry. Skills and knowledge from even five years ago are insufficient to meet the demands of managing today's cloud-hosted resources. IT professionals throughout the industry are asking what new skills will help them stay relevant as cloud computing technologies emerge and mature. The following list explores areas of professional growth most needed for cloud computing professionals—or any IT professional interacting significantly with the cloud:

- **Cybersecurity skills**—Traditionally, IT security focuses on maintaining a protective perimeter around the on-prem data center, managing all traffic into and out of that secure perimeter (Figure 1-3). As data, applications, and other resources move to the cloud, security must be built into the resources themselves so it travels with each resource. Administrators must also carefully consider what data can be outsourced to the cloud, as some types of regulated data are restricted to on-prem storage only. Cloud service providers offer service options with built-in security compliance measures, which can relieve customers of needing to configure these requirements themselves. While the cloud customer is still responsible for enabling cloud security measures and configuring them in ways that best fit their needs, most cloud providers offer many tools—some that are free—to help with this effort. A cloud IT professional must become familiar with these tools and best practices in using them.

Figure 1-3 Cloud security can't rely on a secure perimeter



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Cross-discipline expertise**—The broad possibilities inherent in cloud computing require that cloud technicians understand the business and organizational context of services provided through the cloud. The more you understand, the more valuable you'll make yourself to potential employers. Important areas include the following:
  - The organization's business goals and the business processes and workflows that affect every department within an organization
  - Software development processes, including automation, configuration management, virtualization, monitoring, and CI/CD (continuous integration/continuous delivery) pipelines
  - Infrastructure concepts, skills, and tools—networking in the cloud transitions the concrete configurations of a local network into a more abstract, virtual layer, which will best make sense to you if you already understand the principles of local networking
  - Security vulnerabilities, technologies, and best practices that are specific to cloud-hosted resources
- **Automation scripting skills**—One of the biggest advantages of moving to the cloud is gaining a greater degree of **automation**, which is the use of technology to perform a task with minimal human intervention. The underlying virtualization layer and the encompassing monitoring techniques allow for granular and responsive automation. Developing the skills required to script extensive automation processes (called **orchestration**) in a way that is both efficient and reliable, and then continuing to refine them, is where cloud network admins often are most challenged.
- **Lifelong learning**—Cloud technologies are in constant flux, even more so than on-prem technologies. CSPs are innovating at an astonishing rate. A service that adequately meets your needs when you first migrate to the cloud could be overshadowed by a much improved—and cheaper—service a few months later. Staying relevant requires monitoring your own learning by identifying what you don't know, finding the information resources you need, and following through on your learning process. The most important IT skill is the ability to teach yourself.

Self-Check

1. Which cloud platform does the CompTIA Cloud+ certification target?
   a. AWS
   b. Azure
   c. GCP
   d. None

2. Which of these certifications requires the least technical skill?
   a. CompTIA Cloud+
   b. Associate Cloud Engineer
   c. AWS Cloud Practitioner
   d. VMware Certified Professional

3. Which of the following is *not* an essential characteristic of cloud computing technologies, according to NIST?
   a. Adaptive security
   b. Broad network access
   c. Rapid elasticity
   d. On-demand self-service

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 1-2: Cloud Deployment Models

Certification

1.2 Explain concepts related to service availability.

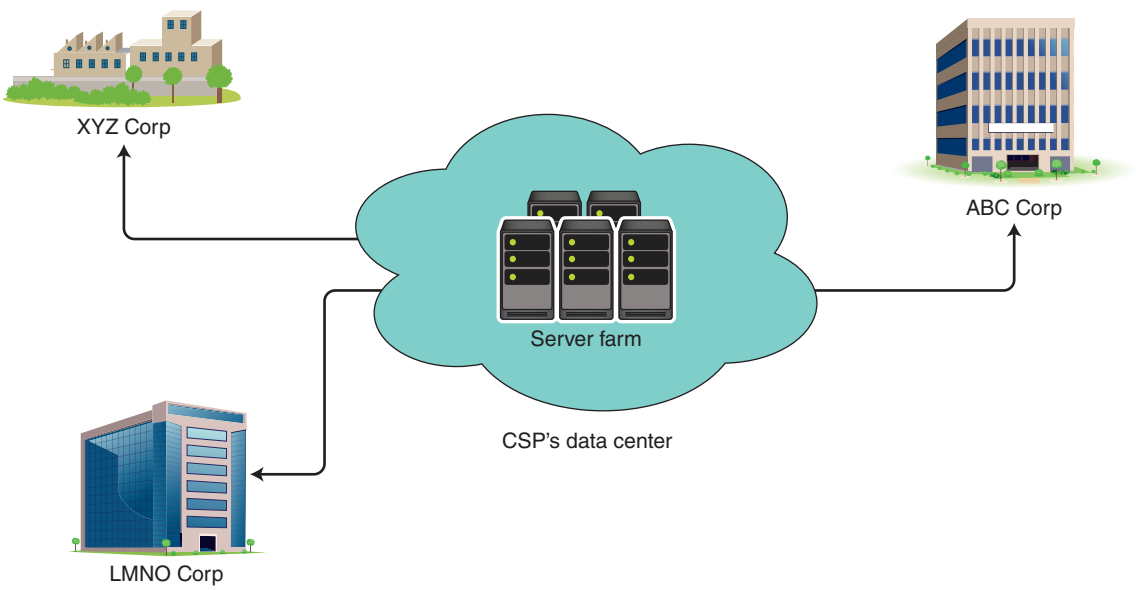
2.1 Compare and contrast cloud deployment models.

One advantage of cloud computing is that a business can use someone else's hardware to host its applications, data, and network infrastructure. One disadvantage of cloud computing is that a business must typically rely on someone else's hardware to host its applications, data, and network infrastructure. These two statements aren't redundant or contradictory—cloud computing adds convenience while also reducing an organization's control of the hardware supporting their IT resources. This creates specific security, monitoring, and configuration concerns that vary according to the type of cloud services an organization decides to use. Most organizations don't jump directly into a cloud-only deployment, so you should be aware of the various options for who owns what resources in the cloud, who is able to use the hardware hosting those resources, and how various responsibilities are distributed between the CSP and cloud customer. In this section, you learn about variations based on public cloud services, private cloud services, and hybrid options.

Public Cloud

When people think of cloud computing, the public cloud is generally what they have in mind. Figure 1-4 illustrates **public cloud** services hosted on hardware resources at the CSP's location, and those physical resources can be shared among many customers. The CSP might be a business, an academic organization, or a government entity. It provides cloud services (paid or free)—such as storage space, applications, compute capacity, or network functions—that are available to the public.

Figure 1-4 A public cloud is hosted on hardware shared by multiple customers



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

With a public cloud, the CSP manages the hardware. In fact, customers can't access the hardware or the CSP's data centers. Instead, the customer relies deeply on the CSP's security, availability, and compliance measures to protect the customer's data and other resources hosted in the public cloud. However, this doesn't mean the customer has no responsibility for the well-being of its resources or that all CSPs provide sufficient data center security. The following is a list of precautions that public cloud consumers can take to help ensure the security and availability of their public cloud-hosted resources:

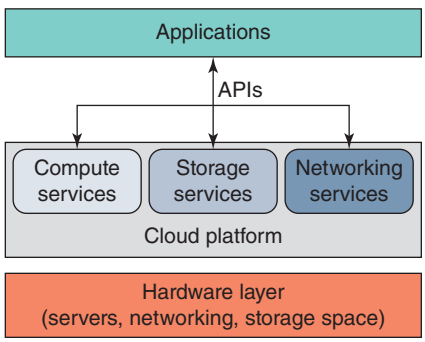
- Check the requirements defined by the CSP's **SLA (service level agreement)**, which is a contract defining service standards from a company to its customer and options for recourse should those service levels not be met. The SLA might include terms of agreement such as requiring encryption of all data in transit or in storage, geographic location of data storage (which might be regulated by law, depending on your organization and the type of data), consistency of service availability, and penalties for failure in any of these areas. Security and other types of failures can be quite costly for your organization, and the SLA should include some kind of compensation for damages resulting from problems on the CSP's end.
- Investigate the CSP's security, monitoring, and availability measures. For example, ask about how the CSP protects against data leakage between tenants on their multitenant infrastructure. Ask whether the CSP relies on third-party vendors, what measures are in place to ensure those organizations' compliance with standards and governance measures, and how those services are handled if there's an availability or security concern with those vendors.
- Understand the CSP's recommendations and requirements for your own organization's standards of operation. This might include access monitoring, license management, self-service troubleshooting, trouble ticket reporting requirements, or encryption processes that you're responsible for.

Private Cloud

Private cloud services are hosted on hardware resources used exclusively by a single organization. This hardware might be located in a CSP's data center and dedicated to one customer, or the hardware might be located in the organization's own data center. What makes a cloud private is that no one else is allowed to use the hardware for their own cloud, regardless of where that hardware is located. This increases security and the organization's control over the exact configuration of the hardware used. In fact, the need to keep sensitive data on-site is often the motivating factor in deploying a private cloud on-prem instead of using public cloud services.

You might wonder how a private cloud hosted on-prem is different from the traditional data center that most organizations already have. A traditional data center might use virtualization for network or data services. **Virtualization** is a time-tested technology designed to emulate parts or all of a computer or network, such as when hosting many services on a single, physical server. While virtualization provides some flexibility in how the hardware interacts with virtual services, the configurations for these services are still closely tied to the underlying physical hardware. A cloud is more abstracted from the physical hardware in the data center, relying on a layer of communication that is managed by comprehensive virtualization software. These communications are enabled by **API (application programming interface)** connections, which are requests in a specific format between applications or other types of systems. As illustrated in Figure 1-5, APIs provide the communication interface between various components of a deployment, such as a web server with serverless code or a database with an application. This sophisticated virtualization layer more specifically defines what the "cloud" is, not necessarily access over the Internet. If you deploy cloud technology on your own hardware, you can have both at the same time: an on-prem data center that is running a private cloud. To qualify as cloud technology, all of NIST's five essential characteristics must apply, which is not true of older virtualization technologies but is true of modern private clouds.

Figure 1-5 Cloud services and underlying hardware communicate using APIs



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 1-2 | If you'd like more information about virtualization and hypervisors, use your Cengage Unlimited subscription to go to *CompTIA Network+ Guide to Networks*, 10th edition, Chapter 7, and read the section titled "Virtual Architecture."

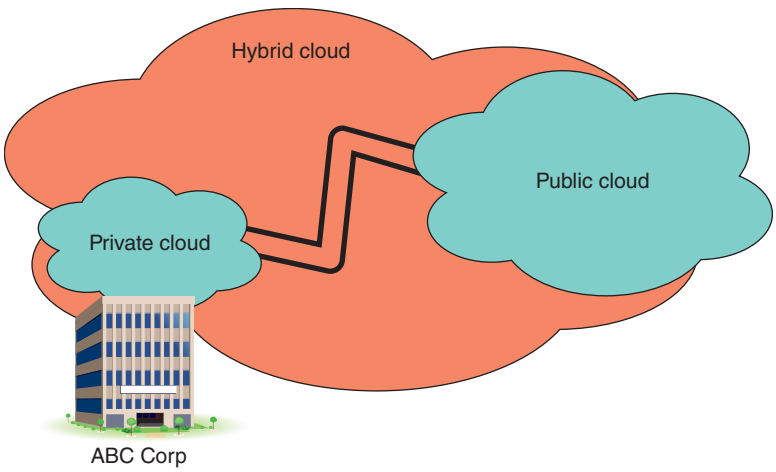
If you don't have a Cengage Unlimited subscription, you can find more information at cengage.com/unlimited.

Note 1-3 | Private clouds aren't only for businesses. An individual might create a private cloud for their home network. For example, Nextcloud is an open-source cloud storage and application service that can be installed on a home server and accessed through the Internet. It functions similarly to Dropbox, except the cloud is owned and managed by the consumer rather than a third-party organization.

Hybrid Cloud

A **hybrid cloud** is a mix of both public and private cloud components, or a combination of some cloud and some traditional on-prem services, where those components and services interact with each other in a direct and seamless manner (Figure 1-6). This scenario might occur as an interim phase during a company's transition to the cloud, or an organization might maintain a hybrid arrangement indefinitely.

Figure 1-6 A hybrid cloud is a mix of public and private clouds or a combination of cloud and on-prem services



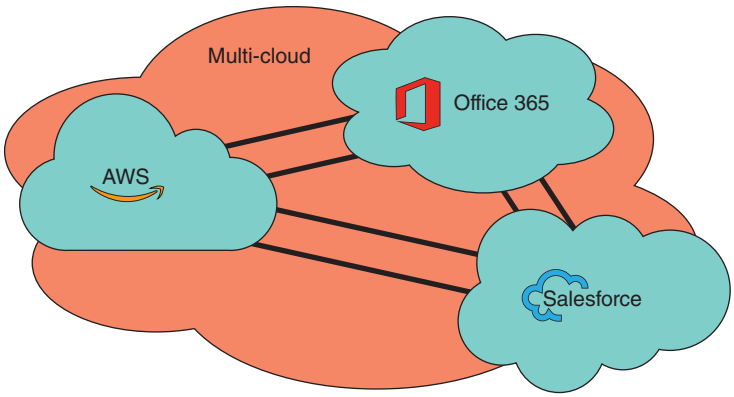
Many organizations take a "cloud-first" or "cloud-centric" approach, where they migrate every application and service that reasonably can be moved to the cloud. However, this migration most often must occur in phases. The organization might have begun its transition from its own hardware and software resources to cloud-based resources, but it is not yet a "cloud-only" network—some of its data and network resources are still hosted locally in its own, traditional data center. Thus, the organization temporarily relies on a hybrid cloud.

Many other organizations, however, find that not every application, dataset, or service is suitable for the cloud. In these cases, organizations make a partial transition with the intent of retaining some of their traditional data center indefinitely. At the same time, the resources across the different hosting environments are integrated and directly interact with each other. Perhaps a database hosted on-prem communicates with a web server hosted in the cloud. Or compute resources hosted in a private cloud can rely on temporary, virtualized hardware resources in a public cloud to accommodate short spikes in demand, which is called **cloud bursting**.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In reality, a hybrid cloud model is the most common cloud deployment because relatively few organizations are ready or able to commit fully to the cloud, and even fewer organizations avoid the cloud entirely.

Figure 1-7 A multi-cloud intertwines cloud resources from multiple cloud providers



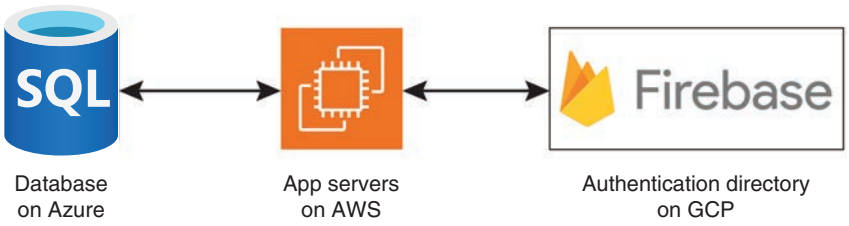
Multi-Cloud

A **multi-cloud** model involves using cloud services from multiple vendors at one time (Figure 1-7). Some organizations commit to a single-cloud environment, such as AWS or Azure. All the cloud services they use are hosted by that one CSP. However, as cloud services mature and the demand for cross-environment functionality increases, many organizations choose the best-in-breed option for each service or system individually. This means working with multiple CSPs to custom build a cloud best suited to the organization's needs.

On the simple side, a multi-cloud might appear like a small business using two or more cloud applications from different vendors—such as email and office suite applications from Google, a file storage and sharing service from Dropbox, and an online presentation platform like GoTo Meeting from LogMeIn, Inc. On the more complex side, an organization might shift its payroll software to a cloud hosted by AWS, and it might use a cloud-based **CRM (customer relationship management)** solution from Salesforce to help manage communications with customers, potential customers, vendors, and associates. Additionally, the organization might run websites on Azure and use GCP for application development. What makes it a multi-cloud setup is that multiple CSPs are involved, even if the services themselves don't directly interact with each other.

In contrast, whereas a multi-cloud deployment model incorporates services from multiple cloud vendors, multi-cloud tenancy implies more integration between these services. In other words, the services running on each cloud platform interact directly with each other even though they are hosted by different cloud providers. For example, for a single app, you might host a database on Azure, run your mobile app servers on AWS, and run an authentication directory from GCP (Figure 1-8). These various components are hosted on different cloud platforms, but they communicate with each other to run a functional mobile app. Whether a system is deployed on a single cloud platform or multiple platforms, the components communicate via APIs. So long as the system relies on components hosted on different cloud platforms, the system is said to have a multi-cloud tenancy.

Figure 1-8 Resources on different cloud platforms as part of the same system infrastructure



Note 1-4 | The rise in popularity of hybrid and multi-clouds necessitates an expansion of skills for IT professionals. No longer is it sufficient to develop expertise in only one cloud platform. Your skill set should cross several cloud platforms with many levels and categories of available services. This is a major advantage offered by the vendor-neutral CompTIA Cloud+ certification, which emphasizes cloud concepts and skills across platforms rather than focusing on specific features and services from a single cloud platform.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Community Cloud

A **community cloud** is accessible to multiple organizations with similar concerns but is not available to the general public. One example might be several regional hospitals that share access to patient records. The patient database in the cloud would be accessible to member hospitals and certain other healthcare providers but would not be available to other organizations or the public. One of the member organizations might host and manage the community cloud resources either on- or off-premises, or it might be provided by a third party.

Self-Check

4. What factor limits the capacity of a private cloud to provide resources to users?
   - **a.** Operating system
   - **b.** Security settings
   - **c.** Hardware
   - **d.** Geographic location
5. An AWS-hosted network linked to Google Workspace productivity apps is an example of what kind of cloud?
   - **a.** Community cloud
   - **b.** Private cloud
   - **c.** Hybrid cloud
   - **d.** Multi-cloud
6. Which of the following scenarios is most likely to require cloud bursting?
   - **a.** A database after a cloud migration
   - **b.** A firewall during application testing
   - **c.** A web server during an advertising campaign
   - **d.** An application at the end of the workday

Check your answers at the end of this chapter.

Section 1-3: Cloud Service Models

Certification

1.1 Given a scenario, use the appropriate cloud service model.

1.5 Explain the purpose of cloud-native design concepts.

1.10 Compare and contrast methods for optimizing workloads using cloud resources.

Cloud computing takes functions and resources that would normally happen on a local network, abstracts those functions to a software-defined level, and then provides those services back to the local network or to remote users from across the Internet or another network. For example, consider the traditional firewall appliance. These physical devices are installed in data center racks and positioned on the network to filter traffic between the internal network and the outside world. They can require a significant amount of upkeep to manage firewall rules, logs, and updates. It is also expensive to upgrade these devices, and they have a limited life expectancy.

Cloud computing allows network administrators to export this firewall function to the cloud as a virtual firewall service, called FWaaS (Firewall as a Service). FWaaS can be expanded rapidly as needed, has no limits on life expectancy, and automates much of the maintenance and upkeep. Firewall coverage for remote office locations or work-from-home employees is more consistent. And firewall functions can be more granularly customized throughout the network. Cato Networks and Palo Alto Networks both offer FWaaS product options.

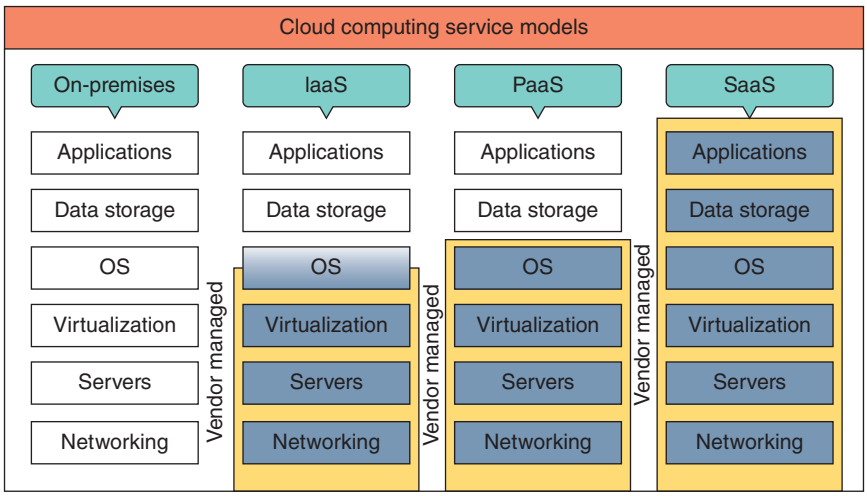
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

While FWaaS is a specific type of cloud computing service, other service types are more encompassing. This section discusses the most common cloud service models.

Common Cloud Service Models

The usual classifications of cloud service models, as defined by NIST and compared in Figure 1-9, are as follows:

Figure 1-9 At each progressive level, the vendor takes over more management and configuration responsibility



- **SaaS (Software as a Service)**—Most cloud consumers are most familiar with **SaaS** (**Software as a Service** and pronounced "sass"), which is the provision of software through the cloud. Applications can be accessed from different types of devices without the customer having to manage any of the underlying infrastructure such as the network, servers, operating systems, or storage. Google's office productivity suite (Google Workspace) is a great example of SaaS. Google Docs, for example, is a document processing application that is accessed through a web browser and can be used on nearly any web-enabled computer or mobile device, regardless of the device's operating system or other installed apps. Dropbox, an online file storage app, and GoTo Meeting, an online conference call app, are also popular examples of SaaS. The consumer might be able to manage user-specific configuration settings, such as sharing options, file storage location, view settings, contacts, permissions, or even default printer.
- **PaaS (Platform as a Service)**—**PaaS** (**Platform as a Service** and pronounced "paz") is an intermediate layer of cloud capability that allows consumers to deploy applications on various platforms without having to manage the lower-layer infrastructure such as the network, servers, or storage. The consumer can manage applications and possibly some of the environment hosting those applications. For example, Google's App Engine can host a custom mobile app accessed by an organization's sales force, while another company uses App Engine to develop new games.
- **IaaS (Infrastructure as a Service)**—The more technically challenging **IaaS** (**Infrastructure as a Service** and pronounced "i-az") allows consumers to deploy a cloud-based network. For example, this network might contain storage, user desktops, network infrastructure devices (such as routers and load balancers), network security devices (such as firewalls and access control lists), and network services (such as DNS and authentication). The consumer doesn't control the underlying physical hardware but can manage operating systems, network configurations, and virtual devices. For example, a manufacturing plant might transition its entire server infrastructure to Google's Compute Engine, including web servers, Active Directory, and a database server.

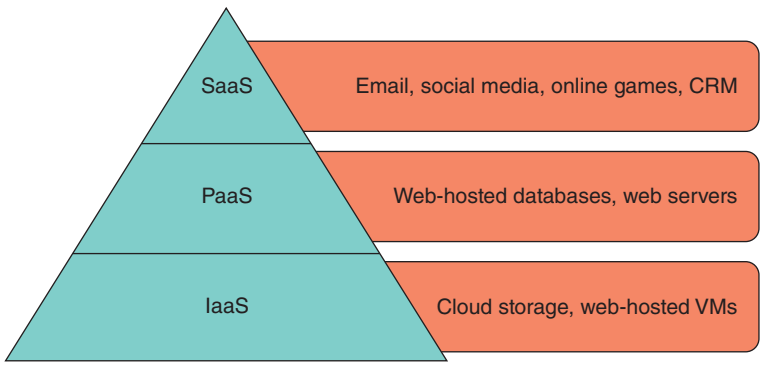
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 1-5 | The "-aaS" acronym can be used to refer to nearly any service offered through the cloud where the consumer does not have to manage the underlying infrastructure. To refer to Anything as a Service, you might use the acronym XaaS, where "X" represents an unknown, just as it does in algebra. (And you thought you would never again use algebra.)

Some other -aaS examples include DBaaS (Database as a Service), DaaS (Desktop as a Service), DRaaS (Disaster Recovery as a Service), CaaS (Communication as a Service or Computing as a Service), BPaaS (Business Process as a Service), and even RaaS (Ransomware as a Service)—because hackers are jumping on the cloud bandwagon, too!

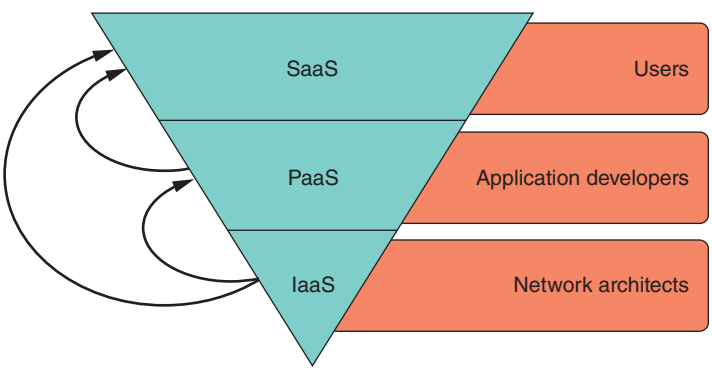
Another way to think about these service models is according to how much a customer must know about configuring and managing cloud networks to use the service, and how much control that customer has over the exact configurations, features, and security measures throughout the technology stack. Consider the service models as they're compared in Figure 1-10. The smaller, upper end of the pyramid indicates how little a SaaS customer needs to understand and interact with a cloud provider's infrastructure to perform their work. In contrast, an IaaS customer interacts more heavily with the service provider's infrastructure for every aspect of their computing needs. The IaaS customer has more control and flexibility with configurations and security features, along with more responsibility for design and security.

Figure 1-10 IaaS customers must understand more about configuring their cloud infrastructure—and they have more control—than do SaaS customers



At the same time, consider how accessible each type of cloud service is to users. In Figure 1-11, the triangle is upside down. Users, the largest group of cloud computing consumers, can easily access and use SaaS products without much setup, whereas IaaS products require extensive preparation by a much smaller group of more skilled network architects and administrators, who provide systems for their users. In the middle of this pyramid is PaaS, which is typically used by application developers—both professionals and laypersons—for developing and testing their products. Customers at the lower layers of this pyramid build products that support customers at the higher layers, such as when a company subscribes to an IaaS product, on which it offers its own PaaS or SaaS products to its own unique market of customers.

Figure 1-11 SaaS services are more immediately accessible to a wide market of users than other categories of cloud services

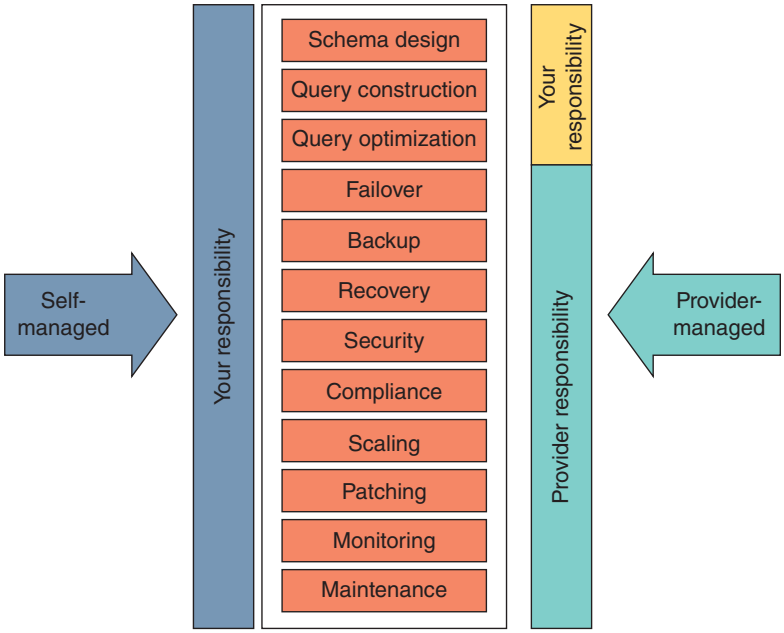


Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Managed Services

There is another way of categorizing services—ask the question, who is in charge of managing the underlying infrastructure that hosts the service components? The answer is not a clearly defined "this" or "that" situation. Instead, the division of responsibility for cloud resource management exists on a spectrum with a variety of available options (Figure 1-12), as described next.

Figure 1-12 Shifting responsibility to the cloud provider



- **Self-managed**—Self-managed cloud services are exactly what they sound like: a cloud service where the customer manages their resources in the service along with the virtual infrastructure hosting it. Typically, the host is a virtual server configured and managed by the cloud customer. For example, you can create a VM instance and install a database such as MySQL on the instance. In terms of cloud service models, this is an IaaS design.
- **Fully managed or provider-managed**—With a cloud-provided managed service, someone else manages the virtual infrastructure hosting the cloud resources. As the customer, you can still make some choices as to what types of servers host the resource, how much capacity and bandwidth these servers have, and any limits you might want to place on the number of servers in the host cluster to control costs or ensure performance. In some cases, such as with AWS's RDS (Relational Database Service), you can even choose serverless hosting. However, the provider manages failover configurations, backup and recovery, security, patching, monitoring, and maintenance of any host servers. In terms of cloud service models, this is still an IaaS design.
- **Serverless**—With a **serverless service**, the customer does not manage the servers at all. There are servers, it's just that the customer does not manage them. In some cases, the customer can choose the type or size of servers in relation to performance and cost, but the customer is not responsible for updating, patching, scaling, or otherwise maintaining the servers hosting the customer's cloud resources. **Scaling** is when the capacity of a resource is increased or decreased. **Auto scaling** is scaling that is done automatically. The host cluster auto scales by adding or removing instances as demand shifts. For very large deployments, this option is often more responsive to large fluctuations in demand. In terms of cloud service models, this is a PaaS design. In AWS, two popular serverless database services are DynamoDB and Amazon Aurora Serverless, and a popular serverless storage service is S3 (Simple Storage Service). Other cloud providers offer similar serverless services, such as Azure SQL Database, Azure Blob storage, and Google Cloud Firestore.

Self-Check

7. Which cloud service model requires the least skill to use?

   a. DRaaS
   b. IaaS
   c. PaaS
   d. SaaS

8. Which cloud service model requires the most skill to adequately secure?

   a. IaaS
   b. PaaS
   c. SaaS
   d. RaaS

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9. Which database deployment option is a PaaS?

a. Serverless
b. Fully managed
c. Self-managed
d. Provider-managed

☐ Check your answers at the end of this chapter.

Section 1-4: Cloud Services

Certification

1.11 Identify evolving technologies in the cloud.

CSPs offer a wide variety of cloud services, some specific to a particular market niche, and others more generalized to meet a wider base of consumer needs. For example, Salesforce hosts a popular SaaS-based CRM system that makes information regarding customer relationships readily available to employees in sales, customer service, marketing, and other organizational departments. Because the data is accessed through an online dashboard, employees throughout the company can analyze extensive and updated customer data to better perform their job duties.

Salesforce focuses primarily on its SaaS CRM products. Although Salesforce also offers a few PaaS products, these are not the focus of their business. Other CSPs have emerged as leaders in the PaaS and IaaS markets, with AWS, Microsoft Azure, and GCP the top three contenders for market share for many years now.

PaaS and IaaS Providers

As you learn about cloud technologies, it's important to follow changes in which providers offer what kinds of services and where they're positioned competitively. The following list explores some of these options:

- **AWS**—The market-leading cloud provider and the first major contender in the public cloud computing market, AWS is a subsidiary of Amazon and was founded in 2006. AWS offers a huge line of cloud computing products with more options and features added frequently. If anything, its plethora of options can be overwhelming and is sometimes considered a weakness. The company offers extensive training tutorials through their website, along with many highly respected AWS-specific certifications.
- **Azure**—Azure is the cloud computing arm of Microsoft's enterprise technologies. Designed for seamless compatibility with Microsoft's other tools, including the widely popular Active Directory, Azure also supports many other operating systems and cloud products. Azure first went live to the public in early 2010.
- **GCP**—Hosted on the same infrastructure as other Google products, GCP brings IaaS and PaaS services to Google's extensive line of cloud products, such as the SaaS-based Workspace collection. App Engine, first released in 2008 and providing basic web application hosting, was Google's first entry into the public cloud market. Soon after, Google began adding other cloud services, which are now collectively referred to as Google Cloud Platform, or GCP.
- **Alibaba Cloud**—Based in Hangzhou, China, Alibaba Cloud is a relative newcomer and yet offers a fast-growing public cloud platform primarily to Asian markets. Alibaba Cloud's parent company, Alibaba Group Holding Ltd., is the largest online commerce company in China and typically ranks in the top five global retailers by revenue.
- **IBM Cloud**—IBM's cloud computing services suite, IBM Cloud, is a rebranding of earlier services from SoftLayer (which IBM acquired in 2013) and IBM's own Bluemix products. IBM Cloud was originally built on the open-source Cloud Foundry, which was originally developed by VMware. With IBM's

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

purchase of Red Hat in 2018, IBM has taken a notable shift toward open-source technology that is intended to target markets across all public cloud platforms.

- **Oracle Cloud**—Based in Redwood Shores, CA, Oracle is known primarily for its database software. In 2010, Oracle began rebranding its existing Oracle On Demand services, which allowed for running Oracle applications in an Internet browser. They also added new cloud computing services, all under the Oracle Cloud umbrella. The slower-growing Oracle Cloud platform takes advantage of Oracle's extensive success with databases and other software products to differentiate itself from other CSPs for Saas, PaaS, and IaaS services.

All these CSPs primarily provide public cloud platforms. You can also host your own private cloud—essentially DIY (do-it-yourself) cloud computing—using cloud management software of some kind. A few examples include the following:

- **OpenStack**—One of the most popular private cloud platforms, OpenStack is an open-source cloud computing platform designed by Rackspace and NASA to create DIY compute, storage, and networking cloud services. OpenStack can be deployed on the company's own data center, through an OpenStack-based public cloud, or in a hosted OpenStack private cloud.
- **CloudStack**—An open-source platform from Apache, CloudStack supports large deployments of VMs. With CloudStack, an organization can deploy a private cloud on their own hardware and hypervisor (such as VMware or XenServer), offering their users similar features to what the public cloud offers: compute on demand, virtualized networking configuration, account management, and an easy-to-use interface.
- **OpenShift**—Whereas CloudStack and OpenStack support IaaS private clouds, OpenShift by Red Hat works more at the PaaS level to support and automate application development in hybrid cloud deployments. OpenShift is used to develop and deliver applications throughout the life cycle in underlying containers. It can be deployed in a private data center or on top of public cloud infrastructure, such as AWS.
- **VMware**—A leading provider of data center virtualization solutions such as vSphere and ESXi, VMware is also a respectable force in the cloud market. VMware Cloud allows an organization to create a private cloud in its own data center. Alternatively, VMware solutions such as vSphere and vCenter can be installed over an existing public cloud infrastructure. Because so many organizations already use VMware virtualization products, VMware's partnerships with some of the major public cloud providers offer a natural extension to a customer's existing VMware infrastructure to create a seamless hybrid cloud or multi-cloud. VMware was acquired by Broadcom in November 2023, but it still operates under the VMware name at the time of this writing.
- **Eucalyptus**—Eucalyptus (Elastic Utility Computing Architecture for Linking Your Programs To Useful Systems—yes, it's a long acronym) is a partially open-source platform that is designed specifically for building private or hybrid clouds to interact with AWS products. Originally developed by researchers at the University of California, Eucalyptus became a for-profit business in 2009, was purchased by Hewlett-Packard in 2014, and is now managed by AppScale Systems.

Exam Tip | Although the CompTIA Cloud+ exam is vendor neutral, you do need to be familiar with major CSP platforms and understand the basic ways they manage cloud services. This background knowledge will give you a more concrete understanding of the relatively abstract cloud computing concepts covered by the Cloud+ exam.

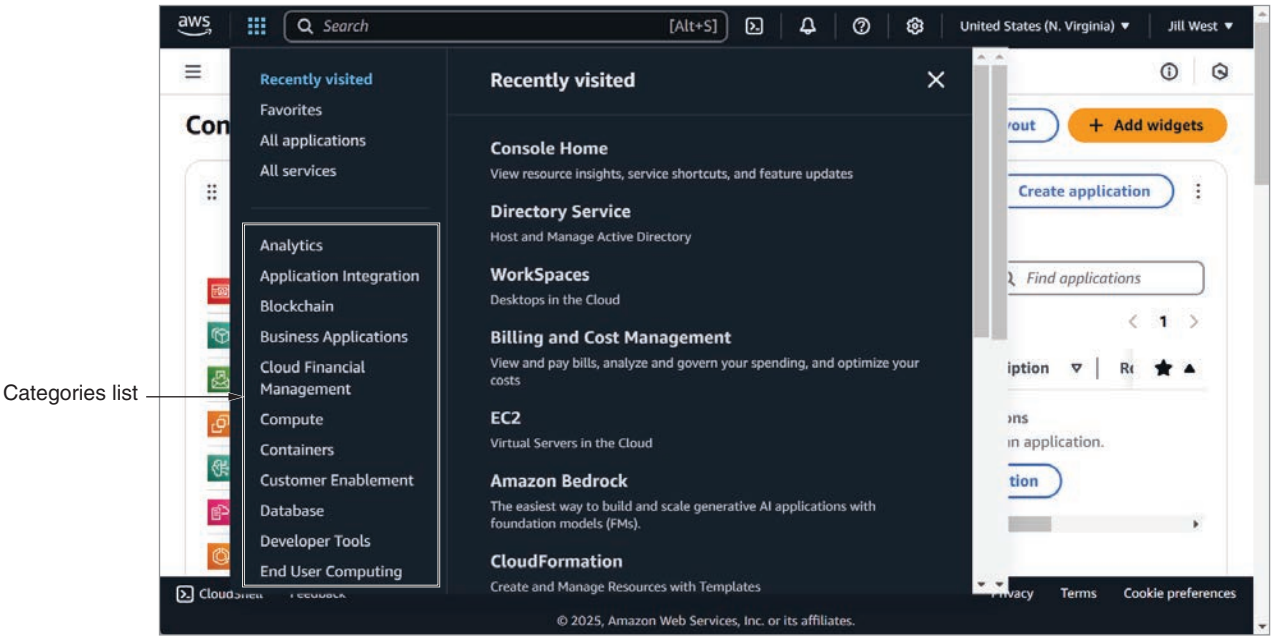
Common Cloud Services

While each CSP offers different cloud services and configures their user interfaces and product offerings in different ways, the major CSPs offer the same basic product types. These are the services you might use to configure your own cloud network, to host your resources on the cloud (such as applications, servers, or data), or to run the processes you need (such as authentication, machine learning, or batch processing). You'll

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

encounter many service categories in various cloud computing platforms, such as AWS's service categories displayed in Figure 1-13. The larger CSPs like AWS use more specific categories to make it easier for their customers to choose the products they need. These categories and the products in them change frequently as cloud technologies continue to mature.

Figure 1-13 AWS service categories



Source: Amazon Web Services, Inc.

The following list describes some of these major categories of cloud services:

- **Compute**—Compute refers to cloud functions that process data in some way. The term primarily refers to VMs, code, or applications that run in the cloud and can also refer to developer tools used to create, test, and deploy applications. When working with compute services, you'll find many options for configuring the virtualized hardware on which these processes run—such as CPU cores, memory size, and I/O (input/output) compatibilities. You'll also find newer, cloud-native technologies for running code in the cloud, such as containers and serverless services. Grouped under the compute category, you might also find tools related to these processes, even though these tools might technically perform functions from other categories. For example, you might find a repository of OS images (a storage function) used to create VMs.
- **Storage**—The storage category covers cloud services that store or preserve data. This is a large category that is often subdivided into smaller categories, including databases (which is arguably a completely separate category from storage), block storage, object storage, file storage, data optimization services such as DFR (data footprint reduction), and data protection services such as backup and disaster recovery. You'll find virtualized hardware options, such as choosing storage volumes of various speeds and storage capacities, RAID (redundant array of independent disks) options, data transfer speeds, and storage protocols (such as Ethernet, iSCSI, or InfiniBand, which bleeds over into networking concepts), as well as pricing options for accessing stored data. You might also find tools grouped under the storage category that enable mass migration of large amounts of data.
- **Networking**—Networking functions provide ways of moving data within the cloud or between networks and resources. Included in this category are options for configuring network functions such as connection configurations, subnets, IP addressing, VPNs (virtual private networks), trunk lines, DHCP (Dynamic Host Configuration Protocol) services, load balancing, and more. Other tools include migration resources and content delivery products, which might be grouped in their own categories.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Security**—As with a traditional data center, security functions in the cloud include perimeter security (firewalls, intrusion detection and prevention, and proxies for the virtual network), identity management, certificate or key management, authentication services, access controls, and token or federation services. Cloud services, by definition, present a larger attack surface than do traditional networks, where access is often handled primarily at the perimeter. Therefore, the cloud security tools you'll typically find at your disposal, especially when configuring a public cloud, will take the needs of this abstracted, remote environment into account.
- **Application components**—Tools from many of these categories can be used for building and hosting cloud-based applications. Cloud services optimized for app development might also be available, such as testing or hosting environments, microservices, templates, security auditing, e-commerce, machine learning, and mobile services.
- **Management tools**—Cloud platforms include a variety of management and monitoring tools, including dashboards (such as OpenStack's Horizon or Amazon's QuickSight), monitoring services (such as Google's Operations or Amazon's CloudWatch), and data analytics services (such as Azure's Stream Analytics or Google's BigQuery).

Table 1-2 shows some products in three of these categories from some of the major cloud platforms. You'll learn more about each of these products in later chapters. This table will help you begin to get familiar with the kinds of service names you'll encounter in this course.

Table 1-2 Popular cloud products and features from major cloud platforms

CSP	Compute	Network	Storage
AWS	EC2 (VMs) ECS (containers) Lambda (serverless functions)	Virtual Private Cloud (networking) Direct Connect (dedicated cloud connection) Route 53 (DNS)	EBS (block storage) Glacier (archive storage) Snowball (petabyte-scale data transport)
Azure	Virtual Machines (VMs) Container Instances (containers) Azure Kubernetes Service (container orchestration)	Azure DNS (DNS) VPN Gateway (cross-premises connectivity) ExpressRoute (dedicated cloud connection) Traffic Manager (traffic routing)	Blob Storage (object storage) Files (file shares)
GCP	Compute Engine (VMs) App Engine (PaaS) Kubernetes Engine (containers)	Virtual Private Cloud (networking) Cloud Load Balancing (scalable load balancing) Cloud Interconnect (dedicated cloud connection) Cloud DNS (DNS)	Cloud Storage (object storage) Persistent Disk (block storage) Filestore (file storage)

Note 1-6 | In this course, you'll need access to at least one major CSP's public cloud (for example, AWS, Azure, or GCP). You might want to sign up with more than one of these CSPs. The projects in this chapter show you the sign-up options for AWS, Azure, and GCP so you can choose the account type that best fits your needs and situation. Many students are concerned about having to provide a credit card number to create a cloud account. For this reason, some CSPs offer student-focused free trials that do not require a credit card.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Projects throughout the course that use these public clouds give instructions for AWS first. Additional projects provide instructions for performing the same or similar tasks in Azure and GCP. You can also watch author-produced videos for projects in AWS or Azure , which are available in MindTap. The idea is to give you lots of hands-on experience with at least one major CSP's public cloud and widespread exposure to other platforms.

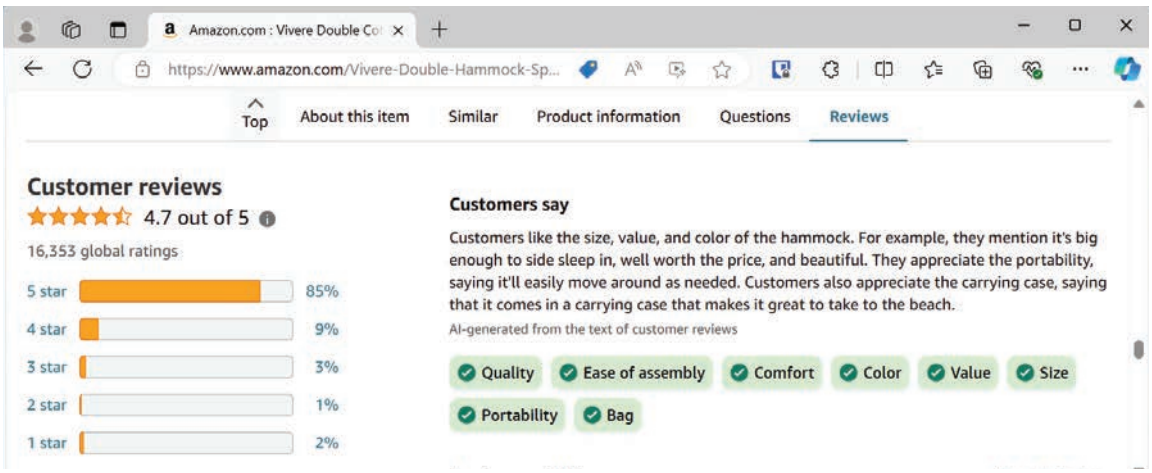
AI/ML (Artificial Intelligence/Machine Learning) Services

AI (artificial intelligence) is the ability of computers to adapt to changing circumstances. AI is a broad set of technologies that support the ability of machines to emulate human intelligence in some ways, create content, interact with the world, collect and analyze data, and solve problems. One subset of AI is **ML (machine learning)**, which is the ability of computers to learn from data analysis rather than only from specific programming.

Many popular cloud platforms offer a variety of AI/ML services that can be incorporated into your cloud-hosted systems, websites, and applications. The CompTIA Cloud+ certification objectives list several AI/ML services that are mostly related to written or spoken language. Consider the following AI/ML service types:

- **Text recognition** services, also known as OCR (optical character recognition) services, can interpret text in documents, prompts, transcripts, images, and even handwriting. Text recognition makes this text searchable, and it can analyze data contained in the text. For example, AWS's Textract can be used to process data from government forms, such as loan applications or federal tax forms. Once the data is extracted, that information can be used for a variety of purposes, such as performing data analytics or providing a Q&A (question and answer) service that helps employees locate archived information.
- **Text translation**, as you might expect, translates text from one language to another. Older translation services relied on algorithmic rules for language translation, often resulting in a loss of intent or nuance. Newer text translation services, such as AWS's Translate, rely on more advanced ML technologies that consider the context of each full sentence to increase accuracy and fluency.
- **Sentiment analysis** evaluates text to extract information about the emotional tone of the text. For example, sentiment analysis can analyze a transcript from a customer service call to determine if the customer was frustrated, angry, satisfied, excited, and more. Sentiment analysis can be helpful for other scenarios as well, such as transcripts from chats, social media comments, or product reviews (Figure 1-14).

Figure 1-14 AI-powered summary of customer reviews

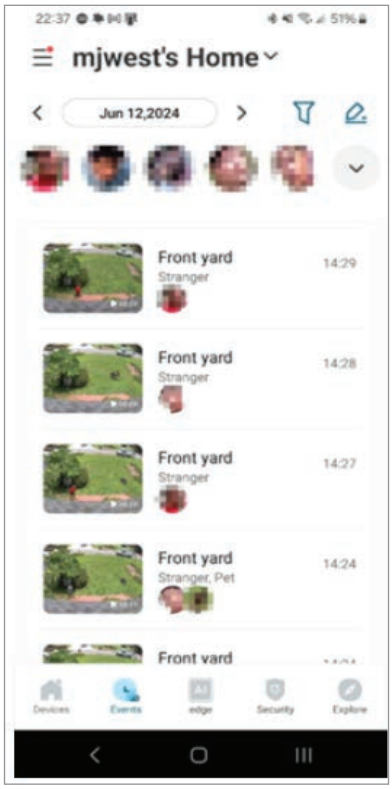


Source: Amazon.com, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Voice-to-text** interprets speech into text. This is the process that allows closed captioning to be generated in real time or that lets you generate text by speaking, such as when dictating a text message while you drive. In AWS, this process is provided by the Transcribe service. Powerful and automated services such as Transcribe can be used to analyze conversations, videos, or phone calls. For example, Transcribe can generate a transcript of meetings for later review, convert doctors' verbal notes to text for each patient's medical record, or detect toxicity in a gaming conversation to ensure a safe playing environment.
- The reverse process is **text-to-voice**, which converts text to human-like speech. AI-powered text-to-voice systems, such as AWS's Polly, use deep learning technologies to naturalize generated voices in a variety of languages. The speaking style can be customized for culturally appropriate characteristics, such as rate, pitch, phrasing, and intonation—you can even add breathing sounds to increase its authenticity.
- **Visual recognition** involves analyzing high volumes of images and videos (even streaming video in real time, such as from traffic cameras or cameras on a smart car) to identify desired information. For example, AWS's Rekognition service can determine the identity of a person in a video or photo if an image of their face is already identified in another photo and is available to the system. This can be helpful for scenarios such as security cameras where you don't need to be alerted when a known family member approaches the camera, but you do want a notification if the person approaching is a stranger (Figure 1-15). Additionally, Rekognition can be trained to label images and activities in visual media—including objects such as a bike, mountains, or a sunset, and activities such as smiling, running, or driving. Visual recognition technology can even be used to diagnose areas of concern in X-rays, MRI (magnetic resonance imaging) scans, or CT (computed tomography) scans.

Figure 1-15 Security camera captures people not identified as family or friends



Source: Fantasia Trading LLC

Several other types of AI/ML services exist, such as Lex in AWS, which provides a natural language model to support human interactions with background systems, or Amazon Kendra, which can search stored data across a wide variety of data types to find requested information. As you expand into various cloud platforms and vendor-specific certifications, you'll need to learn more about the AI/ML services offered by each cloud platform.

Exam Tip | Because the CompTIA Cloud+ exam is a vendor-neutral exam, you likely will not be required to know vendor-specific service names for any of these AI/ML service types. However, you will need to be prepared to identify each AI/ML service type when given a description or a specific need.

Generative AI in the Cloud

The emergence in November 2022 of ChatGPT and similar tools drew significant public interest in AI. This process also greatly increased the impact of AI technologies on how applications are designed, how the cloud is used to support these technologies, and even how work is performed in many roles across nearly every industry.

What made ChatGPT such a significant development? ChatGPT made public and widely available **GenAI (generative AI)**, which is any of several technologies used to generate content (such as text or images) based on training from one or more datasets.

While you might have used some of these GenAI tools to draft an email, outline a paper, research a project, or create an image or video, you also need to understand how GenAI is used in the cloud, and how

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

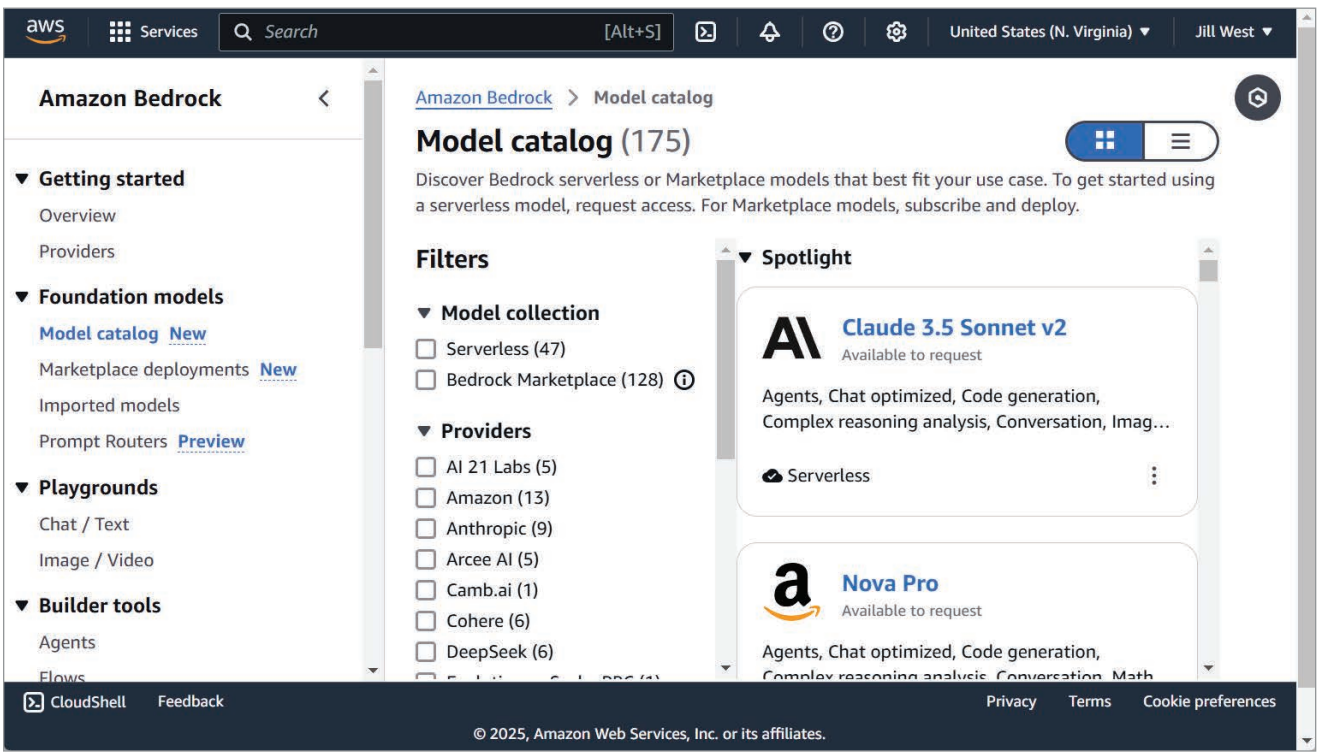
GenAI technologies are incorporated into backend systems to add features to applications or websites. For example, what services in AWS and other cloud platforms are powered by GenAI? What services are used to incorporate GenAI into your own applications or websites? What resource configurations are needed in the cloud to support AI on your applications or websites?

For GenAI to work, it needs a trained dataset, called an **FM (foundational model)**, as an information source. Some organizations have developed FMs that can be used by other organizations, which reduces the need to train new models. Many of these FMs, called LLMs (large language models), are trained on a massive, generalized dataset, which makes them widely adaptable to many industries or needs. Other models, called SLMs (small language models), are more specific and curated in their training and source data, which can increase their accuracy while reducing the time needed to train them and the infrastructure resources required to run them. While LMs are trained on text and can only generate text, other types of FMs can be trained to generate images, speech, other audio (such as music), video, and other kinds of outputs.

To give you a basic idea of the types of GenAI services you'll encounter in the cloud, the following list describes a few of the most popular GenAI services in AWS:

- Bedrock (Figure 1-16) provides single-point API access to a variety of FMs, which makes it easier for a developer to change FMs on the backend without affecting frontend design or programming.
- PartyRock is a sandbox environment where you can experiment with building AI-generated applications using the FMs available through Bedrock.
- SageMaker can be used to develop, train, and deploy your own ML models.
- Trainium offers VM configurations that are optimized for ML training.
- Amazon Q (Figure 1-17) is an AI-powered assistant you can use within the AWS Management Console or embed in your own applications.

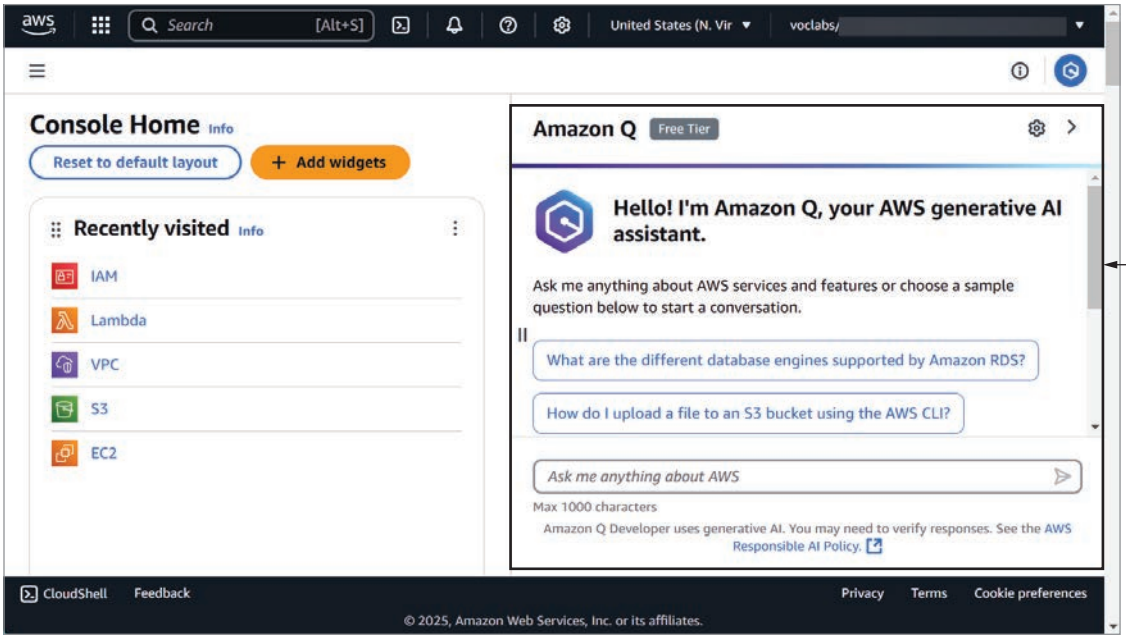
Figure 1-16 Choose from several FMs in Bedrock



Source: Amazon.com, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 1-17 Amazon Q in the AWS Management Console



Source: Amazon.com, Inc.

Applying Concepts 1-1: Generative AI on the Cloud

Estimated time: 15 minutes
Objective 1.11: Identify evolving technologies in the cloud.
Resources:

- Internet access
- If on a smartphone, a generative AI app, such as ChatGPT

Description: The emergence of ChatGPT and similar tools in the past few years largely has been enabled by cloud infrastructure. Not only is cloud capacity capable of expanding almost indefinitely to accommodate skyrocketing rates of GenAI use, it also provides low-latency availability to global markets and adaptive pricing structures for startup companies. In this activity, you will research the cloud platforms supporting your favorite AI apps.
Complete the following steps:

1. What AI tools do you use most often? List GenAI chat apps, such as ChatGPT or Microsoft Copilot. Also list AI tools incorporated into other platforms you use, such as AI features in your favorite music streaming service.
2. Use a search engine, such as Google or DuckDuckGo, to research which cloud platform three of these tools are hosted on and what kinds of services the AI tools use to function. Be sure to check actual links provided in the search, not just the AI summary at the top of the search.
3. Use a GenAI chat tool, such as ChatGPT or Perplexity, to research which cloud platform and cloud services these same three tools use. Compare the information to what you found when using the search engine.
4. Which of these research techniques gave you the most helpful information?
5. What kinds of references were provided to verify the information given to you?

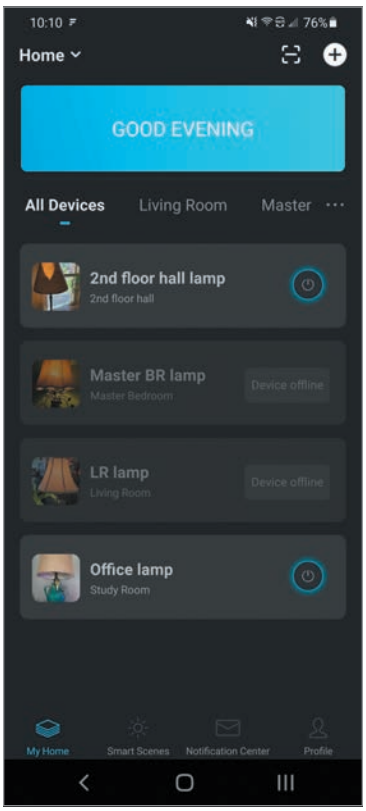
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

IoT (Internet of Things)

The **IoT (Internet of Things)** is a collection of all devices connected to the Internet. These aren't only your normal computing devices, like laptops, smartphones, and printers; IoT includes anything that can communicate with the Internet. All sorts of things can be connected to a network, from toasters, refrigerators, bathroom scales, and garage doors to watches, lamps, cars, thermostats, doorbells, and even the kitchen sink.

One of the fastest-growing areas of IoT is personal monitoring devices, such as health monitors, exercise equipment, GPS locators, and smartwatches. Another exploding IoT market interconnects smart home devices. You might already be familiar with Amazon Echo, Apple HomePod, or Google Home. These voice-controlled smart speakers and their embedded personal assistant apps (such as Alexa, Siri, and Google Assistant) can interlink a plethora of devices, from locks and lights to security cameras and coffee pots. You can control these devices through voice commands while interacting with a smart speaker, or you can control IoT devices through an app on your smartphone (Figure 1-18). All these connected devices within a home create a type of network called a HAN (home area network).

Figure 1-18 App-controlled IoT devices



Source: Merkury Innovations

IoT additionally includes technology with a much greater impact on industries and businesses. Consider manufacturing plants that incorporate sensors and actuators, which control machinery to ensure optimized production, improved efficiency, and predictive maintenance. A **sensor** is a device that can detect and monitor metrics related to its environment, such as temperature, light, motion, humidity (such as air humidity or soil moisture), or pressure (such as air pressure or water pressure). Sensors convert the detected inputs into digital data that can be communicated, stored, and monitored.

An **actuator** is a device that converts digital commands into physical actions. Where sensors are the "eyes and ears" of an IoT system, actuators are its "hands and feet." Actuators use electric motors, air or hydraulic pressure, fluid flow, or magnetic energy to create movement. They can be used to open a door, turn on a light switch, or operate a robotic arm. They can also apply oil to a machine that is vibrating when it shouldn't, trigger an emergency shut-off valve when water levels rise too high, activate water pumps when crops start to get too dry, or release precise amounts of insulin to optimize a patient's blood sugar level.

While IoT devices can function without cloud technology, most of these devices rely on the cloud to optimize their functionality in two ways, as follows:

- **Communication**—IoT devices generally communicate over the Internet with a cloud service of some kind, such as a SaaS, that allows the user to control their devices, especially when using a personal assistant app like Alexa.
- **Data**—IoT generates massive amounts of data, which is often stored, monitored, and sometimes processed in the cloud. This is beneficial because you only pay for the storage space you use, and you can be confident the cloud's storage space can expand to accommodate whatever amount of storage you're willing to pay for.

Self-Check

10. Which public cloud platform offers the most streamlined integration with Windows Active Directory?

a. AWS
b. Alibaba
c. Azure
d. Oracle

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11. In which cloud category would you most likely find a service for configuring DNS records?

a. Networking
b. Security
c. Compute
d. Storage

12. Which IoT device provides access to a personal assistant app?

a. GPS locator
b. Smart speaker
c. Security camera
d. Smart toaster

☐ Check your answers at the end of this chapter.

Section 1-5: Troubleshooting Methodology

Certification

6.2 Given a scenario, troubleshoot network issues.

Throughout this course—and throughout your career—you'll be required to troubleshoot various problems. Taking a methodical approach to this process helps ensure efficiency and prevent costly delays and mistakes. As you become more familiar with the networks and other resources that you manage, you'll likely begin to rely more heavily on your experience and intuition. In the meantime, however, a troubleshooting method can help you organize your thoughts and your approach to solving problems. And when tackling an especially stubborn problem, you can fall back on this established method to help you work through the process one logical step at a time.

Exam Tip | The Cloud+ exam might give you a troubleshooting scenario that requires you to identify the *next* step according to a reasonable and logical troubleshooting method. To do this, you must use the information given to identify which steps of the troubleshooting process have been completed so far in the scenario and then identify which step should come next to give you relevant information for your next decision and which of the task options would accomplish that step.

Common Cloud Computing Problems

At this point in your career, you've probably already developed a sense of how to troubleshoot a computer or network problem. In this course, you'll learn more about how troubleshooting processes apply specifically to cloud computing. You'll first explore a few common issues you might experience when working in cloud environments. When the troubleshooting steps are discussed later in this section, you'll explore how those steps can be applied to common cloud issues. Consider some of these common issues:

- **Connectivity issues**—Using public cloud services relies heavily on access to the Internet. If your connection to your ISP (Internet service provider) goes down, you won't have access to your cloud services. Problems could also originate from your local network hardware or from endpoint configurations. Or you might have dedicated connections to a public exchange where connectivity can suffer from a routing misconfiguration. You'll need to carefully investigate where the connectivity is failing or suffering compromised performance to determine who is responsible for addressing the problem.
- **Latency**—You might have connectivity, but it's really slow. This delay is called **latency** and is often caused by poor routing, insufficient bandwidth, or a bottleneck. Latency might be the ISP's responsibility, or you might need to upgrade your service with either your ISP or your CSP.
- **Capacity**—While you can certainly face capacity issues with a traditional data center, cloud services can be designed to auto-scale in order to avoid capacity-related problems. If this auto-scaling is not adjusting to current demands, you might need to adjust the related configurations or check for security issues that might be causing unexpected problems.
- **Security**—Cloud computing brings its own set of security vulnerabilities. When troubleshooting cloud services, you'll need to become familiar with the unique indicators of cloud-based security issues.

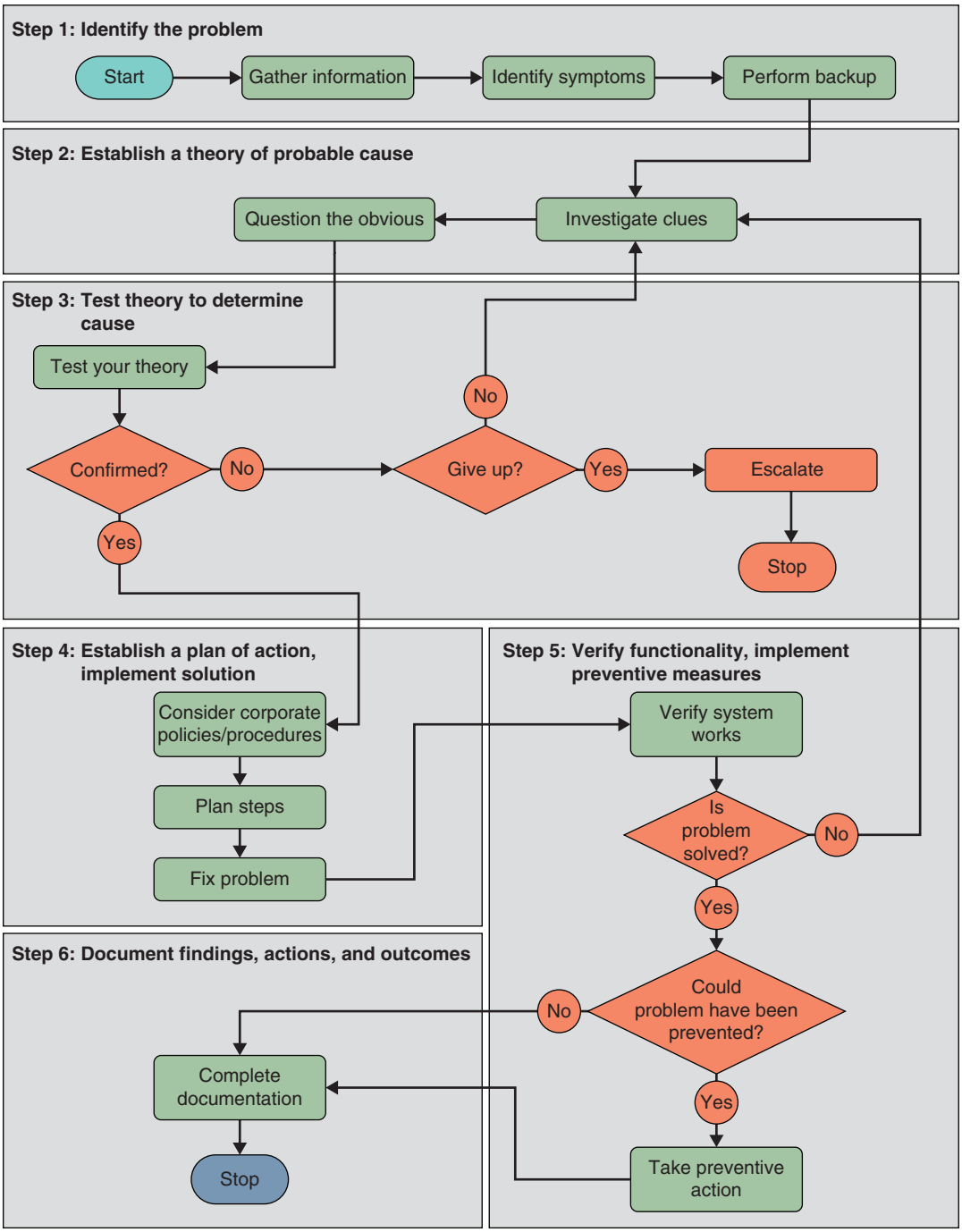
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

You'll learn more about each of these problems and how to solve them later in this course. In the meantime, the following troubleshooting steps will help you develop a deeper understanding of how various components and services can affect each other in a system.

Troubleshooting Steps

The following is a list of the steps in the CompTIA troubleshooting methodology, which are also diagrammed in Figure 1-19. Although the steps cannot be followed rigidly without evaluating how best to approach a specific problem, they can provide guidance when you need to handle a complex or perplexing problem. As you read through the steps, think about how you've used these steps yourself in troubleshooting problems you've encountered:

Figure 1-19 General approach to solving cloud problems



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1. **Identify the problem**. This step includes questioning the user and identifying any changes the user has made recently to the computer or changes that have been made to the network or environment. Also, back up data on the computer or other system before making any changes to it yourself.
2. **Establish a theory of probable cause**. Beginning with the symptoms, search for internal and external clues for what might be causing the problem. Remember that the cause is not always complex—question the obvious!
3. **Test the theory to determine cause**. After you've confirmed your theory, decide what steps you should take next. If you can't confirm your theory, try again or escalate the problem to the next support tier.
4. **Establish a plan of action to resolve the problem and implement the solution**. However, don't make any changes without first considering corporate policies and procedures and the impact the change might have on other systems. Keep in mind that some changes must first be approved at higher levels.
5. **Verify full system functionality, and, if applicable, implement preventive measures**. This might include educating the user.
6. **Document findings, actions, and outcomes**. Most organizations keep a **knowledge base** of some kind for systematic documentation of information, configurations, troubleshooting scenarios, and helpful tips relevant to a specific network or service. Be sure to use professional communication skills, such as good grammar and thorough explanations, so your documentation will be easy for others to understand later.

Exam Tip | Although this particular troubleshooting methodology describes a logical series of steps to evaluate a problem, not all problems can be evaluated in the same way. You'll need to use your intuition and professional expertise to adapt your approach to troubleshooting a specific problem. Even on the CompTIA Cloud+ exam, you might be given a scenario where the classic approach to troubleshooting is not a good fit, requiring a customized and creative approach to the problem.

Preventive Measures

You can take several measures to help minimize the likelihood of problems cropping up in your cloud environment. Consider the following tips:

- Maintain good monitoring and analysis techniques.
- Thoughtfully configure your dashboards to show the most helpful information.
- Follow effective change management processes.
- Understand your cloud services and how to identify where problems are likely to occur, how to locate those problems, and how best to address them.

These best practices will be covered in more detail throughout this course.

Self-Check

13. What cloud configuration technique can be used to compensate for capacity challenges?

a. Latency
b. Endpoint configurations
c. Change management
d. Auto-scaling

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14. Most often, what will the last step be when troubleshooting a problem?

a. Documentation
b. Verify functionality
c. Test the theory
d. Implement the solution

☐ Check your answers at the end of this chapter.

Summary

Section 1-1: Characteristics of Cloud Computing

- The CompTIA Cloud+ certification is a vendor-neutral, entry-level cloud computing certification that requires demonstration of technical skills.
- Popular cloud vendor certifications are published by CSPs such as Amazon, Microsoft, and Google.
- Cloud computing consists of five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.
- Skills most needed by cloud professionals include cybersecurity skills, cross-discipline expertise, automation scripting skills, and lifelong learning.

Section 1-2: Cloud Deployment Models

- Public cloud services are hosted on hardware resources at the CSP's location.
- Private cloud services are hosted on hardware resources used exclusively by a single organization.
- A hybrid cloud is a mix of both public and private cloud components, or a combination of some cloud and some traditional on-prem services, where those components and services interact with each other in a direct and seamless manner.
- A multi-cloud model involves using cloud services from multiple vendors at one time.
- A community cloud is accessible to multiple organizations with similar concerns, but not to the general public.

Section 1-3: Cloud Service Models

- The usual classifications of cloud service models, as defined by NIST, are SaaS, PaaS, and IaaS.
- The division of responsibility for cloud resource management exists on a spectrum: self-managed, fully managed, and serverless.

Section 1-4: Cloud Services

- Market-leading IaaS and PaaS CSPs include AWS, Azure, and GCP. Other popular CSPs include Alibaba Cloud, IBM Cloud, Oracle Cloud, OpenStack, VMware, and Eucalyptus.
- Cloud services are grouped into categories, such as compute, storage, networking, security, application components, and management tools.
- Popular AI/ML services hosted on many cloud platforms include text recognition, text translation, sentiment analysis, voice-to-text, text-to-voice, and visual recognition.
- The IoT is a collection of all devices connected to the Internet.

Section 1-5: Troubleshooting Methodology

- Common problems encountered with cloud technologies include connectivity issues, latency, capacity, and security.
- The steps for CompTIA's troubleshooting methodology are as follows:
  1. Identify the problem.
  2. Establish a theory of probable cause.
  3. Test the theory to determine cause.
  4. Establish a plan of action to resolve the problem and implement the solution.
  5. Verify full system functionality, and, if applicable, implement preventive measures.
  6. Document findings, actions, and outcomes.
- Preventive measures include good monitoring practices, thoughtful dashboard design, effective change management, and a clear understanding of a company's cloud infrastructure.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Key Terms

For definitions of key terms, check the Glossary.

actuator
AI (artificial intelligence)
API (application programming interface)
auto scaling
automation
AWS (Amazon Web Services)
Azure
broad network access
cloud bursting
community cloud
CRM (customer relationship management)
CSA (Cloud Security Alliance)
CSP (cloud service provider)
FM (foundational model)
GCP (Google Cloud Platform)
GenAI (generative AI)
hybrid cloud
IaaS (Infrastructure as a Service)
IoT (Internet of Things)
knowledge base
latency
measured service
ML (machine learning)
multi-cloud
multitenancy
NIST (National Institute of Standards and Technology)
on-demand self-service
orchestration
PaaS (Platform as a Service)
private cloud
public cloud
rapid elasticity
resource pooling
SaaS (Software as a Service)
scaling
sensor
sentiment analysis
serverless service
SLA (service level agreement)
tenant
text recognition
text translation
text-to-voice
virtualization
visual recognition
voice-to-text

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

AI artificial intelligence
API application programming interface
BYOD bring your own device
CRM customer relationship management
CSA Cloud Security Alliance
CSP cloud service provider
DBaaS Database as a Service
IaaS Infrastructure as a Service
ISP Internet service provider
ML machine learning
PaaS Platform as a Service
SaaS Software as a Service
VM virtual machine
VPN virtual private network

Review Questions

1. Which of the following technologies most directly supports and enables multi-cloud tenancy?
   a. SLA
   b. API
   c. CRM
   d. GenAI

2. Who is responsible for ensuring the security of cloud-hosted resources? **(Select two.)**
   a. The CSP
   b. The ISP
   c. The cloud customer
   d. The government

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. When you speak to Alexa, you initiate processing of code that runs in the cloud without creating a new server to do so. Which cloud service model most likely provides the service that processes this code?

   a. SaaS
   b. PaaS
   c. IaaS
   d. DBaaS

4. Which type of service is most likely to have auto-scaling embedded in the design of the service?

   a. Fully managed service
   b. Self-managed service
   c. Serverless service
   d. Provider-managed service

5. To deploy a series of test environments in the cloud, you want a script that will design and configure several servers, some network infrastructure, and firewall rules. What process will produce the script you need?

   a. Automation
   b. Self-healing infrastructure
   c. Multitenancy
   d. Orchestration

6. Which document can ensure you will receive compensation if your company's Internet service experiences an extended outage?

   a. CI/CD
   b. Script
   c. SLA
   d. Dashboard

7. Multitenancy enables which cloud characteristic?

   a. Rapid elasticity
   b. Broad network access
   c. On-demand self-service
   d. Resource pooling

8. Which service type would best be used to determine what percentage of traffic captured by security cameras consists of commercial trucks?

   a. Sentiment analysis
   b. Text recognition
   c. Generative AI
   d. Visual recognition

9. For what purposes does the IoT often use cloud technologies? **(Select two.)**

   a. Virtualization
   b. Storage
   c. Compute
   d. Communication

10. During which step of the troubleshooting process should you back up data?

    a. While testing your theory as to the probable cause
    b. While identifying the problem
    c. While verifying full system functionality
    d. While establishing a plan of action

Scenario-Based Questions

Question 1-1

Your company transitions a dozen servers to the cloud instead of purchasing several new servers to replace aging equipment. The monthly cost of these cloud-hosted servers that can grow with the business is significantly lower than the anticipated cost of purchasing new hardware that would have, at first, greatly exceeded the current needs of the business. What feature of the cloud has your company taken advantage of?

a. Broad network access
b. Measured service
c. Multitenancy
d. Cross-platform compatibility

Question 1-2

As you're developing your hybrid cloud, you move a handful of servers and two applications to the cloud. Both applications need to communicate with a database that will remain on-prem. What technology will allow the applications to communicate with the database without human intervention?

a. IoT
b. NIST
c. OCR
d. API

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Question 1-3

Raul's company is preparing for a cloud migration and has narrowed its options to a small handful of CSPs. All these CSPs offer a wide variety of cloud services and meet all the company's security requirements. Cheri, Raul's boss, is now asking him to recommend what type of cloud services would best fit their needs. She explains they want to start with cloud-based email and an office productivity suite that enables easy file sharing among team members. Which cloud service model should Raul recommend they use to accomplish these goals?

a. SaaS
b. IaaS
c. RaaS
d. PaaS

Question 1-4

Anika has run into a problem with the VPN connection to her company's cloud-hosted database. She has a good Internet connection and can access her work email account, which uses the same sign-in credentials as the VPN. But she still can't pull reports from the database service. She calls Patrick, who works for the company's help desk. As Anika describes the problem to Patrick, he tells her to try again and make sure she's using the right password, then restart her router, and if that still doesn't fix it, to restart her computer. Which troubleshooting step did Patrick fail to complete, which will likely result in wasted time and increased frustration for both Anika and Patrick?

a. Test the theory to determine cause.
b. Document findings, actions, and outcomes.
c. Establish a plan of action and implement the solution.
d. Establish a theory of probable cause.

Hands-On Projects

Note 1-7 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Project 1-1: Research Cloud Certifications

Estimated time: 30 minutes (+20 minutes for group work, if assigned)
Objective 1.11: Identify evolving technologies in the cloud.
Group work: This project includes enhancements when assigned as a group project.
Resource:

- Internet access

Description: This course prepares you to take the CompTIA Cloud+ CV0-004 exam, which is a vendor-neutral exam covering a broad array of foundational and intermediate cloud computing concepts within the context of an organization's entire IT system. Other cloud certifications are offered mostly by cloud vendors, including Amazon, Google, VMware, OpenStack, Cisco, Microsoft, and IBM. Other vendor-neutral certifications are offered by Cloud Security Alliance (CSA) and Cloud Credential Council (CCC).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Passing the CompTIA Cloud+ exam will help prepare you to target more specialized and advanced cloud certifications later, such as those published by cloud vendors. Understanding what those certifications cover, how they're organized, and how they're related will help you make better choices about which certifications are the best fit for you and your career aspirations.

Use the web to research and answer the following questions:

1. What are the six domains of objectives for the CompTIA Cloud+ CV0-004 exam?
2. What other certifications (or equivalent knowledge) does CompTIA recommend you have before taking the Cloud+ exam?
3. What is the primary difference in purpose between the CompTIA Cloud+ exam and the CompTIA Cloud Essentials+ exam?
4. What are the domains of the AWS Cloud Practitioner certification?
5. What kind of experience does AWS recommend you have before taking the Cloud Practitioner exam?
6. What skills does the Microsoft Azure Fundamentals exam cover?
7. What are the six sections of the Cloud Digital Leader certification by Google?
8. **For group projects:** Each member of the group should research online for practice questions for a single cloud certification, either one discussed in this project or a different certification. Group members should compare their sources to ensure no one uses the same source as someone else. Each group member quizzes the other group members with five to ten practice questions and tallies their performance. Each group member then lists which exam objectives the group collectively demonstrates sufficient knowledge as required by each objective covered by that exam. Submit the name of the exam, the sources of the practice questions, a brief summary of the group's performance on the practice questions, and the list of exam objectives currently mastered by the group collectively.

Project 1-2: Create an Account in AWS

Estimated time: 30 minutes
Objective 4.3: Given a scenario, implement identity and access management.
Resources:

- Internet access
- If creating a standard AWS account, a credit card
- If creating an AWS Academy account, an invitation from your instructor to join a Learner Lab in AWS Academy

Description: AWS offers multiple options for creating and accessing free resources for various purposes. The account itself is always free. While a typical AWS account requires a credit card (even if you don't accrue charges), most readers do have another option. Specifically, AWS Academy does not require a credit card and can be managed by your instructor. The following are the two types of accounts you can use for this course:

- **AWS account**—This is a standard account type open to the public. It requires a credit card and allows full access to standard AWS services, including free-tier services. So long as you carefully follow the project steps and reliably shut down and delete the cloud resources you create for your projects, you should be able to complete all the required portions of the AWS projects in this course without incurring any charges on your credit card.
- **AWS Academy account**—This account type is available to educators and students from accredited educational institutions, and it does not require a credit card. Instructors can customize the AWS environment students use, and instructors can enter each student's AWS console to help with troubleshooting or grading work. AWS Academy offers many appealing benefits, including free credits allotted to each course a student takes in AWS Academy (this amount varies by course).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 1-8 | Note to Instructors: At this time, the AWS Academy process begins with an application from your institution. There is no cost to join–it just requires filling out an application and waiting for approval. The application should be completed by an instructor or administrator who will oversee your school's participation in the program; this person becomes the CPOC (central point of contact). This process should be completed before the course begins, as there might be a delay while your school's application is processed.

Once the institution is accepted, the CPOC invites instructors to join. The instructor then sets up a course and invites students. Students get free credits they can use during each course they join to complete assigned activities. The best course type to choose for this Cloud+ course is the "Learner Lab" course. Learner Labs include a sandbox environment in the AWS cloud where students can experiment with certain AWS services using their credits.

Additionally, students can get discounts on AWS certification vouchers. Instructors also have access to their own training and voucher discounts for AWS certification exams. For more information, visit aws.amazon.com/training/awsacademy/. Note that this information changes often. For additional updates, please email the author at jillwestauthor@gmail.com.

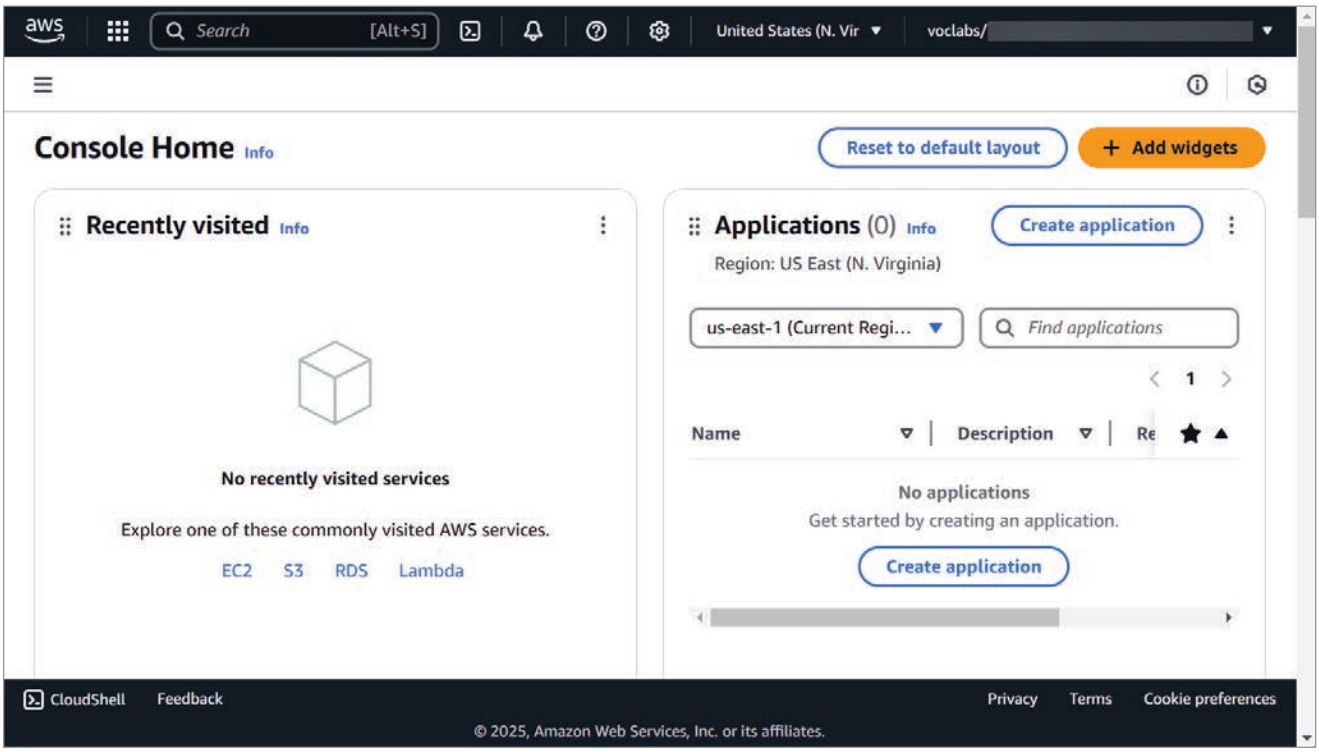
Check with your instructor to determine which type of account you should create. From the following steps, choose the set of steps that matches the type of account you will create, and then complete the steps to create your account.

Part 1: Standard AWS Account Option

If you're creating a **standard AWS account**:

1. Go to aws.amazon.com, click **Sign in**, and then click **Create a new AWS account**.
2. Complete the steps on the screen. You'll need a credit card number and an email address you check regularly.
3. Once your account creation process is complete, return to **aws.amazon.com** and sign in using your new AWS account. Make sure you can get to the AWS Management Console (Figure 1-20). Spend a few minutes clicking through the guided tour, if available, and exploring the console, and then continue with Part 3.

Figure 1-20 AWS Management Console



Source: Amazon.com, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Part 2: AWS Academy Account Option

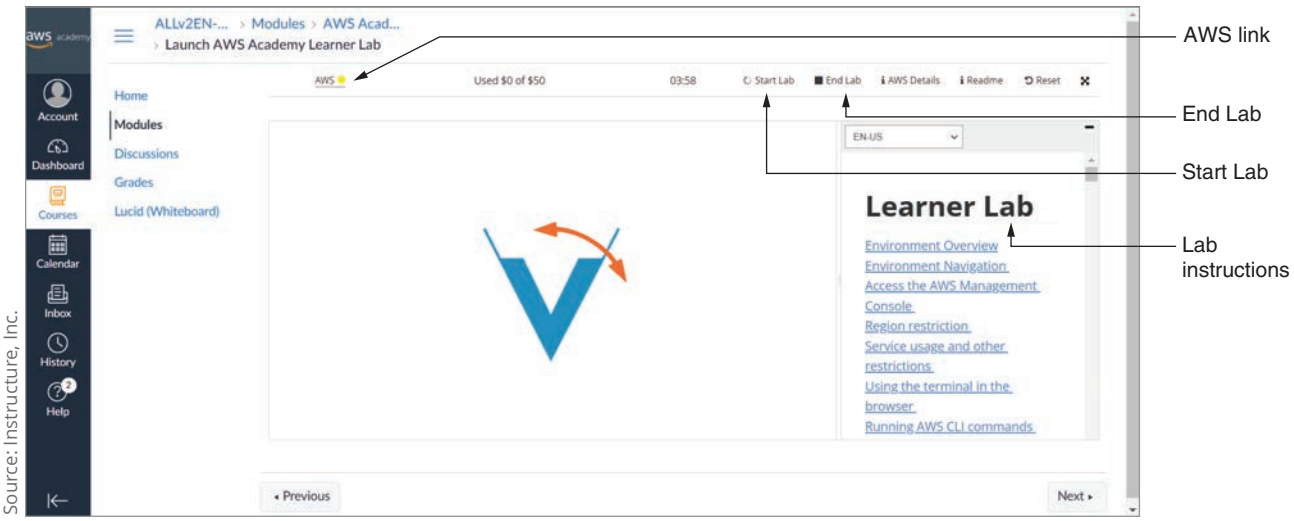
If you're creating an **AWS Academy account**:

1. Check your student email account for an invitation email from your instructor through AWS Academy.
2. Follow the instructions in the email to complete your account creation.
3. Once your account creation process is complete, use the portal address provided by your instructor or visit **awsacademy.instructure.com** to return to the AWS Academy portal and click **Student Login**. Sign in using your new AWS Academy account.

Note 1-9 | Make sure to bookmark your AWS Academy portal address for future reference. You will use the page again later in this course.

4. If you clicked the link in your email invitation, the platform might take you to your course automatically. If not, on your Dashboard, click the tile for your Learner Lab course.
5. On the Learner Lab course's page, click **Modules**, and then click **Launch AWS Academy Learner Lab**.
6. The right pane contains instructions on how to use the lab environment. Take a few minutes to read through that information.
7. When you're ready, click **Start Lab** (Figure 1-21). It might take a minute or two for the lab environment to load. When the AWS dot turns from yellow to green and the information in the right pane updates, click the **AWS** link next to the colored dot.

Figure 1-21 Launching the AWS Academy Learner Lab



Source: Instructure, Inc.

8. A new tab should open to your AWS Management Console (Figure 1-20). Spend a few minutes clicking through the guided tour and exploring the console, and then continue with Part 3.

Part 3: Exploring AWS for Both Account Options

As you complete the next few steps, notice the differences between service-specific dashboards, such as the EC2 Dashboard. **Do not create any new cloud resources.**

1. In the upper-left corner of the console, click the **Services** button, which appears like a grid. How many categories of services are listed in the left pane?

IaaS services are used to build virtual infrastructure in the cloud, such as virtual networks or virtual servers. In AWS, you use the VPC (Virtual Private Cloud) service to create and manage virtual networks, and you use the EC2 (Elastic Compute Cloud) service to create and manage virtual servers.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. Click the **Networking & Content Delivery** category, and then click **VPC** to go to the VPC dashboard. How many VPCs currently exist in your account for this region?
3. Return to the **Services** list. Click the **Compute** category, and then click **EC2** to go to the EC2 dashboard. How many EC2 instances currently exist in your account for this region?

PaaS services are often used to build and test software. In AWS, you use the CodePipeline service to manage code throughout an application's life cycle. Alternatively, you can run code for short periods of time in Lambda. Lambda is especially useful when you need to run code to handle data in AWS and you don't want to manage a server for running that code.

4. In the **Services** menu, click the **Developer Tools** category, and then click **CodePipeline** to go to the CodePipeline dashboard. What service in CodePipeline is used to securely store, publish, and share software packages?
5. In the **Services** menu, click the **Compute** category, and then click **Lambda** to go to the Lambda dashboard. What are the two primary types of resources you can create in Lambda?

SaaS services are typically used directly by end users, both technical and nontechnical. AWS offers a few SaaS services, though this is not really the focus of the AWS platform. These services require some configuration from cloud administrators before they are ready for end users to access.

6. List four AWS services in the End User Computing category.
7. When you're finished exploring, return to the Console Home page.
8. In the upper-right corner of the window near your account name, notice that a region is listed, such as United States (N. Virginia). What region is currently selected in your account?
9. **Take a screenshot** showing your region and account name in the upper-right corner of your AWS Management Console; submit this visual with your responses to this project's questions.
10. When you're finished, sign out of your AWS account. If you are using an AWS Academy account, on the Learner Lab tab, click **End Lab** and then click **Yes**. Completing cleanup steps at the end of each project will help preserve your AWS Academy course credits.

Project 1-3: Create an Account with Azure

Estimated time: 30 minutes
Objective 4.3: Given a scenario, implement identity and access management.
Resources:

- Internet access
- Microsoft account
- If using a free trial or pay-as-you-go subscription, a credit card
- If using an Azure for Students subscription, a school-issued email address

Description: Microsoft offers two options for a free Azure account that work for this course. To use Azure, you create a Microsoft account and then create one or more subscriptions within the account. If your school uses Office 365, you might already be able to authenticate to Azure using your existing Microsoft account. The two types of Azure subscriptions (there are others) most relevant to this course are as follows:

- **Free trial subscription**–This subscription is available to the public and requires a credit card. The $200 free credit expires after 30 days, whereas the free trial services last for 12 months. The subscription itself is free; any services you use beyond the free offers will be charged to your credit card. You will also be offered a technical support plan, starting at $29/month, which you can decline.

Note 1-10 | Microsoft offers a generous amount of credit for the first 30 days of the free trial. A challenging disadvantage of the free trial is that, after 30 days, you must upgrade to a pay-as-you-go subscription without any remaining credits, and at that point, it can be difficult to identify which services

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

are free and which are not. Azure doesn't mark the free services as clearly as AWS does. Those free services are only available for a year, at which point you're left with the "always free" services, which are somewhat limited. Therefore, if you use Azure, carefully pay attention to resource configurations, run times, and potential charges.

- **Azure for Students subscription**–This subscription is available to full-time faculty and students (older than age 18) at an accredited, degree-granting educational institution. You must use a school-issued email account, but you do not need a credit card. You might also be able to generate an Azure for Students verification code through your school's Microsoft Azure Dev Tools for Teaching platform. Ask your instructor for additional information on this robust resource. You get $100 free credit that lasts 12 months, and you can renew your student subscription after a year for another disbursement of free credits. However, if you run out of your free credit before the 12 months is over, you will have to sign up for a pay-as-you-go subscription using a credit card to retain access to your account. Be very careful to manage your credits and avoid running out. With Azure, your instructor has no options for renewing your credit if you use it up before the class is over.

Note 1-11 | Note to Instructors: At this time, Microsoft's Azure for Students is managed by a team at Microsoft. Students apply directly through their program to confirm their status as a qualifying student.

Your school might already offer the Azure Dev Tools for Teaching program. If so, students might have an invitation from your program administrator, or they might be able to use their existing student email (if these emails are Microsoft accounts) to access their Azure resources. According to the Microsoft site, students must be at least 18 years old and must be a full-time student.

For more information, visit https://azure.microsoft.com/en-us/free/students or search for updated information to the Azure for Students program. Note that this information changes often. For additional updates, please email the author at jillwestauthor@gmail.com.

Check with your instructor to determine which type of account you should create. From the following steps, choose the set of steps that matches the type of account you will use, and then complete the steps to create your account.

Option 1: Standard Azure account

If you're creating a **standard Azure account using the free trial subscription**:

1. Go to **azure.microsoft.com**, activate the free trial, and create a Microsoft account or sign in with an existing Microsoft account. Complete the identity verification and agreement process. You can decline any technical support options. Once you're able to sign into your Azure account, navigate to the Home page (Figure 1-22).

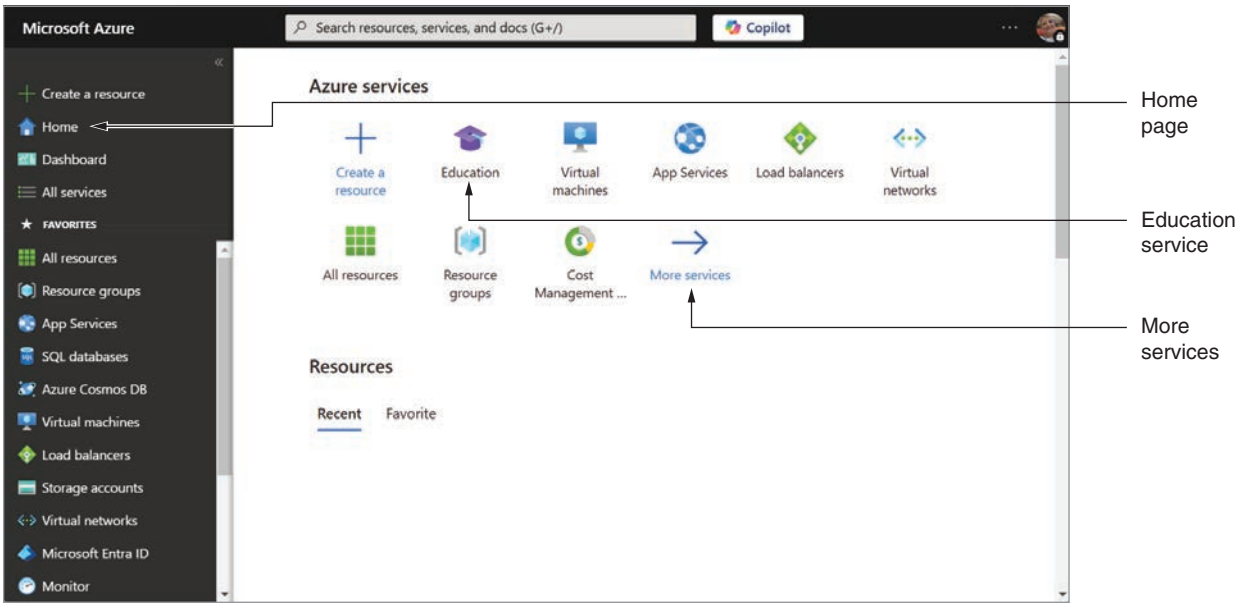
Option 2: Azure for Students subscription

If you are creating an **Azure account using the Azure for Students subscription**:

1. Go to **azure.microsoft.com/en-us/free/students**, activate the Azure for Students offer, and create a Microsoft account or sign in with an existing Microsoft account. Complete the identity verification and agreement process. You're automatically taken to the Education Overview page, which displays your available free credits. Revisit this page periodically to monitor your credit usage. For now, navigate to the Home page (Figure 1-22).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 1-22 Azure Home page



Options 1 and 2: Exploring Azure for Both Account Options

After you sign in, complete the following steps to explore the Azure portal. **Do not create any new cloud resources.**

2. Spend a few minutes exploring the portal. Under Azure services, click **More services**. How many categories of services are listed in the left pane?
3. Click a few services to display service-specific dashboards, such as the Virtual machines dashboard.
4. Return to the **All services** page, and click the **Compute** category.
   a. What are two resource types listed in the IaaS section?
   b. Which section contains the services you would need to create containers?
5. Click the **Networking** category.
   a. Which service do you think you would need to create your own network space in Azure?
   b. What are two services in the Hybrid connectivity section?
   c. What are two services in the Network security section?
6. Return to the **Home** page. **Take a screenshot** showing your account name in the upper-right corner of your Azure portal; submit this visual with your responses to this project's questions.
7. When you're finished, sign out of your Azure portal.

Project 1-4: Create an Account with GCP

Estimated time: 30 minutes
Objective 4.3: Given a scenario, implement identity and access management.
Resources:

- Internet access
- Google account
- If not using an education option, a credit card
- If using an education option, information from your instructor on how to sign up

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Description: Google Cloud offers an attractive free trial option available for all new accounts, which provides a generous amount of credit for 90 days. Its biggest drawback is that it requires a credit card or bank account information in case you exceed the free credit amount. Projects in this course will not require that you use all your free credit so long as you follow the steps carefully and remember to delete resources when you're not using them. GCP's education options must be organized through your school. Your instructor will give you this information if relevant for your specific course.

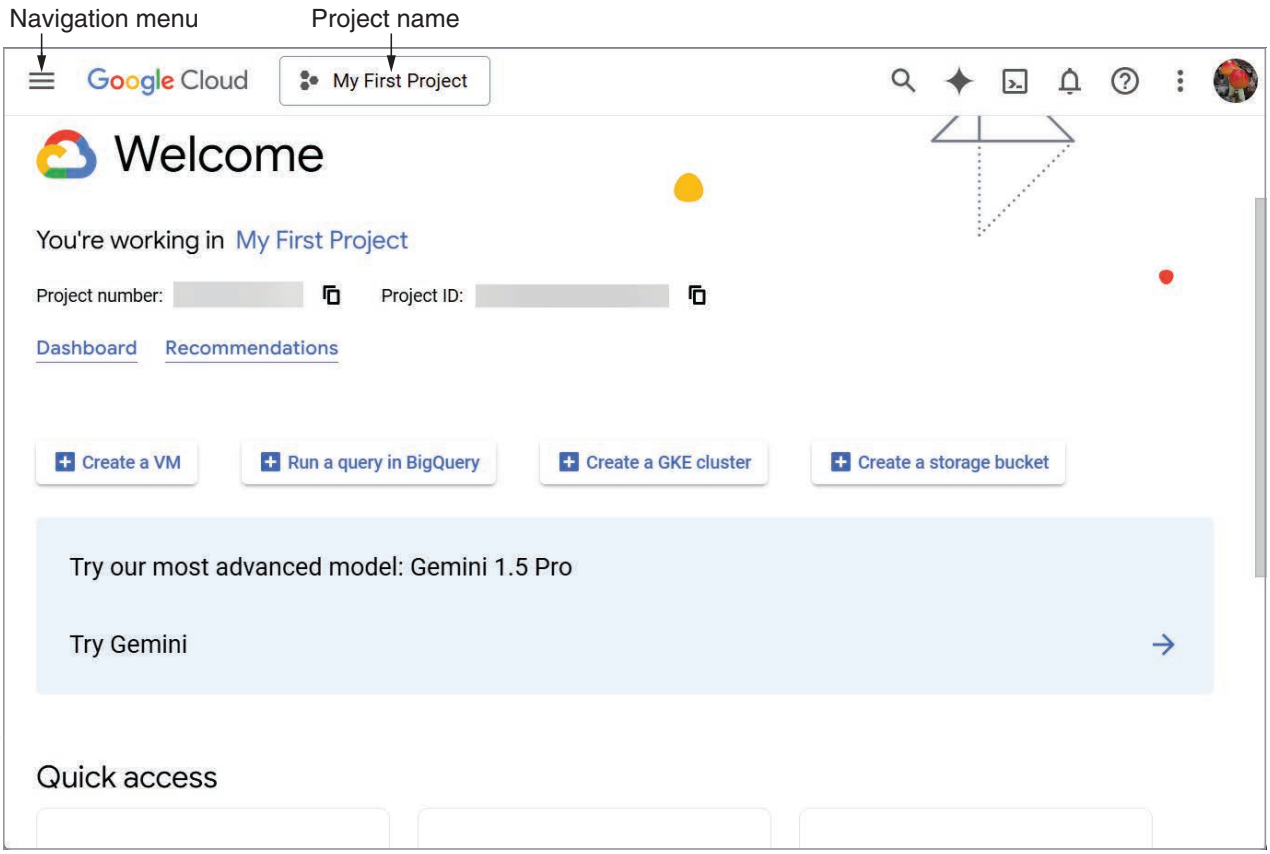
Note 1-12 | Note to Instructors: At this time, Google offers the Google Cloud for Education program. The program offers free Google Cloud credits to students, which can be distributed by the instructor. Faculty members must apply for the program and allow 15 business days to process their application.

For more information, visit https://cloud.google.com/edu/faculty or search for updated information to the Google Cloud for Education program. Note that this information changes often. For additional updates, please email the author at jillwestauthor@gmail.com.

Complete the following steps to create your account. **Do not create any new cloud resources.**

1. Go to **cloud.google.com** and sign up for the free trial.
2. Spend a few minutes exploring the GCP console (Figure 1-23). An initial project has been created for you, which is listed at the top of the window just left of center. What is the name of your project?

Figure 1-23 GCP console



Source: Google LLC

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. Click the **Navigation menu** button to open the navigation pane, and then click **VIEW ALL PRODUCTS**. Notice the All products page has its own navigation pane. How many categories are listed?
4. Click a few services to view service-specific dashboards, such as the Compute Engine dashboard.
5. Return to the **All products** page, and click the **Compute** category.
   - What are two services listed in the Compute category?
   - Which service would you need to create containers?
6. Click the **Networking** category.
   - Which service do you think you would need to create your own network space in GCP?
   - Which service do you think you would need to set up hybrid connectivity in GCP?
7. Use the **Navigation menu** to go to the **Dashboard**. Which tile displays your estimated charges?
8. In the upper-right corner, click to show your account name and email address. **Take a screenshot** showing your account name in the upper-right corner of your Google console; submit this visual with your responses to this project's questions.
9. When you're finished, sign out of your Google console.

Challenge Projects

Note 1-13 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 1-1: Cloud Service Types

Estimated time: 30 minutes
Objective 1.1: Given a scenario, use the appropriate cloud service model.
Resources:

- Internet access
- Cloud platform account

Description: These Challenge Projects are intended to give you practice figuring out how to do tasks in the cloud without step-by-step guidance. While most of the popular cloud platforms offer extensive and well-written instructions, sometimes it's faster and easier to poke around and figure it out for yourself. Further, the tasks in these Challenges will help you practice recalling the steps you performed in the Hands-On Projects. The steps tell you what task to accomplish and ask questions along the way to help you think about what you're doing. However, the steps intentionally do not provide detailed instructions on how to perform the task.

Complete the following steps. **Do not create any new cloud resources.**

1. You can use AWS, Azure, or GCP for this Challenge project; alternatively, if allowed by your instructor, you can use another cloud platform such as Alibaba Cloud or IBM Cloud. Sign in to the cloud platform. Which platform are you using?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. Study the service categories provided in this cloud platform, and complete Table 1-3.

Table 1-3 Cloud services

Service model	Name of a service that fits in each service model	Why do you think this service is included in this service model?
IaaS		
PaaS		
SaaS		

3. What is one category of services you think would be interesting to learn more about?
4. Choose one of the services listed in that category. What service did you choose?
5. Read about the service in the cloud platform or on the CSP's website. In your own words, what does this service do? Be sure to provide sufficient detail to show you understand the purpose of the service and the basics of its role in a cloud deployment.
6. Sign out of your cloud account.

Solutions to Self-Check Questions

Section 1-1: Characteristics of Cloud Computing

1. Which cloud platform does the CompTIA Cloud+ certification target?
   Answer: d. None
   Explanation: The CompTIA Cloud+ certification is vendor neutral and so **does not target any specific cloud platform**.

2. Which of these certifications requires the least technical skill?
   Answer: c. AWS Cloud Practitioner
   Explanation: The **AWS Cloud Practitioner** certification is a nontechnical, entry-level AWS certification appropriate for professionals in technical, managerial, sales, purchasing, or financial roles. The other certs all emphasize technical skills.

3. Which of the following is *not* an essential characteristic of cloud computing technologies, according to NIST?
   Answer: a. Adaptive security
   Explanation: Cloud providers often describe additional characteristics and benefits to cloud computing, including **adaptive, intelligent security** that takes advantage of AI technology to improve built-in security defenses. However, this is not an essential characteristic of cloud computing as defined by NIST. According to NIST, the five essential characteristics of cloud computing are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.

Section 1-2: Cloud Deployment Models

4. What factor limits the capacity of a private cloud to provide resources to users?
   Answer: c. Hardware

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Explanation: Because a private cloud runs on **hardware** dedicated to only that cloud, its capacity is limited to the capabilities of the hardware on which it runs. The OS and security setting options for a private cloud are similar to those of a public cloud. Although the geographic location of a private cloud might limit some of the capabilities of a private cloud, it is not the limiting factor for a private cloud's capacity.

5. An AWS-hosted network linked to Google Workspace productivity apps is an example of what kind of cloud?

Answer: d. Multi-cloud

Explanation: A **multi-cloud** model involves using cloud services from multiple vendors at one time, such as AWS and Google. A community cloud's services are accessible to multiple organizations with similar concerns but are not available to the general public. Private cloud services are hosted on hardware resources used exclusively by a single organization. A hybrid cloud is a mix of both public and private cloud components, or a combination of some cloud and some traditional on-prem services, where those components and services interact with each other in a direct and seamless manner.

6. Which of the following scenarios is most likely to require cloud bursting?

Answer: c. A web server during an advertising campaign

Explanation: An **advertising campaign** is temporary in nature and can create highly fluctuating traffic to a web server; in this case, cloud bursting could help alleviate some of the demand on the regular web server by shifting some of the traffic to cloud resources. The demand on a database will likely remain steady or slowly increasing after a cloud migration. The demand on a firewall during application testing is not likely to increase significantly. The demand on resources hosting an application is likely to decrease at the end of the workday rather than increase.

Section 1-3: Cloud Service Models

7. Which cloud service model requires the least skill to use?

Answer: d. SaaS

Explanation: With **SaaS (Software as a Service)**, applications can be accessed from different types of devices without having to manage any of the underlying infrastructure such as the network, servers, operating systems, or storage. This type of service requires little technical skill to configure and use. IaaS (Infrastructure as a Service) and PaaS (Platform as a Service) services require more technical skill to set up, maintain, and secure. DRaaS (Disaster Recovery as a Service) also requires more technical knowledge than a SaaS application to configure and use.

8. Which cloud service model requires the most skill to adequately secure?

Answer: a. IaaS

Explanation: IaaS (Infrastructure as a Service) customers have more control and flexibility with configurations and security features than do PaaS (Platform as a Service) or SaaS (Software as a Service) customers. Therefore, IaaS customers retain greater responsibility for the security of their IaaS deployment. RaaS (Ransomware as a Service) is used to attack, not secure resources.

9. Which database deployment option is a PaaS?

Answer: a. Serverless

Explanation: With a **serverless** database, the customer does not manage the servers at all—this is a PaaS (Platform as a Service) design. With a self-managed database, the customer manages the database and the server hosting it—this is an IaaS (Infrastructure as a Service) design.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

With a managed database (called fully managed or provider-managed), someone else manages the servers hosting the database—this is also an IaaS design.

Section 1-4: Cloud Services

10. Which public cloud platform offers the most streamlined integration with Windows Active Directory?

Answer: c. Azure

Explanation: Azure is designed for seamless compatibility with Microsoft's other tools, such as Windows Active Directory. Although other public cloud platforms, such as AWS and GCP, offer Active Directory compatible services, it is not natively integrated into these platforms.

11. In which cloud category would you most likely find a service for configuring DNS records?

Answer: a. Networking

Explanation: DNS is a **networking** function and will be found in the Networking category in most cloud platforms. Security category services offer features such as firewall type protection. Compute category services offer options for running applications, such as VMs or containers. Storage category services offer options for saving data.

12. Which IoT device provides access to a personal assistant app?

Answer: b. Smart speaker

Explanation: Voice-controlled **smart speakers** offer access to embedded personal assistant apps such as Alexa, Siri, and Google Assistant. Security cameras and smart toasters can often be controlled via personal assistant apps. GPS locators typically don't offer features for making changes to the device, other than registering it with a specific account.

Section 1-5: Troubleshooting Methodology

13. What cloud configuration technique can be used to compensate for capacity challenges?

Answer: d. Auto-scaling

Explanation: Cloud services can be designed to **auto-scale** in order to avoid capacity-related problems. Latency is a delay often caused by poor routing, insufficient bandwidth, or a bottleneck. Connectivity issues could originate from local network hardware or from endpoint configurations. Change management processes to track and control configuration changes can help minimize the likelihood of problems in a cloud environment.

14. Most often, what will the last step be when troubleshooting a problem?

Answer: a. Documentation

Explanation: Once a problem has been resolved, the symptoms, findings, actions, and outcomes should all be **documented** in a knowledge base for future reference. Testing your theory, implementing a solution, and verifying functionality all happen earlier in the troubleshooting process.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Chapter 2

Cloud Compute

Chapter 2 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Explain compute virtualization technologies.
2. Compare on-prem virtualization with VMs in the cloud.
3. Describe serverless computing and container technologies.

Chapter 2 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 2-1: Virtualization Technologies
- VMs (Virtual Machines)
- VM Configuration
- Network Connection Types

Section 2-2: VMs in the Cloud
- VM Instance Types
- Cloning VM Instances
- Billing Models
- Resource Allocation Factors
- Affinity

Section 2-3: VM Alternatives
- Serverless Computing
- Containers

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 2 Scenario Three days after your first cloud-centric staff meeting, Kendra and Malik shuffle into the conference room where you've been setting up a brief presentation on the projector. You've collected notes from the others and from your own research, and you've merged all the information into this overview so everyone will be up to date on what you've all learned. Part of the presentation covers services you can access in the cloud, and part of it gives a survey of popular cloud providers and the types of services they're best known for. It's a lot of good information, and you're actually kind of excited about it—but you dread doing any public speaking, even in front of your two coworkers. You feel your heart pounding in your chest, and imagining your boss as a purple unicorn isn't helping you relax.

As Kendra and Malik chat about lunch plans, you knock over your coffee cup, squelch a few swear words, and quickly wipe up the mess. Thankfully, there wasn't much left in the cup, but you're annoyed to notice your hands shaking. Deep breath. It's gonna be okay.

You start the presentation, and almost immediately Kendra and Malik jump in with questions. Forty minutes later, you realize none of you paid much attention to the PowerPoint because you all had so much information you wanted to share and talk about and even more questions to explore. Kendra pauses and takes the metaphorical step back as she asks, "Okay, so what's next? We know we want to start moving to the cloud. How do we do this? Where do we begin?"

Malik says, "Well, I know I'm the newbie here. But I wonder, what assets do we have in our network that could be moved to the cloud? I mean, how do we go from a server that's sitting in our rack to some kind of service in the cloud? I guess I don't really understand how that happens."

Kendra responds, "We don't actually ship the server computer to the cloud provider. We can shift some of our virtual servers from our on-prem network to the cloud. So basically, instead of running these servers on our own hardware, we would use the cloud provider's hardware."

You jump in as you get the idea of where she's going with this train of thought. "Right, that's the essence of the cloud—using someone else's hardware. Right?"

Kendra smirks a little, "At a very basic level, yes, this is true. But in reality, the cloud also enables system design opportunities and technologies that are difficult or impossible to implement on-prem. It's not just a question of whose hardware is being used—it's really a whole new way of thinking about system architecture."

Malik seems a little confused. "How does all of that work? Why does using different hardware change how you might design a system?"

"In the on-prem environment, we usually think in terms of servers," says Kendra. "We create virtual servers that each have their own operating system, and a lot of resources actually go to waste. In the cloud, it's easier to set up servers that start and stop in seconds, that don't need entire operating systems. We can even run code directly in the cloud without needing a server at all."

As she scrunches her face in thought, she adds, "I think we need to do some research into computing options in the cloud so the two of you have a thorough understanding of this concept before we move forward." She scratches out a list on her notepad and hands you a sheet of paper. Then she adds, "Here's what you'll need to research before our next meeting."

- How does virtualization work on-prem?
- How do you create a virtual server in the cloud?
- What other kinds of resources can run applications and host network services?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 2-1: Virtualization Technologies

Certification

1.7 Compare and contrast virtualization concepts.

1.10 Compare and contrast methods for optimizing workloads using cloud resources.

2.5 Given a set of requirements, provision the appropriate cloud resources.

Virtualization is a virtual, or logical, version of something rather than the actual, or physical, version. Think about a house and the people who live there. A house's purpose is to provide a living space. Then take the logical *idea* of "living space for people," and put a lot of those living spaces in one building, like an apartment complex. There's still only one building, but many people and families have their own living spaces within that building, each complete with a kitchen, living room, bedrooms, and so on. They share physical resources, such as electric and plumbing services, but they don't all have to furnish, decorate, or use their apartments the same way as everyone else in the building. Some apartments might have one bedroom, while others have two or three. Some apartments might be organized and minimalistic, while others are cluttered or have a more lived-in appearance.

Virtualization works with much the same premise—the physical machine can be divided into pieces that support several virtual systems. You don't give one virtual system the CPU (central processing unit) and another virtual system the RAM (random access memory). Instead, you take the capacity each hardware component offers and divide that capacity among the virtual systems running on the physical host. These virtual systems are called virtual machines, as described next.

VMs (Virtual Machines)

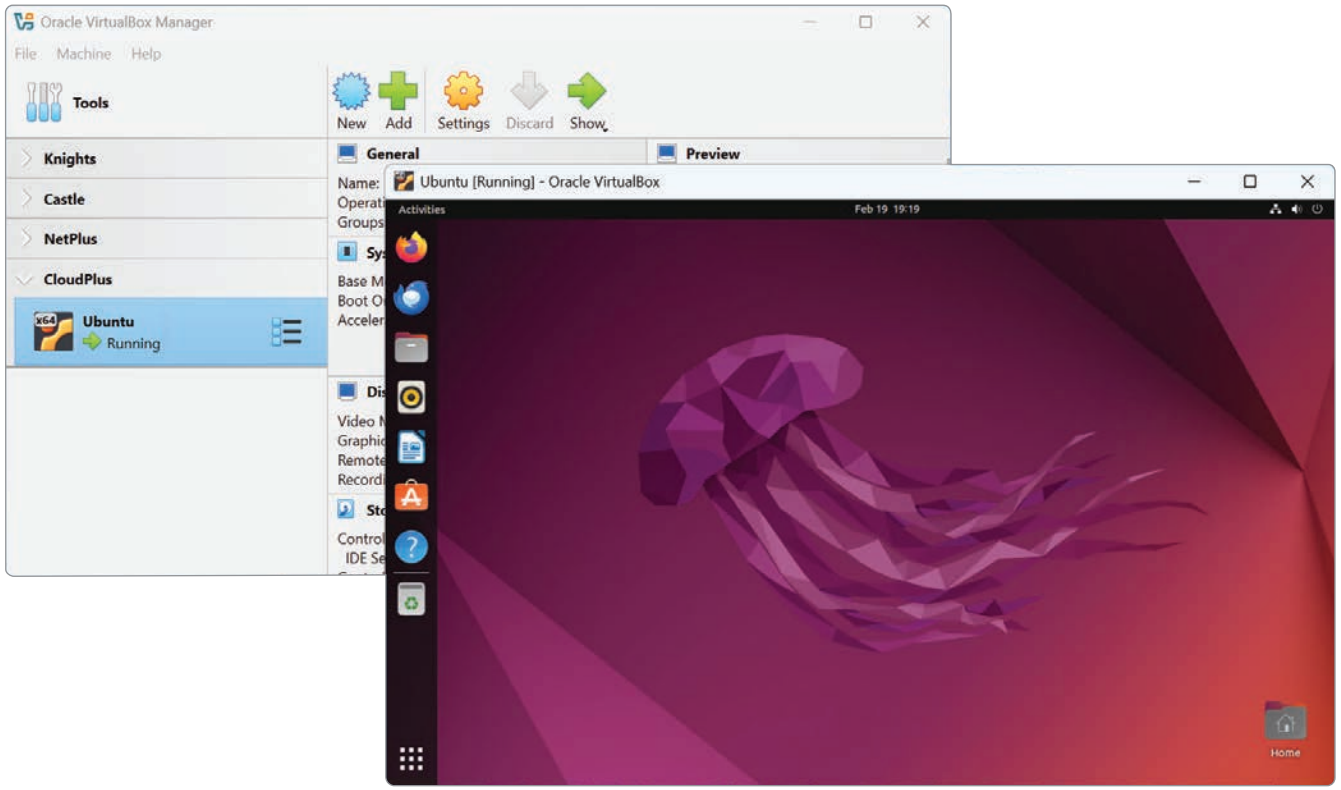
VMs (virtual machines) are similar in concept to the apartments in an apartment building. All the VMs on a physical computer share the same hardware resources, but each VM can have its own OS (operating system). For example, when you create an Ubuntu server VM on a Windows PC, the Windows machine is the physical computer, or host. The Ubuntu machine is the virtual, logical computer, or guest, that is hosted by the physical computer. This is made possible through the use of a hypervisor, which creates and manages each VM. It also manages hardware resource allocation and sharing between a host and any of its guest VMs. In Figure 2-1, a Windows 11 host machine runs a hypervisor with an Ubuntu guest VM.

Note 2-1 | A type 2 hypervisor is installed over an existing OS as an application, and then guest OSs are installed inside the hypervisor. In contrast, a type 1 hypervisor is installed directly on the firmware of the physical machine and is sometimes erroneously referred to as a "bare-metal" hypervisor. In reality, a type 1 hypervisor is itself a minimal operating system. Like any OS, the hypervisor relies on firmware to enable communication with the underlying hardware. Because there is no underlying host OS using hardware resources or limiting the type 1 hypervisor's access to those resources, a type 1 hypervisor is more powerful than a type 2 hypervisor. It's also faster and more secure.

A popular hypervisor is Hyper-V, which is embedded in Window Server 2008 and beyond, and in Windows 10 and beyond (sometimes called Client Hyper-V). Hyper-V has elements of both categories of hypervisors. Although it appears to run more like an application, in reality, Hyper-V creates a virtualization layer underneath the existing Windows installation, thereby establishing its role as a type 1 hypervisor. Native to Linux OSs is a hypervisor called KVM (Kernel-based Virtual Machine), which also has elements of both a type 1 and type 2 hypervisor. Popular type 2 hypervisors include Oracle VirtualBox, VMware Workstation, and Parallels Desktop.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

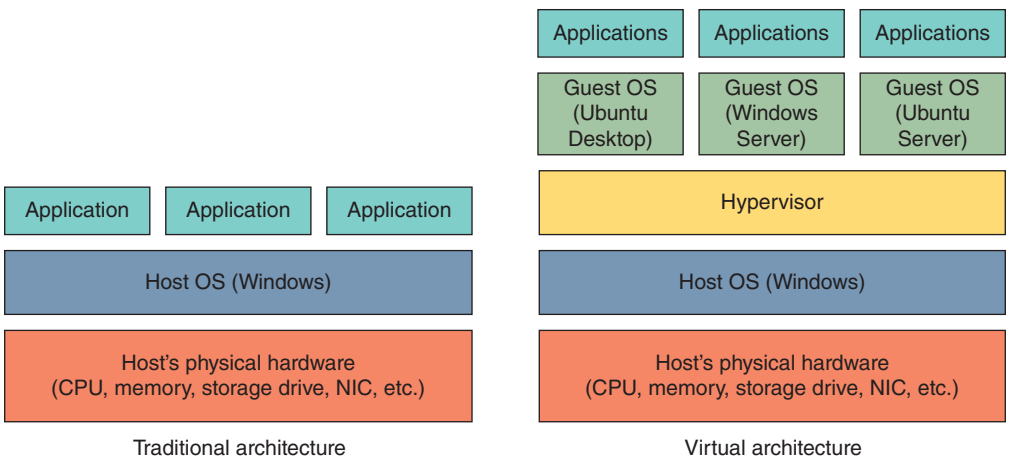
Figure 2-1 A Windows 11 host with Ubuntu Desktop VM in Oracle VirtualBox hypervisor



Sources: Oracle Corporation; Canonical Group Limited

Together, all the virtual devices on a single computer share the same CPU, storage disks, DIMMs (dual in-line memory modules), and physical network interfaces. Yet each VM can be configured to use a different operating system, and it can run its own applications. Figure 2-2 illustrates an example of some of the elements of virtualization when hosted on a user's computer as compared to the traditional, nonvirtual installation of a single OS and its applications.

Figure 2-2 A hypervisor allows for installation of multiple guest VMs



A single physical machine with robust hardware that is hosting several VM servers can take the place of an entire rack of physical servers. The hypervisor dashboard, then, provides central control of all the VMs.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 2-2 | A server can also host a **VDI (virtual desktop infrastructure)** environment where users sign in remotely and are presented with a virtual desktop, such as Windows or a Linux distro. This eliminates much of the need for expensive workstations because with VDI, most of the processing is done on the server and not on the user's local, physical computer. VDI can also provide either persistent or nonpersistent instances. With a persistent instance, when the user remotes back into the desktop, any changes they made in previous sessions will still be there, including files they saved. With a nonpersistent instance, the desktop is reset each time someone signs in. Top-rated VDI solutions include Amazon Workspace, Microsoft Azure VDI, Citrix Workspace, and Parallels RAS (Remote Application Server).

VM Configuration

When configuring VMs inside of a hypervisor, you're given options for hardware resource allocation. Important criteria for deciding among these resources are the number and size of CPUs and GPUs (graphics processing units), amount and type of memory, storage type and space, and access to device drivers. These decisions are important whether you're building the VM in your own hypervisor or whether you're building VMs in the cloud. Let's briefly discuss how these decisions work with a local hypervisor, and then you'll learn about cloud-based CPU and memory resources for cloud-hosted VMs.

CPUs

The number of CPU cores dedicated to a specific VM is measured by **vCPUs (virtual CPUs)**. It might seem that the more vCPUs a VM has, the better, but this isn't necessarily true. Through a process called CPU scheduling, the hypervisor schedules a VM's access to the physical CPU cores available on the host machine. This scheduling gives each VM and the host the access they need to the available processing capacity on the host CPU. Giving a VM more vCPUs than it really needs can cause unnecessary delays. This is because additional scheduling is required to give every allocated vCPU processing time even though they are not really needed.

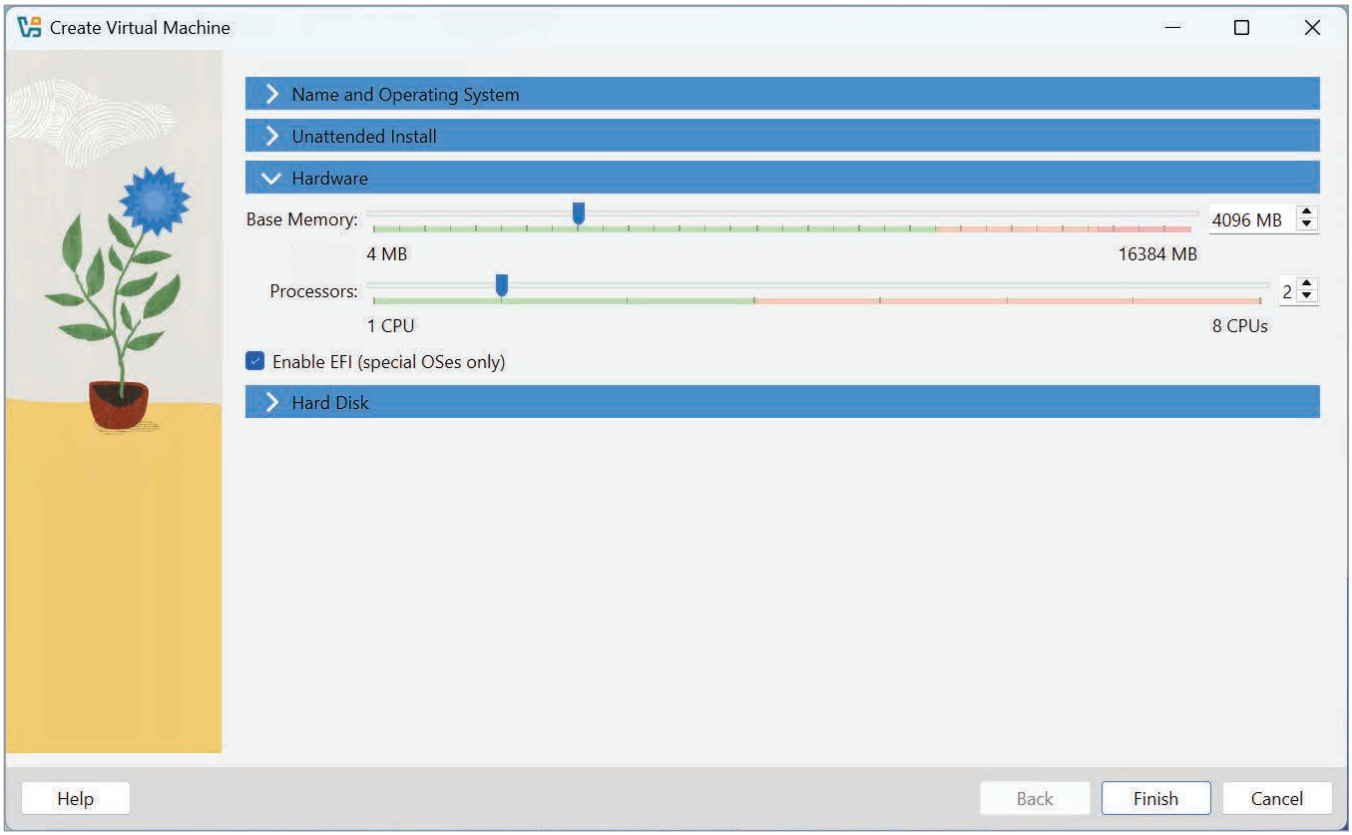
The situation is like having too many people in the kitchen when you're trying to prepare a large meal. If each cook is working on different dishes and needs different tools, it can be helpful to have several people in the kitchen at one time because the available resources rarely limit any individual's work. However, if multiple cooks frequently need the same tools, such as when only one oven is available for five cooks who each need to bake a casserole at different temperatures, then having all five cooks work on their dishes at the same time will cause conflicts. If only one casserole can bake at a time, there's little benefit to having more than one cook working on casseroles. Similarly, with CPU cores, if multiple cores are frequently trying to use the same resources (such as cache, RAM, files, or network interfaces) when performing parallel tasks, it might be better to have fewer cores that perform those tasks in a more serial manner, one after another.

Unless you already know a VM will need a high number of vCPUs, the best practice when creating a VM is to start with one or two vCPUs and then test the system when it's running its normal processes for all VMs on the server (Figure 2-3). If you do increase the number of vCPUs for any VM, be sure to carefully plan the ratio of physical CPU cores to vCPUs needed at any given time.

There are other reasons to minimize vCPU allocation to your VMs. For example, licensing could be an issue. Most people are familiar with obtaining software licenses per user or per device. The technical term for this arrangement is CAL (client access license). In many cases, however, organizations need **volume-based licensing**, which is designed for many users under one license. With volume-based licensing, costs are calculated not according to the number of users or the number of machines on which the software is installed but, instead, according to the number of CPU sockets in a system or (more recently) according to the number of logical CPU cores used to execute that software. Especially in the case of cloud-hosted VM instances, this is the same as the number of vCPUs. When using core-based licensing, administrators might need to minimize the number of vCPUs for instances hosting certain applications so they can manage software license expenses more efficiently.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-3 Assign a number of vCPUs to a VM



Source: Oracle Corporation

GPUs

When you're building a high-performance computer such as a gaming system, you'll typically include a graphics card to increase its graphics processing capability. The graphics card's onboard GPU brings additional cores to the computer's hardware resources. While GPU cores tend to run a bit slower than CPU cores, they are designed to handle high volumes of parallel functions. Additionally, a GPU can typically offer more cores than a CPU can within the same price range. A GPU's parallel processing is especially helpful for tasks such as machine learning and scientific computation of large data sets.

AI in IT | The demand for GPUs has increased substantially over the past few years as more companies are investing in training, fine-tuning, and running their own AI models. AI model training requires massive amounts of parallel processing resources, which GPUs are designed to support. GPUs can also boost performance during the AI inference phase, which is when an AI tool responds to user requests.

Similar to CPUs, you can also allocate GPU capacity to a VM on a hypervisor or in the cloud. This can be accomplished using one of two approaches:

- **GPU pass-through**—In this scenario, the GPU's entire capacity can only be used by one VM at a time.
- **Virtual or shared GPU**—In this scenario, the GPU's resources are artificially divided into portions that can be allocated at the same time to multiple VMs. Each partial allocation of available GPU resources is sometimes referred to as a **vGPU (virtual GPU)**.

Which approach to use is typically decided by the physical GPU in your system (that is, whether it's capable of supporting shared GPUs) and the virtualization software (that is, which of these options it offers).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 2-3 | Hardware pass-through configurations in a hypervisor can also support other types of hardware devices, such as a network card connected via PCIe (peripheral component interconnect express) or a printer connected via USB (universal serial bus).

Memory

Whether virtual or physical, more memory is better in nearly all cases. The challenge when configuring VMs on a physical machine is to make sure that each system—including the host—gets the memory it needs. When a VM is running, it will typically reserve the amount of physical memory allocated to it whether or not it's using that much memory at any given time. In contrast, some hypervisors allow for dynamic allocation of memory, which reserves only what it needs for currently running processes. Reserved memory is called **vRAM (virtual RAM)**. It's important that you leave enough physical memory for the host machine even when you have multiple VMs running.

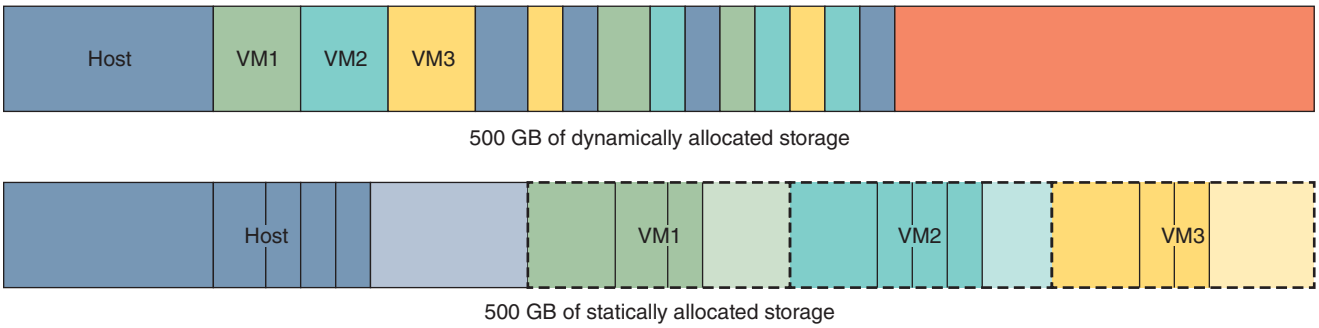
Similarly, make sure each VM can reserve sufficient memory for the OS and applications installed in it. Because of these requirements, the memory allocated to VMs on your local machine will necessarily be limited by the amount of physical RAM you have installed on the physical host. This is why a workstation or server customized for hosting VMs should be given as much physical RAM as your budget will allow.

Storage

A VM needs storage space not only for saving files but also for running software such as an OS and applications. Most hypervisors allow for dynamic storage configuration for VMs. This means the VM will only use as much storage space as it actually needs for its existing data. For example, you might provision 100 GB of the host's 1 TB hard disk for a VM, but if the VM only has 50 GB of data, the other 50 GB of space will remain available to the host or other VMs. This works fine until the VMs or the host start to use up most of their allotted storage space. To accommodate these slowly increasing storage needs, you can overcommit the currently available storage space and then add more physical disks to expand the available space at the time it's needed.

However, another complication arises from using dynamic storage space. Suppose you have a host with three VMs. The host's hard disk has 500 GB available, and each VM is allocated 100 GB of dynamic storage space to be reserved only as needed. As illustrated in the top portion of Figure 2-4, the host has already used the first 75 GB for its own purposes, and then each of the VMs uses another 20–25 GB as they're created and installed.

Figure 2-4 Statically allocated storage space is accessed more efficiently



Over time, the disk space is used a little here and there by the host and each VM. Accessing data spread randomly across the disk becomes increasingly time-consuming. Before too long, you start to notice performance issues in accessing stored data. This is because each machine's data is distributed throughout various storage areas on the disk. This problem could have been avoided by reserving a consolidated portion of the disk to each VM from the beginning, as illustrated in the bottom portion of Figure 2-4. This way, each VM stores its data in consecutive locations on the disk, making it faster and easier for each VM to locate its stored

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

data on the disk. When configuring multiple VMs on the same host, carefully consider which of these options might be best for your situation.

You've found here how the processing, memory, or storage resources allocated to a VM can be very different from the physical resources available on the host. Similarly, when you move these VMs into the cloud, you'll find even more built-in flexibility. Cloud computing services are known for their **elasticity**, meaning that a service can be expanded and contracted as needed. For example, a VM might be configured with up to 500 GB of storage space. As it approaches that maximum, its allocated storage space might be set to auto-scale to a larger number, ensuring the VM continues to operate with optimal efficiency. Some of the cost savings of cloud services are due to the difference between the initially proposed resources for a service versus the available potential capacity of these services, which are auto-scaled up or down as needed.

Network Connection Types

Each VM on a host can have one or multiple virtual NICs (network interface controllers), or **vNICs (virtual NICs)**, that can connect the VM to other machines, both virtual and physical. The maximum number of vNICs on a VM depends on the limits imposed by the hypervisor. For example, VirtualBox allows up to eight vNICs per VM. Upon creation, each vNIC is automatically assigned a MAC (media access control) address.

A vSwitch (virtual switch) in the host's hypervisor connects the vNICs to a network. One host can support multiple virtual switches, which are controlled by the hypervisor. Each VM can be configured to network with other machines on the physical network or only to its host and other VMs on the same host. The most common of these networking modes (Figure 2-5) are as follows:

- **Bridged mode**—As illustrated in Figure 2-5a, a vNIC in bridged mode accesses a physical network using the host machine's physical NIC. In other words, the virtual interface and physical interface are bridged. The vNIC obtains its own IP (Internet Protocol) address, default gateway, and subnet mask information from the physical network's **DHCP (Dynamic Host Configuration Protocol)** server, which is a network service that assigns IP addresses to clients on the network. For example, if the physical network is using the network address block 192.168.0.0/24, the host might be given an address such as 192.168.0.10, and the VM might be given an address such as 192.168.0.20. Both IP addresses are located in the same network. When connected using bridged mode, a VM appears to other network nodes as just another client or server on the network. VMs in your data center that must be available at a specific IP address, such as mail servers or web servers, should be assigned bridged network connections.

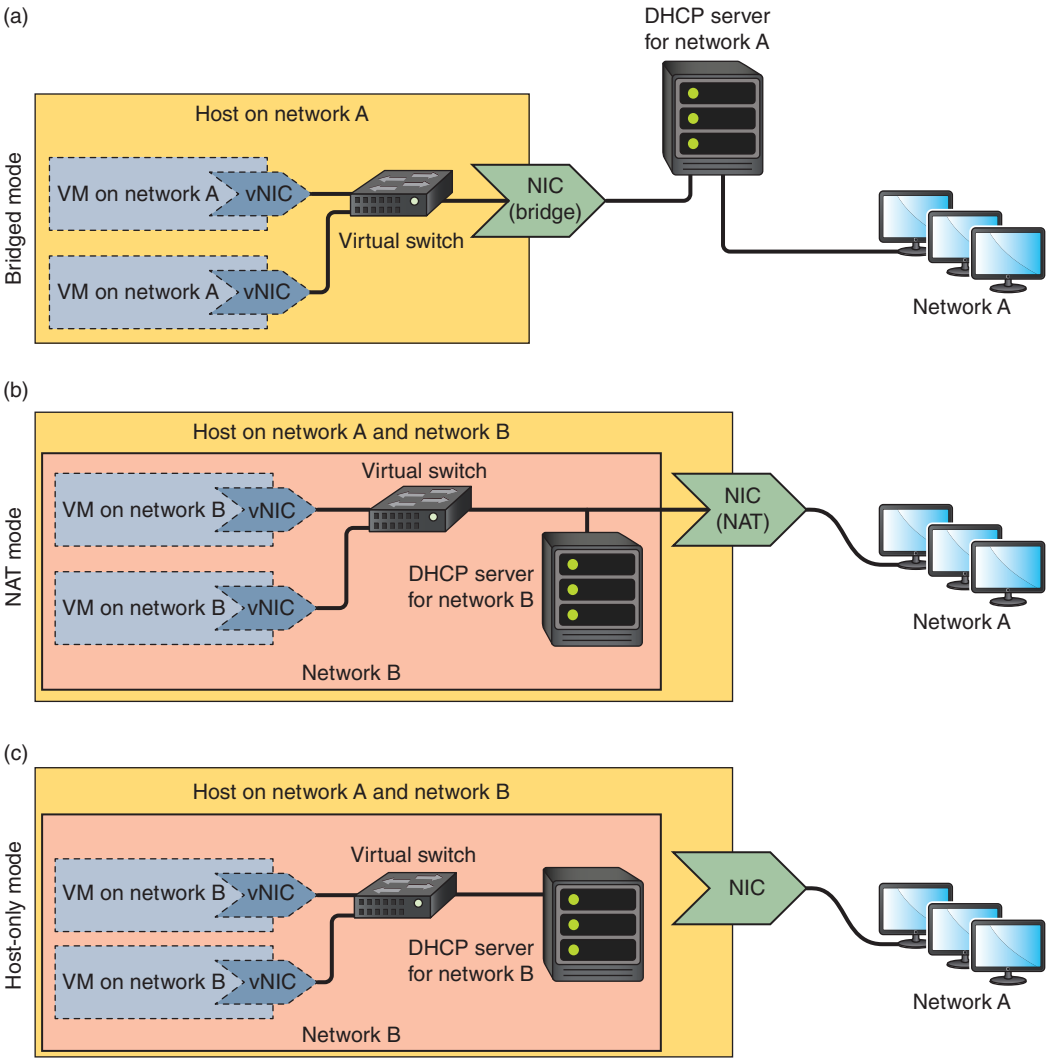
Note 2-4 | If you'd like more information about how DHCP works, use your Cengage Unlimited subscription to go to *CompTIA Network+ Guide to Networks*, 10th edition, Chapter 3, and read the section titled "IPv4 Addresses."

If you don't have a Cengage Unlimited subscription, you can find more information at cengage.com/unlimited.

- **NAT mode**—As illustrated in Figure 2-5b, a vNIC in NAT (network address translation) mode relies on the host machine to act as a NAT device. In other words, the VM obtains IP addressing information from its host (rather than from a server or router on the physical network), and the NAT feature performs address translation on communications between the VM and the physical network. To do this, the hypervisor must act as a DHCP server. For example, if the physical network is using the address block 192.168.0.0/24, the host might be given an address such as 192.168.0.10. However, the VM would be given an address from the host's internal, virtual network. For example, if the virtual network's address block is 172.16.0.0/24, the VM might be given the address 172.16.0.10. The host's physical interface and the VM's virtual interface are on two different networks. Despite having an address from the internal network, a vNIC operating in NAT mode can still communicate with other nodes on the network and vice versa. However, other nodes communicate with the host machine's IP address to reach the VM; the

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-5 The network mode determines which network the VM is part of and which devices (physical or virtual) it can reach



VM itself is invisible to nodes on the physical network. The NAT network connection type is appropriate for VMs that do not need to be accessed at a known address by other network nodes. For example, virtual workstations that are mainly used to run standalone applications, or that serve as test beds for application development or OS installations, are good candidates for NAT network connections.

- **Host-only mode**—Figure 2-5c illustrates host-only mode where VMs on one host can exchange data with each other and with their host, but they cannot communicate with any nodes beyond the host. In other words, the vNICs never receive or transmit data via the host machine's physical NIC. In host-only mode, as in NAT mode, VMs use the DHCP service in the host's virtualization software to obtain IP address assignments from the internal network address block. The host's physical interface still receives an address from the physical network's address block, but VMs cannot use the host's physical interface to communicate with the outside world. Host-only mode is appropriate for test networks or if you simply need to install a different OS on your workstation to use an application that is incompatible with your host's OS. Host-only networking is less commonly used than NAT or bridged mode networking because it won't work for virtual servers that need to be accessed by clients on the local network or for virtual workstations that need access to LAN (local area network) or WAN (wide area network) services such as email or webpages.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

1. Which of the following license types do you use when you purchase and install Microsoft Office on your personal computer?
   - **a.** Core-based license
   - **b.** Volume-based license
   - **c.** Socket-based license
   - **d.** Client access license
2. Your hypervisor can allocate use of the host's GPU to only one VM at a time. Which deployment technique is the hypervisor using?
   - **a.** Host-only
   - **b.** Pass-through
   - **c.** Shared
   - **d.** Virtual
3. Which network mode requires a VM to receive an IP address from the physical network's DHCP server?
   - **a.** NAT
   - **b.** Dynamic
   - **c.** Bridged
   - **d.** Host-only

Check your answers at the end of this chapter.

Section 2-2: VMs in the Cloud

Certification

1.4 Compare and contrast storage resources and technologies.

1.7 Compare and contrast virtualization concepts.

1.8 Summarize cost considerations related to cloud usage.

2.3 Summarize aspects of cloud migration.

2.5 Given a set of requirements, provision the appropriate cloud resources.

6.1 Given a scenario, troubleshoot deployment issues.

Now that you've learned the basics of how virtualization works on-prem, let's take some of these concepts to the cloud. While virtualization is not required in order to host cloud services, it is by far the more common method of creating and providing cloud resources (as compared to hosting cloud services directly on physical hardware). Virtualization supports the cloud's elasticity and scalability features, allows for workload migration, and increases resource **resiliency**, which is the ability of a system to withstand and compensate for failures or outages while continuing to function.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

VM Instance Types

In the cloud, a VM **instance** is an individual VM created using one or more cloud services. For example, in AWS, VM instances are provided through the EC2 (Elastic Compute Cloud) service. In Azure, you use the Virtual machines service, and in GCP, you use Compute Engine. As you create each instance, you choose its configuration, such as number of vCPUs, amount of memory, disk storage type and size, connectivity options, network placement, and more.

When deploying VM instances in the cloud, CSPs (cloud service providers) such as AWS, Azure, and GCP offer preconfigured instance types in which CPU and memory capacity are automatically allocated in a fixed ratio. Table 2-1 compares popular classes of instance types offered by AWS, Azure, and GCP. Notice that you can determine some information about the instance size by the name of the instance type. For example, AWS uses size names to identify relative instance sizes (such as micro, medium, and large). Azure states the number of vCPUs in the instance name, as does GCP. Also, different optimizations are classed by family, such as the general purpose T2 family in AWS or the D2 series in Azure. Within each family are size combinations that generally keep the same ratio of vCPUs to memory but incrementally increase the actual sizes.

Table 2-1 Instance types for AWS, Azure, and GCP

Instance type	Samples from AWS	Samples from Azure	Samples from GCP
"General purpose" or "standard" for testing, small to medium databases, and low to medium traffic web servers	t3.medium (2 vCPUs and 4 GiB memory) m8g.2xlarge (8 vCPUs and 32 GiB memory)	Standard_D2_v5 (2 vCPUs and 8 GB memory) Standard_A4m_v2 (4 vCPUs and 32 GB memory)	e2-standard-2 (2 vCPUs and 8 GB memory) t2d-standard-8 (8 vCPUs and 32 GB memory)
"Compute optimized" or "high-CPU"	c8g.2xlarge (8 vCPUs and 16 GiB memory)	Standard_F8s_v2 (8 vCPUs and 16 GB memory)	c2-standard-8 (8 vCPUs and 32 GB memory)
"Memory optimized" or "high-memory"	r8g.2xlarge (8 vCPUs and 64 GiB memory)	Standard_E8s_v6 (8 vCPUs and 64 GB memory)	m1-ultramem-40 (40 vCPUs and 961 GB memory)

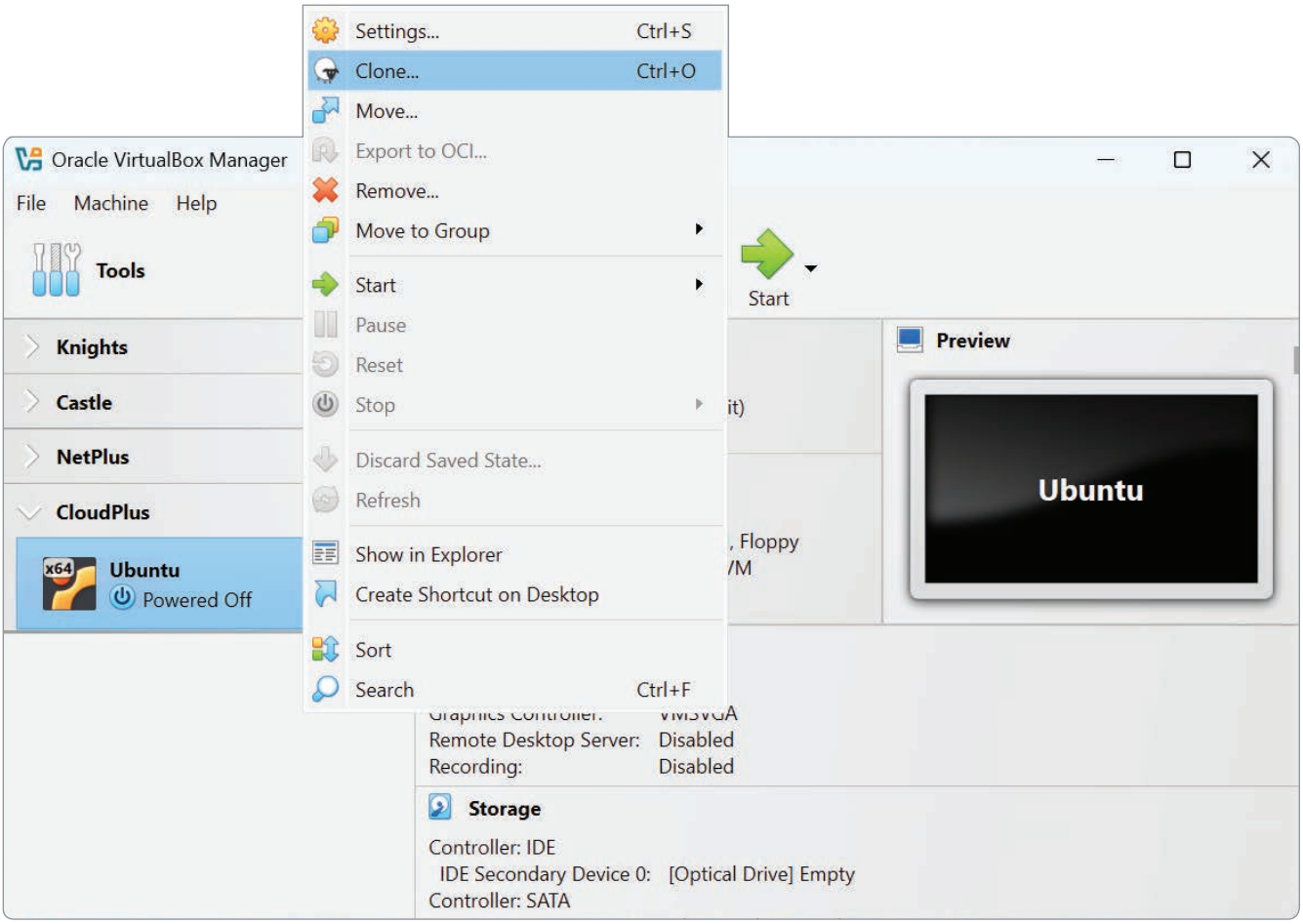
Note 2-5 | Although GB and GiB are sometimes used interchangeably, they are not necessarily the same thing. GiB stands for gibibyte, which is similar to GB (gigabyte) but is calculated by powers of two instead of multiples of 10. For example, 1 GB is 1000 MB, whereas 1 GiB is 2^{10}, or 1024 MiB (mebibytes). Still, according to some standards (such as JEDEC, the Joint Electron Device Engineering Council), GB might still be used to refer to the power-of-two calculation, such as 1024 MB for 1 GB. GCP, for example, uses the "GB" unit to identify memory usage and states in their pricing information that they calculate 1 GB as 2^{30} bytes. In contrast, AWS uses the GiB unit of measurement throughout their website. Ki (kibi-), Mi (mebi-), and Gi (gibi-) are called binary prefixes.

Cloning VM Instances

Either on prem or in the cloud, a VM can work as a standalone system, or several can work together in clusters. When adding new nodes to a cluster or creating a backup of a standalone instance, one approach to creating VM copies is to clone the instance. A VM **clone** is essentially a copy of an existing VM's design and data. Cloning is used to create many identical VMs for a cluster, to copy VMs to different locations, or to create backups of VMs. As an example, Figure 2-6 displays the option to clone a VM in VirtualBox on a local computer.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-6 Clone a VM to create a copy of it



Source: Oracle Corporation

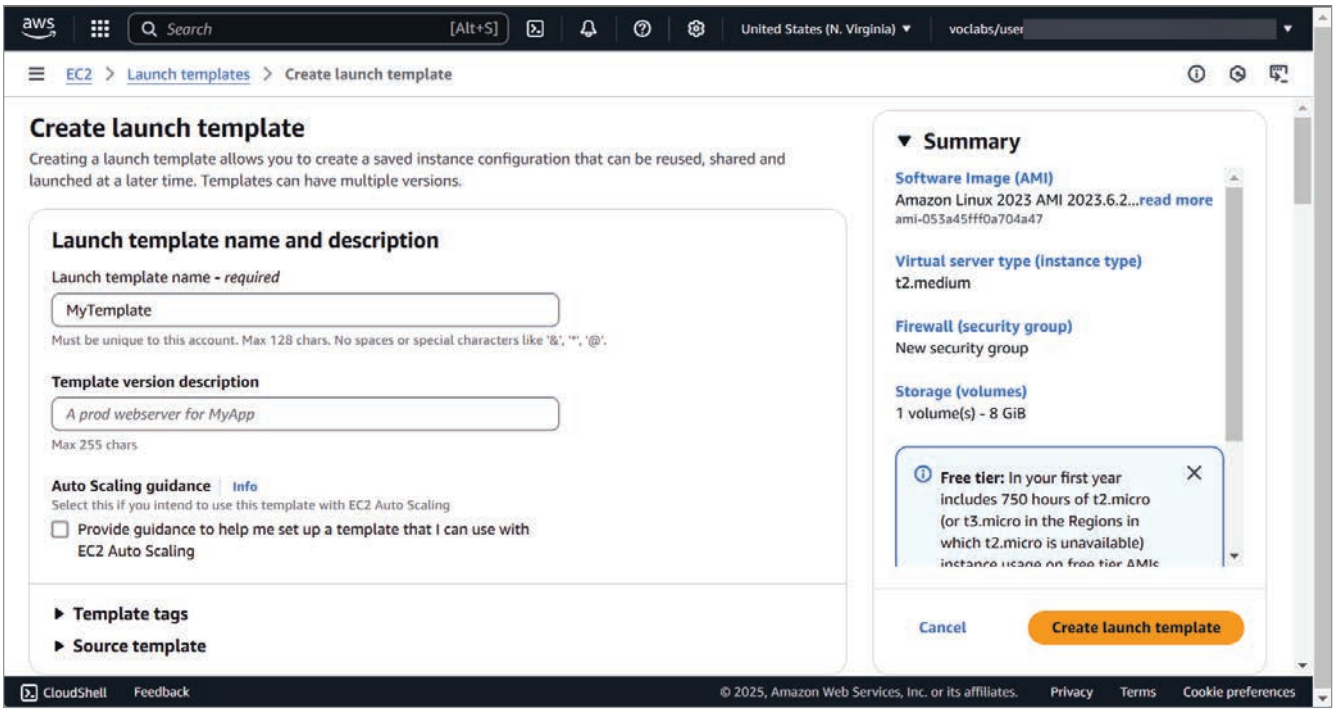
Most hypervisors provide built-in tools for cloning VMs. Although cloning doesn't work quite the same way in the cloud as it does in a hypervisor, most cloud platforms also offer built-in options for copying an existing VM or creating multiple copies of identical VMs from a template or image. For example, in AWS, this process begins by creating an AMI (Amazon Machine Image) with the desired OS, OS settings, and installed applications. You can create many EC2 instances from the same AMI, which ensures their OS settings and applications will be identical. However, keep in mind that many other decisions must be made for each created instance, such as the instance type, key pair, security group, and network settings. All these configuration options can be defined in a launch **template** (Figure 2-7).

In AWS, an EC2 template determines the following parameters:

- **AMI ID**—A template determines the AMI, which applies a standard template for the root storage volume. This determines, among other things, the instance's OS. You can use a prebuilt image provided by AWS, choose a shared image available from the community, buy an image from a third party, or create your own custom image. To create your own image, you might first create an instance from a standard image, make configuration changes to the instance, and then create a new image from the customized instance. This custom image is sometimes called a golden image and should be kept updated, especially with security updates and configurations. Other cloud platforms offer similar options for creating VMs from custom images.
- **Instance type**—The instance type determines the instance's virtual hardware resource allocations, including vCPUs, RAM, and attached storage volumes. The instance type might also determine the instance's network bandwidth, or the vNIC might be configured separately.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-7 Creating a launch template in AWS



Source: Amazon Web Services, Inc.

- **Network and security settings**—Network and security settings can include the instance's subnet, key pair for authentication, security group for firewall type protection, vNICs, and IP configuration for each interface.
- **Storage volumes**—The AMI includes parameters to create a boot storage volume. You can also specify additional storage volumes and related settings, such as size, volume type, and whether it's encrypted.

Additional parameters are also available within the launch template. You might decide to specify only some of these options in the launch template but not all of them. For example, you might want to set up a different process for assigning a key pair to each instance created from the template rather than using the same key pair every time.

In Figure 2-8, the configuration screen displays options for creating an EC2 template. Alternatively, you can create a launch template from an existing EC2 instance. When using templates, be aware that a template could cause problems if it is misconfigured. Especially when you pull a template from an outside source, check the configuration carefully.

Note 2-6 | Part of your EC2 backup strategy might include making copies of the EBS (Elastic Block Store) volumes attached to your instances. A copy of a storage volume is called a snapshot. You can create and update volume snapshots separately, or you can include one or more snapshots in an AMI.

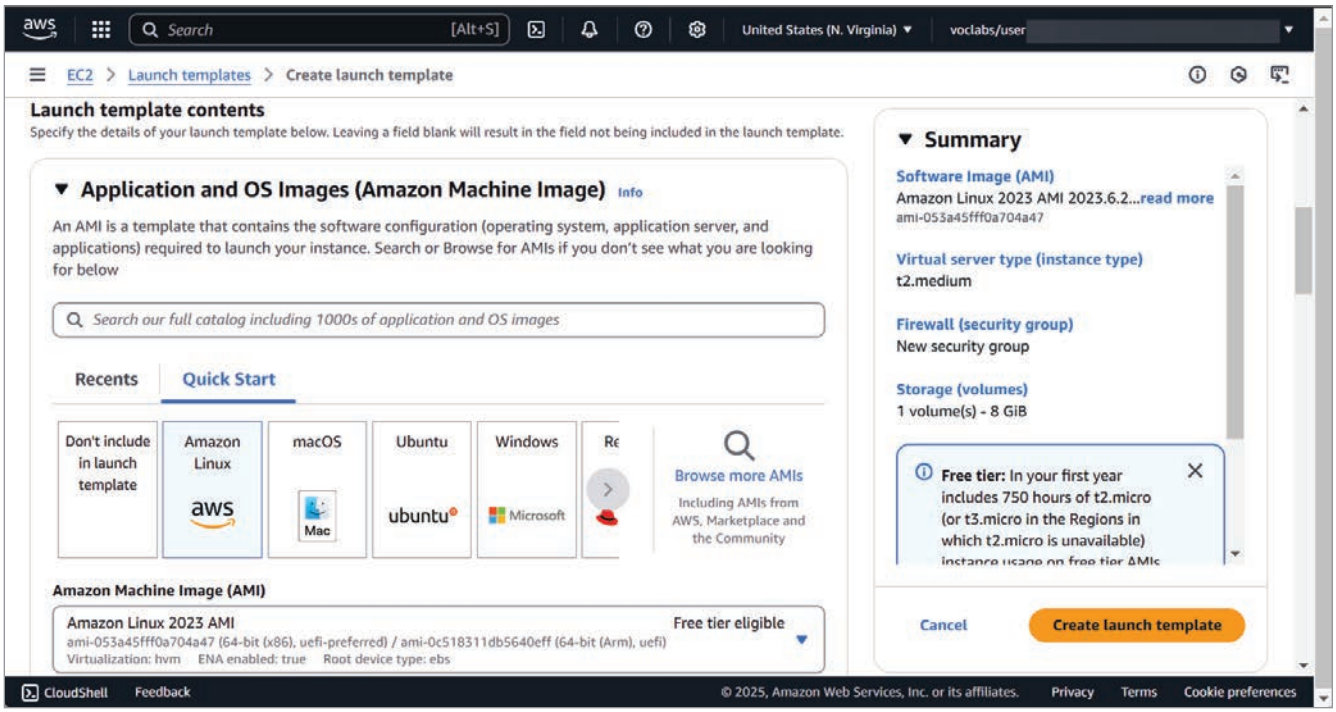
Billing Models

Cloud customers can often get commitment-based discounts and discounts for providing full payment upfront for anticipated resource usage. The following list explores types of discount plans, called billing models, that many cloud providers offer, especially for compute resources such as VM instances:

- **On demand**—Also called **pay-as-you-go**, in this scenario, you pay for the resources you use as you use them with no commitment for future use. For example, you might create a VM instance to host a web

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-8 Specify launch template parameters



Source: Amazon Web Services, Inc.

server for a special event, and you pay for that server when it's running. However, you don't intend to keep the web server after the event is finished, so the standard, on-demand rate for the VM instance likely provides the most cost-efficient billing option.

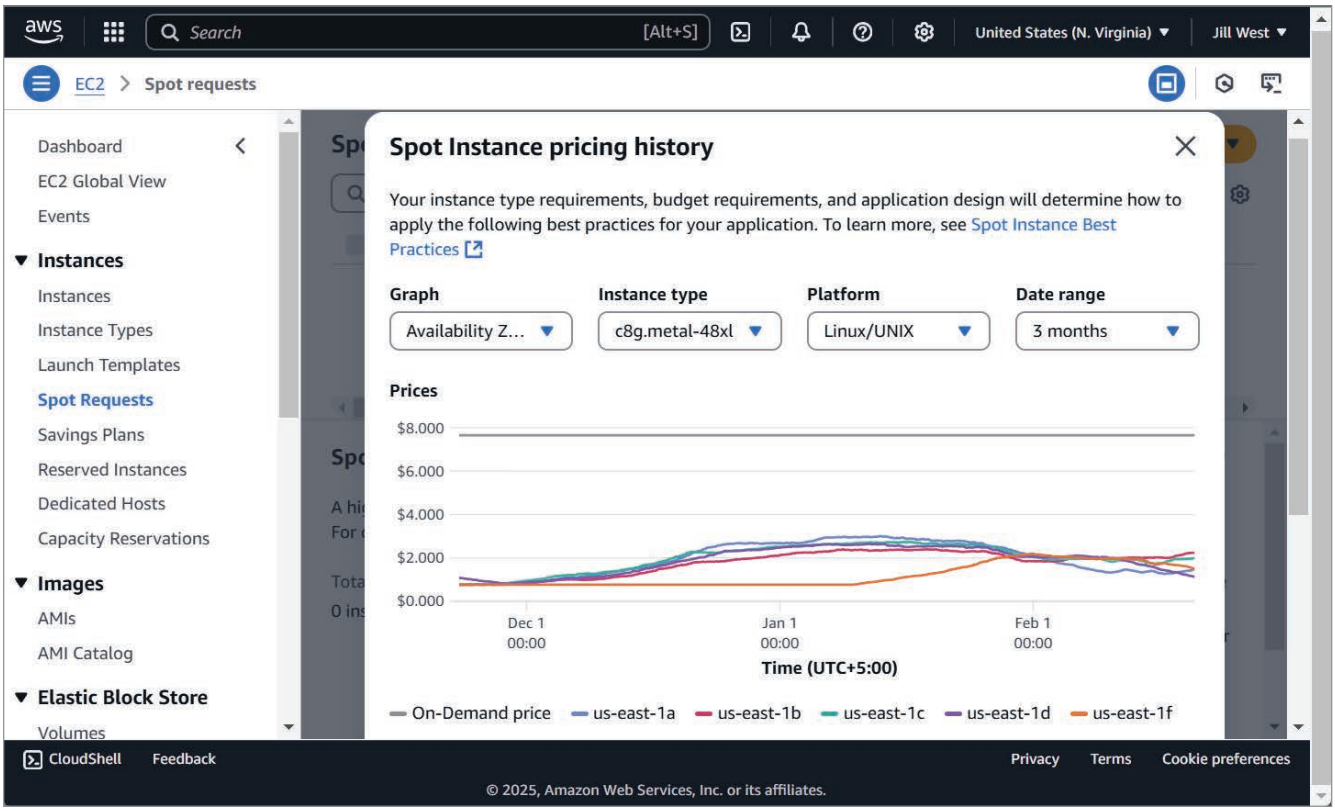
- **RI (reserved instance)**—Also called a reserved resource, in this case, you pay a lower rate in exchange for committing to a certain amount of future use of a resource. For example, suppose you need a VM instance to host a database, and you know you'll need the database to exist for a minimum of three years. AWS, for example, offers a lower rate on instances if you commit to using a specific instance configuration (that is, instance type and OS) for one or more years. You can get an even lower rate if you pay part or all of this cost up front. Reserved resources best fit with workloads whose utilization parameters are predictable over time. With AWS, an RI configuration is set at the beginning of the reservation period and cannot be changed. More recently, AWS has begun offering Savings Plans as an alternative to RIs. Savings Plans include more flexibility for adapting a customer's commitment to use compute resources over the course of the commitment period. For example, with a Savings Plan, you can move your compute commitment and price savings to a different instance type, to a different region, or even to a different compute service (such as moving from EC2 to Lambda, which you'll learn about later in this chapter).

Exam Tip | Another option for saving money on EC2 instances in AWS is the Reserved Instance Marketplace. Suppose you committed to an instance for three years. A year later, you decide you would rather host the database in a different database service, such as DynamoDB, or you decide you need a larger instance to meet your growing needs. You've already committed to the instance, but now you won't need it. You can instead post the instance to the Reserved Instance Marketplace for another AWS customer to take over the commitment. If you need a reserved instance for a new project, you can shop the Reserved Instance Marketplace to determine if an offering might meet your needs for a lower cost or for a shorter time than an RI provided directly from AWS.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Spot instance**—CSPs typically maintain sufficient hardware capacity for many more VM instances than what is running at any given time. There are times when more of their capacity is used, and other times when much less of their capacity is being used. As demand for VM instances waxes and wanes, the relative cost for each instance can also fluctuate. If you have a workload that can be started when costs are low and stopped when costs rise, this workload can be run much less expensively using a spot instance. Figure 2-9 shows how spot prices on AWS can fluctuate over time, potentially providing significant savings. However, the workloads run on spot instances must be able to tolerate interruptions with little warning (such as two minutes) when spot prices exceed the maximum threshold you set for those instances. Workloads that are typically well suited to spot instances include stateless web servers, containerized workloads, HPC (high-performance computing), big data processing, and media rendering processes. You'll learn more about serverless services and containers later in this chapter.

Figure 2-9 Spot instance pricing can offer significant savings over on-demand instances



Source: Amazon Web Services, Inc.

- **Dedicated host**—Not so much a billing model as a configuration option, this feature increases the cost of the instance while also increasing its security. A dedicated host ensures that a customer's VM instances are run on a physical host that is dedicated to that one customer—that is, the physical host is not shared with other customers. This arrangement can be especially helpful for ensuring compliance with software license restrictions or government regulations. A dedicated host also provides the customer with increased visibility into the host's physical configuration, such as the number of sockets and the number of physical cores (information that is sometimes essential to software license management).

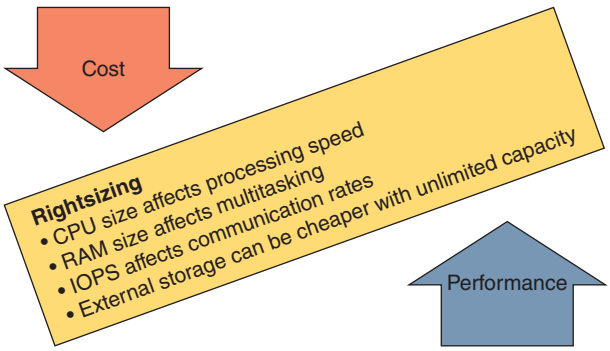
Exam Tip | In preparation for the CompTIA Cloud+ exam, don't just memorize the definitions of these billing models. Additionally, consider scenarios that would best fit each billing model. What key words in a scenario description would indicate that one billing model would be a better fit than another billing model? What features of a billing model would make it a poor fit for potential scenarios? You might try creating some practice exam questions of your own to help you think through these possibilities.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Resource Allocation Factors

While cloud platforms can support huge deployments of virtual hardware resources, budget and other considerations impose certain limitations on these deployments. The trick is to find an optimized balance between meeting performance demands and keeping within budgetary or other restrictions (Figure 2-10). This balance is called **rightsizing**, which helps avoid sizing issues. Oversized resources can cost too much to run, and undersized resources can fail in times of high demand. Rightsized resources accurately reflect the needs of the demands and restrictions placed on them.

Figure 2-10 Rightsizing balances between cost and performance



Many factors directly affect the cost efficiency of virtualized CPU and memory resources for a particular deployment. Let's examine a few of these factors:

- **Security**—As you learn about CPU and memory virtualization technologies, you'll find there are many efficiency advantages to relying on a shared compute environment in the cloud. This arrangement sometimes creates complications for security concerns, however. Organizations might be required to keep some types of data completely on-prem. In other cases, the data can be stored in the cloud but only on dedicated equipment that is not shared with other tenants. Although often more expensive, a dedicated compute environment can also be kept more secure.
- **High availability**—Availability of a network or service is calculated as total uptime over a given period and is expressed as a percentage, such as 99.9%. **HA (high availability)** is a network's or service's maximized potential for being available consistently over time and is often referred to as the number of nines referenced in the percentage. For example, 99.99% HA is called "four nines availability." For virtualized systems, HA features are often built into the hypervisor to manage VMs and their resources. The hypervisor can restart a failed VM or move a VM to another host in the event of hardware failure. In large data centers, multiple host machines working together in a cluster provide failover coverage for guest VMs. The hosts collectively offer more CPU and memory capacity than is needed to run the virtual infrastructure, allowing for sufficient resources to support the VMs if a host fails. Similar configurations can be deployed in the cloud.

Note 2-7 | One way to consider availability is by measuring a system's uptime, which is the duration or percentage of time it functions normally between failures. Realistically, what is actually measured is often downtime, which is the average amount of time a system is out of service during a given period. For example, a system that experiences 99.99% uptime is unavailable, on average, 52 minutes and 35 seconds per year. In contrast, five nines availability averages to only 5 minutes and 15 seconds downtime per year.

- **Disaster recovery**—**DR (disaster recovery)** strategies and tools are designed to facilitate an organization's recovery from an adverse event. In the cloud, these tools and strategies tend to be greatly simplified due to its virtual nature, and many DR-related processes can be fully automated. Compared to physical systems, whose backups are expensive to store and maintain in hot site, warm site, and cold site arrangements, copies of virtual systems can be stored online (even on a different cloud platform from the active system), kept updated alongside running systems, and automatically activated when needed. With data storage hosted in separate services from compute systems, virtual servers can be

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

quickly re-created not just during a full-scale emergency, but also in response to declining performance or other difficult-to-diagnose problems. Virtual systems are considered disposable, meaning it's faster and easier to replace them than to try to fix them. Still, when designing backup solutions as part of a DR plan, consider the additional processing capacity needed for VMs to maintain these backups.

- **Energy savings**—On-prem virtualization provides significant energy savings for an organization compared to running an on-prem data center with no virtualization, as less hardware is needed to support a complex infrastructure. Expanding into the cloud further increases energy savings as the hardware power and cooling burdens are shouldered by the CSP. However, this energy savings can be offset by the increased need for ISP (Internet service provider) bandwidth or dedicated lines to CSP data centers.
- **Cost considerations**—The rates charged for VM instances in the cloud are often determined by the number and type of vCPUs as well as the amount of allocated RAM. Other factors include storage type and size, payment schedules, and added features:
  - **Storage**—Storage fees are calculated according to disk type such as SSD (solid-state drive) or HDD (hard disk drive), input/output performance, and provisioned space, multiplied by time. In calculating costs for storage volumes, you pay for the *provisioning* of space, not the actual space used by the data you're storing on these storage volumes. So if you allocate a 2 TB virtual SSD for a VM instance, you pay for the 2 TB the entire time that SSD is allocated, even if you have no data stored on it.
  - **Features**—Added features such as burstable processing, attached GPUs to accelerate processing, or allocated public IP addresses can affect overall rates for an instance type. For example, if you need a static public IP address for a web server (that is, a public IP address that stays the same even when the instance is power cycled), most public cloud platforms charge an extra fee to reserve a public IP address, which you can then assign to the instance. In another example, you can require that the instance run on hardware dedicated only to your organization's use. Dedicated tenancy increases the cost of the instance, but this expense might be worth it if the instance contains sensitive information requiring an unshared host to increase security.
  - **Network placement**—Data coming into your cloud's network, called **ingress traffic**, generally is free unless using certain services such as a load balancer. Traffic leaving an instance's network, called **egress traffic**, might be charged, depending on where that traffic is going. For example, if the traffic is leaving the CSP's own infrastructure onto the public Internet, typically this traffic incurs charges. Position and organize instances carefully to minimize the cost of network traffic.
  - **Cost model**—Most of the major CSPs are moving toward per-second billing, although calculations vary (such as requiring a minimum of one-minute billing before counting to the nearest second).
  - **Commitment-based discounts**—Committing to a service for a longer period allows for discounted rates. Some commitment plans allow for early cancellation or shifts to different services, and varying levels of ease for accomplishing these changes.
  - **Payment schedule**—Some CSPs offer discounts for full payment upfront for committed usage.

Affinity

Most of the virtualization scenarios discussed earlier in this chapter consist of a single physical host, or standalone host, supporting multiple VMs. What if you want to distribute several VMs across multiple physical hosts? This would offer several advantages, including the ability to scale up your deployment and the reliability of backup hosts taking over in the event of a host failure. When done on-prem in a LAN, this scenario is sometimes called a CAB (cluster across boxes). It relies on some key configurations between the physical

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

hosts and shared storage provided by a SAN (storage area network). While cloud data centers use a more wide-scale version of this concept, the same idea applies—instances might reside on any of several physical hosts collaborating to provide the virtualized environment.

Whether you're responsible for configuring these clustered hosts on-prem or if you're creating VM instances in the cloud, you can choose whether certain VMs reside on the same host or different hosts. Guest VMs residing on the same host are part of an **affinity** group. You accomplish this restriction by setting an affinity rule on the VM instances. If, instead, you prefer these VMs *not* share the same physical host, such as for availability reasons, you can set an **anti-affinity** rule.

When the hypervisor starts these VMs, it will attempt to abide by the affinity or anti-affinity rule, which might be set as required or preferred:

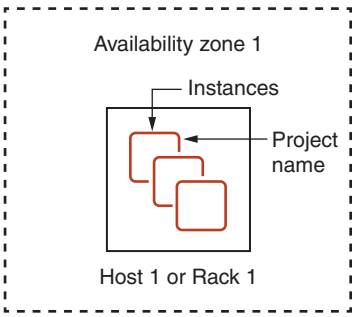
- **Required**—If the affinity or anti-affinity rule is required and the rule cannot be observed (for example, if there are not enough physical hosts available to start the number of requested VMs on separate hosts), then the requested instances are not started.
- **Preferred**—If the affinity or anti-affinity rule is only preferred, the hypervisor will give its best effort at complying with the rule but will still start the VMs even if the rule cannot be fully observed.

Public cloud platforms offer multiple methods to control the affinity of cloud instances. One of the easiest methods is to set a desired geographical location of an instance. In AWS, for example, this is handled through regions and availability zones. An **AZ (availability zone)** represents one or more physical data centers, and a **region** is a collection of AZs in a single geographic area. When you configure an EC2 instance in AWS, you can choose its region and its AZ. If you want another EC2 instance in a different data center to improve availability, you can configure the next instance in a different AZ, which guarantees the second instance is hosted on a different collection of power, cooling, and network infrastructure. Other cloud platforms offer similar options based on geographical locations of data centers.

Most CSPs, such as AWS, automatically attempt to spread your instances across physical hosts even when all these instances are hosted within the same data center. If you need to influence this distribution in AWS and other cloud platforms, you can use **placement groups**, which are logical groups of instances that are all hosted in the same AZ. AWS offers three types of placement groups:

- **Cluster**—A **cluster** is a group of devices or resources that appear as a single device or resource to the rest of the network. In AWS, instances in a cluster placement group are all hosted as closely together as possible within the AZ to support low-latency communications (Figure 2-11). This placement scenario is commonly used in HPC clusters in support of applications that require high-speed, parallel operations.

Figure 2-11 Cluster placement group



- **Spread**—This arrangement distributes each instance within a small group (no more than seven) across diverse physical hosts (Figure 2-12). This is commonly used when both low latency and high availability are required so that, if any one instance's host fails, the others will not be affected by the same host failure. Each instance's physical host relies on separate network connections and power sources within the same data center; however, their proximity within a single data center still minimizes latency that would be caused by traveling longer distances to a different data center.
- **Partition**—These instances are grouped into partitions (no more than seven partitions per AZ), and then the partitions are distributed across different physical host systems (Figure 2-13). Each partition contains multiple instances, and each partition's instances reside on different hosts in different racks than the instances in another partition. With a partition placement group, you can also create partitions in several AZs within a region.

Figure 2-14 shows the option to add an EC2 instance to a placement group as the instance is being configured.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-12 Spread placement group

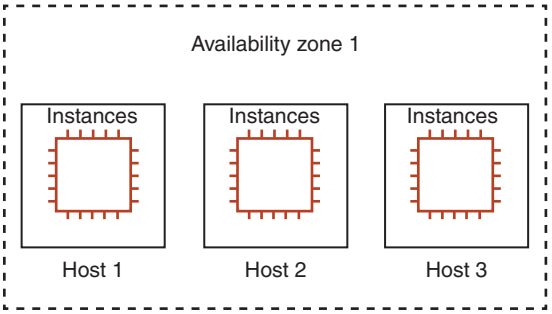


Figure 2-13 Partition placement group

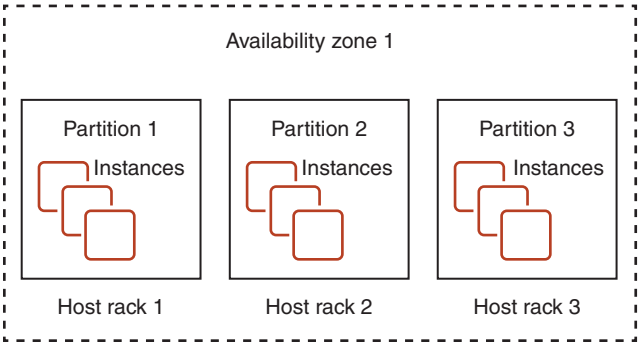
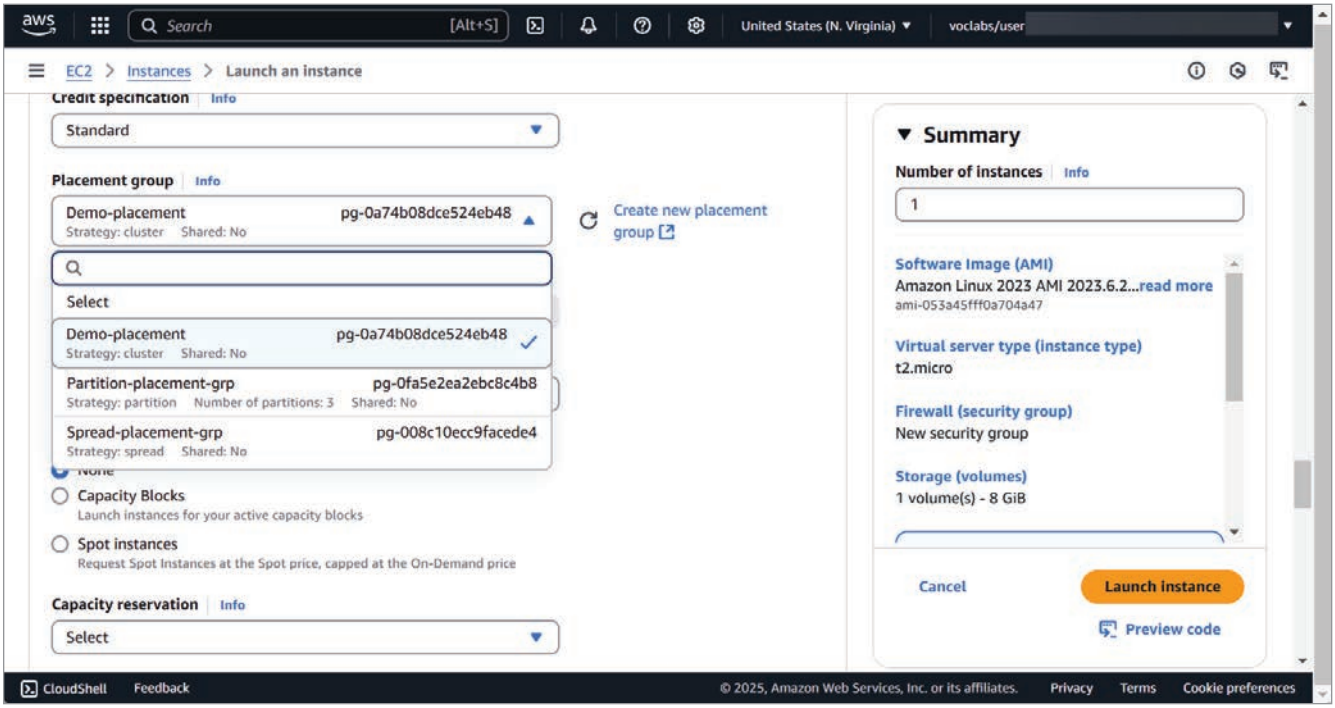


Figure 2-14 This demo placement group uses the cluster strategy



Source: Amazon Web Services, Inc.

Self-Check

4. Which configuration selection determines the amount of RAM a cloud VM can use?
   a. Instance type
   b. Image
   c. Storage volume
   d. Network bandwidth

5. Which configuration parameter can ensure all the VMs for a deployment reside on the same physical host?
   a. Partition
   b. Cluster
   c. Affinity
   d. Template

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6. Which principle ensures you can consistently access a cloud resource?
   a. Placement
   b. HA
   c. DR
   d. Cluster

Check your answers at the end of this chapter.

Section 2-3: VM Alternatives

Certification

1.1	Given a scenario, use the appropriate cloud service model.
1.5	Explain the purpose of cloud-native design concepts.
1.6	Compare and contrast containerization concepts.
1.10	Compare and contrast methods for optimizing workloads using cloud resources.
3.4	Given a scenario, manage the life cycle of cloud resources.
5.2	Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.
5.4	Explain the importance of tools used in DevOps environments.

No discussion of compute in the cloud would be complete without some coverage of alternatives to VMs for hosting applications and other services. Two popular alternatives are serverless services and containers. Let's first consider serverless technology because serverless services are sometimes used to support containers.

Serverless Computing

A cloud-native, streamlined technology for hosting cloud-based applications is serverless computing ("serverless" for short), which is a type of serverless service. Despite the name, servers are still involved with serverless computing. However, the cloud customer doesn't have to bother with configuring or managing the server—the CSP does this part.

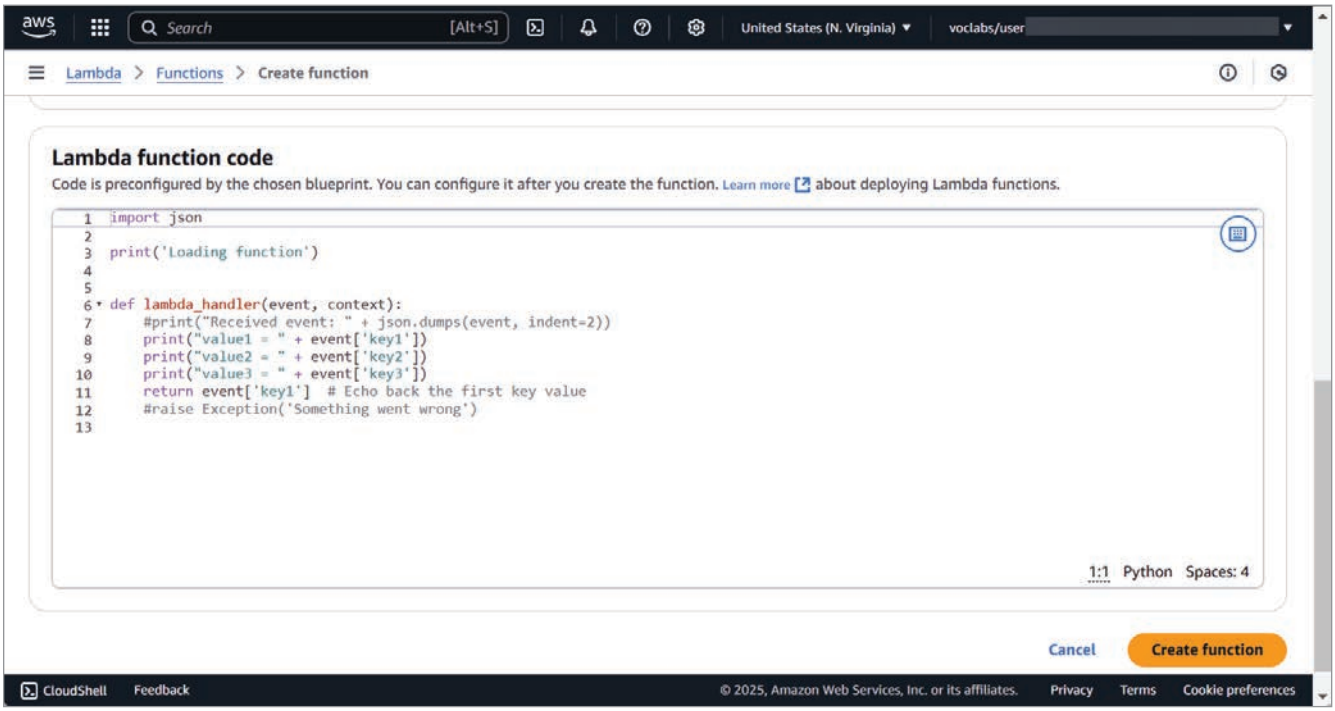
When running serverless compute, the CSP offers short-term or purpose-driven use of a server without access to the server environment. For example, the time needed to run code in AWS's Lambda can often be measured in microseconds (Figure 2-15). This reduces overall costs to the consumer and transfers responsibility for managing the server (such as provisioning, patching, and maintenance) to the CSP. This paradigm is sometimes called **FaaS (Function as a Service)**, which refers to the customer managing individual processes or functions in the cloud rather than managing an entire OS environment.

Other serverless services can store data, run databases, or handle security functions. Popular examples of serverless services include the following:

- **AWS**—Lambda for running code, S3 (Simple Storage Service) for data storage, and DynamoDB for NoSQL databases
- **Azure**—Cosmos DB for databases, Bot Service for intelligent serverless bots, and Blob Storage for data storage
- **GCP**—Cloud Run for managed container deployment, Cloud Run functions for running code, and App Engine for hosting web and mobile apps

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-15 Run code directly in Lambda without managing a server



Source: Amazon Web Services, Inc.

Serverless services are ideal for many databases, backup or data transfer tasks, and apps or code that don't need to run continuously, such as IoT (Internet of Things). Serverless is often used for backend tasks behind the scenes of complex websites or applications. It tends to be cheaper for many types of jobs, is more cloud-native in design (which you'll learn about later in this course), and is often more secure.

Applying Concepts 2-1: Create a Serverless Resource

Estimated time: 15 minutes
Objective 1.1: Given a scenario, use the appropriate cloud service model.
Resources:

- Internet access
- An account for a public cloud platform, such as AWS, Azure, or GCP

Description: Many public cloud platforms provide built-in guidance on how to build a serverless function. Typically, at a very basic level, this is a simple process that is easy to learn. For example, to create an AWS Lambda function, you have the option to use a pre-built blueprint that lets you practice with the process before authoring your own function.

To experiment with a serverless service in your selected cloud platform, complete the following steps:

1. Research your selected platform's serverless services, and choose one of the more popular and mainstream services to practice with in this activity. For example, in AWS, you might choose Lambda. Comparable services in other platforms include Azure Functions and Google Cloud Functions. A serverless compute service is probably the easiest of these services for this activity; however, you could also choose a serverless database service, a serverless storage service, or some other serverless service. Do some basic research on this service so you understand its purpose and the big picture of how it works. Which service did you choose?

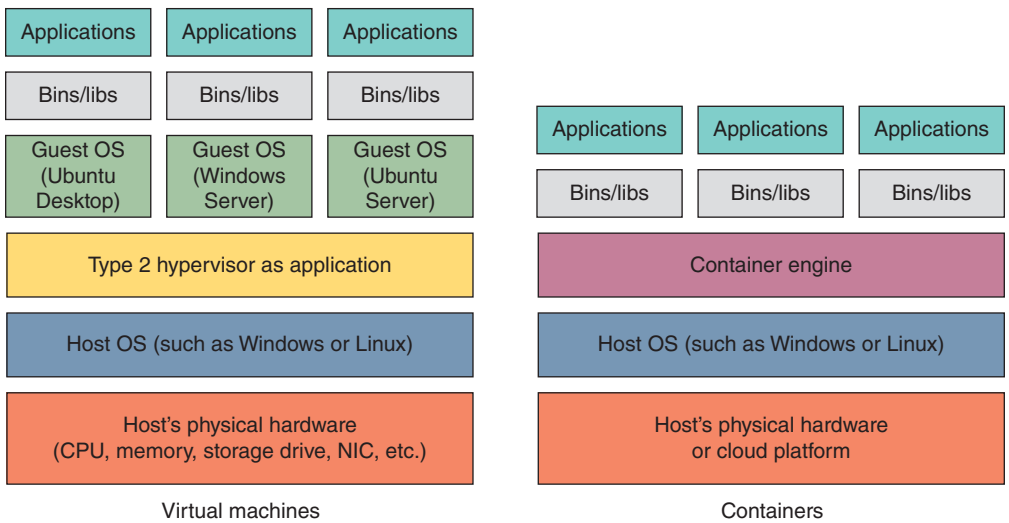
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. Find online training materials or practice activities embedded in the platform of your selected service. Which resource do you think will be the most helpful as you practice deploying a resource, such as a function, in your chosen serverless service?
3. Work through the steps to deploy a resource in this service and troubleshoot any problems or error messages you encounter. For example, if you're working with AWS Lambda, you would create a function using a blueprint, which is allowed with an AWS Academy account if you choose an existing IAM role instead of creating a new role. If you're using AWS Academy or another education-based account, you might encounter permission limitations. If the limitations are too restrictive, consider switching to a different service. Work through the process as far as you can and document your progress and limitations. When you've created the resource (or worked as far as you can before running into insurmountable permissions issues), **take a screenshot** of your progress; submit this visual with your responses to this activity's questions.
4. If you were able to create the resource, consider testing it. For example, in AWS Lambda, you create a test event, run it, and then view the execution results.
5. Delete all resources you created—good cloud hygiene will help protect you from unwanted cloud charges.

Containers

VM technology has been around for many years and has become the de facto operating standard for most data centers. Recall that a VM relies on an underlying hypervisor of some kind to manage communication between the VM's OS and the host's physical hardware. Then, applications of various types are installed inside the VM's OS. This stack is portrayed on the left side of Figure 2-16.

Figure 2-16 A container engine creates a lightweight, self-contained environment



However, the overhead created by installing and then booting an entire OS on a full VM to host applications can be problematic. Suppose you're only hosting one or a few lightweight applications on the VM. You have an entire OS with much of its potential sitting unused by the apps. You also have the processing, memory, and storage resources allocated to the VM that are rarely used to full capacity and, in fact, must maintain some buffer of allocated resources beyond the maximum amounts used at any time.

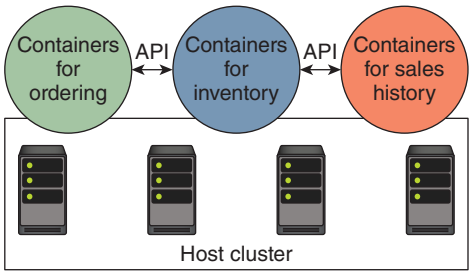
What if, instead, you could take the parts of the OS you need for that specific application and package them into a container with the app itself? What if you could use exactly the amount of hardware resources you

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

need at any given time, and nothing more, without having to boot up an entire operating system? This is the essence of a **container**, which is a lightweight, self-contained environment that provides only the essential OS services needed to run the embedded application.

As illustrated on the right side of Figure 2-16, the OS layer is run by a software platform called a container engine. The market-leading container engine is Docker Engine, and many cloud platforms offer container services that support Docker containers. In this scenario, the app draws on exactly the hardware resources it needs and, as it is moved from one environment to another, takes with it exactly the OS resources that it needs and nothing more; essentially, this is OS virtualization. It abstracts the application's processes outside of the traditional OS. This strategy makes the app more portable to be hosted in different operating environments in a more predictable way, and it makes the container more lightweight than an entire VM with its OS and apps. An advantage to this arrangement is that multiple containers can run on a cluster of hardware resources to achieve high availability and load balancing objectives (Figure 2-17).

Figure 2-17 Many containers can run on a cluster of hosts



Suppose you break down a large application into its smaller parts, called **microservices**. For example, you might separate the ordering processes of an application from the inventory process or the sales history process. Each microservice communicates with other microservices using APIs (application programming interfaces). Next, you design each part of the application to run in a separate container or a cluster of containers. Then, a cluster of physical servers can support a high number of containers; the application's microservices are spread throughout several containers to pull on physical resources (such as CPU, memory, or storage) throughout the cluster.

One of the most significant factors in an organization's adoption of containers is that software can be developed, tested, and deployed in a more stable hosting environment when each application is essentially packaged with its own running environment inside the container. Developers can test changes to the underlying code and accurately predict the effect of those changes on the application, no matter the hosting environment where it's used. Containers are more lightweight than VMs and do not require such robust hardware resources to run. They can also scale more quickly and can more easily handle distributed workloads across containers and across their hosts.

Supporting Containers

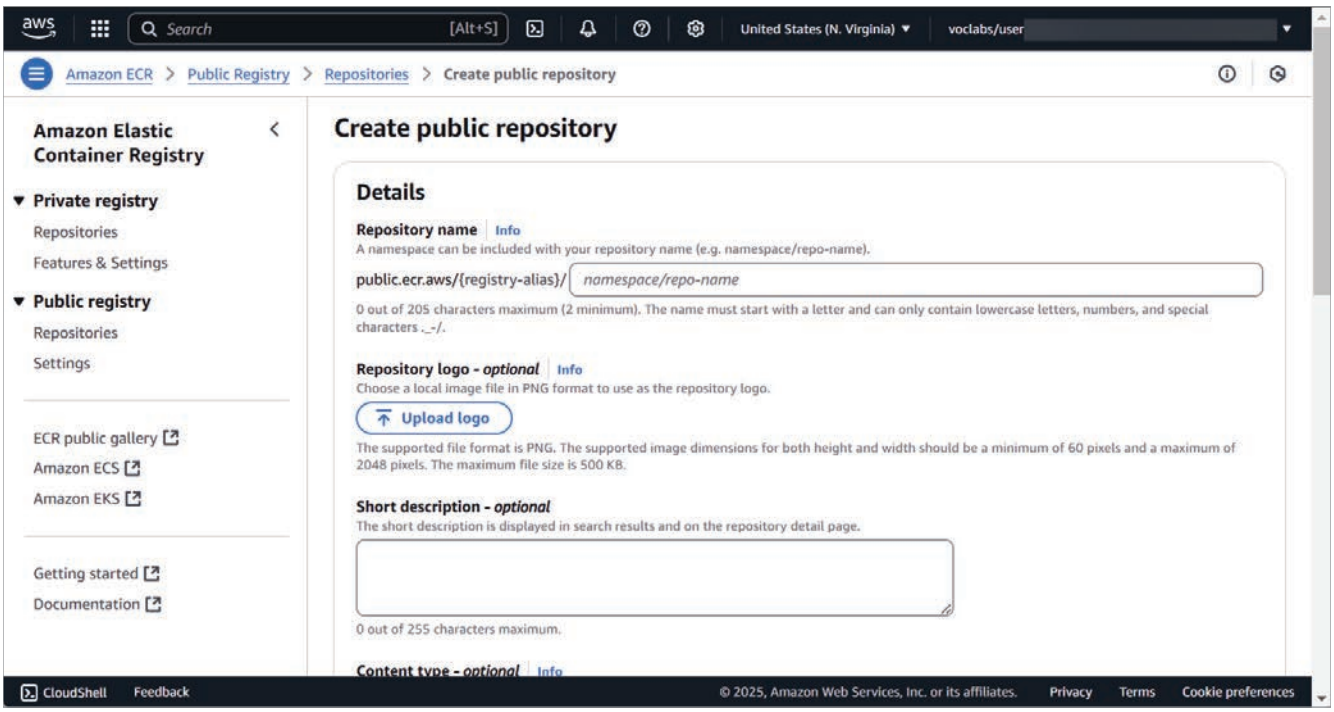
Creating and running containers is a complex process that requires specialized software. Consider the following phases of container creation and management:

- **Create a container image**—When you write code for an application to be run in a container, you include with the code some information, called environment variables, about the dependencies needed for that application to work. These dependencies are created from a specific OS, such as Ubuntu or Windows Server. For example, one environment variable determined by dependency information is the file path. Windows systems use a different file path configuration than do Linux systems, and this information is defined for the container-hosted application. Another environment variable might indicate the location of a secrets file the container needs to do its job, such as passwords or API keys. One way to do this is to store a secrets file in a volume mount. Typically, data storage attached to a container disappears any time the container is shut down; this storage type is called **ephemeral storage** because it doesn't continue to exist on its own. However, sometimes data must exist on its own so many containers can access it, even as these containers are created and deleted. A volume mount connects a container to **persistent storage**, which is a storage device that continues to exist even when the container is terminated. And one of the environment variables might give the container information on how to connect with that storage device. Alternatively, some container services offer built-in secrets storage, and the container would need to know how to find the information listed there. Once the code and all its variables and dependencies are defined, you then package this information into a fairly small file called a container image. This image will serve as a template for all the containers that will run this code.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Store the image in a repository**—Next, you deposit the image in a repository, also called an image registry, so the image can be accessed by a container management service. This registry can be public so that others can use the images or private so you have more control over who can access the images. The Docker repository is called Docker Hub, which offers personal subscribers unlimited public repositories for free with a limited number of image accesses called image pulls. In AWS, this repository can be created in the AWS ECR (Elastic Container Registry) service, which offers more free options and native compatibility with other AWS services. Figure 2-18 shows options when creating a public repository. Notice in the left pane that private repositories and public repositories are listed in separate areas.

Figure 2-18 An image repository can be public or private



Source: Amazon Web Services, Inc.

- **Run the containers**—To run the containers, you also need underlying hosts, whether physical or virtual, to support the container management tool. You can run containers on a single, standalone host or node, such as a single physical server running Docker Engine. However, as the workload volume increases, it eventually requires a cluster of hosts. This is a challenge to coordinate, especially considering that the data storage these containers access must sync across hosts. One option at this point is to upgrade to a tool such as Docker Swarm, which coordinates multiple hosts to run a larger collaboration of containers. A more sophisticated option is to use Kubernetes, which not only is significantly more complex but also offers robust configuration options. **Kubernetes** is an open-source container management service developed by Google and now available in other major cloud platforms such as AWS and Azure. Kubernetes is also sometimes called K8s, where the middle eight letters, "ubernete", are replaced by the number 8. K8s works to optimize management of container workloads and services.

You can run containers in a cloud environment, which is optimized for this purpose. For example, in AWS, you can store container images in ECR and then run containers in ECS (Elastic Container Service) or EKS (Elastic Kubernetes Service). ECS is the AWS-native, managed container orchestration service for handling containers automatically. It can be used to deploy, schedule, scale, and manage containers on top of EC2 instances. ECS creates a cluster of multiple EC2 instances to provide the services needed to host your

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

containers. Alternatively, you could create several EC2 instances yourself directly in EC2 and install something like Docker Swarm on top of those instances to host your containers. However, ECS creates instances that are optimized as members of a host cluster, and AWS provides some additional automated services in ECS that make container management easier for the customer. To accomplish this goal, each instance contains both an ECS agent and a Docker agent for monitoring and communication.

When using ECS, the customer maintains certain responsibilities for managing these EC2 instances, such as the following:

- Create the EC2 instances and join them to the ECS cluster.
- Monitor resource usage and scale resources as needed.
- Manage each host instance's OS, including security and updates.

In AWS, an alternative to EC2 for hosting containers is Fargate, which automates some of these host cluster responsibilities. When using Fargate, it will create the needed instances for your containers. Essentially, Fargate is a serverless service because the customer is not responsible for managing the underlying servers for their containers. As the customer's containers require additional resources, Fargate provisions sufficient EC2 instances to handle the demand. Fargate is more expensive per instance than the customer-managed instances in EC2, but it's also more responsive to shifting demand. This factor can ultimately be less expensive for a customer if the workload is less predictable. The number and capacity of available instances can shift more quickly in response to changing demands, which means the cost of these instances more accurately reflects the actual use of the instances.

Note 2-8 | A **managed service**, such as ECS, is a service where the cloud provider handles most of the management and upkeep of the infrastructure for that service, and the customer also retains some control over where their resources run, how they're configured, and how they scale. For example, ECS is a fully managed service because AWS handles the technicalities of the compute, network, and security configurations even as customers scale up their container deployments. The cloud customer still chooses some of the configurations, such as the compute layer used—you can choose EC2 instances to retain the most control in hosting your containers, Fargate for serverless compute, or Outposts to run containers on prem. But because the service is managed by AWS, the customer can focus more on the business logic of their system design and less on the technicalities of how to make it all work.

In contrast to serverless services, managed services offer customers more control over their deployments. Serverless services, by comparison, function without much input from the customer and generally run in short bursts only when needed, reducing or eliminating costs from idle or underutilized resources. A true serverless service is said to "scale to zero," meaning that when it is not being used, it should scale down to zero cost (although this is becoming less common with more complex services, which might require relatively small costs for some idle resources between uses).

Applying Concepts 2-2: Research Services for Containers

Estimated time: 15 minutes
Objective 1.6: Compare and contrast containerization concepts.
Resource:

- Internet access

Description: You just learned about container-related services in AWS used to store container images and run container instances. Researching cloud services in a specific platform to meet a specific need gives you a good opportunity to gain familiarity with that platform and learn how to distinguish between different types of cloud services. In this activity, you will research similar container services in another cloud platform.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 2-9 | If you already know a specific service in a popular cloud platform like AWS that does the task you need, it can be much easier to find a comparable service in another cloud platform. Use your favorite search engine, such as Google. Ask for a comparison that will suggest similar services in the new platform. For example, the prompt "aws ec2 vs alibaba cloud" yields search results identifying an Alibaba service called ECS (Elastic Compute Service) that works similarly, and the search results include a brief comparison of the two platforms and services. You can try the same approach with the research in this activity.

Complete the following steps:

1. Select a popular cloud platform besides AWS. You've been learning about Azure and GCP along with AWS in this course. However, there are many other good options for this activity, such as Alibaba Cloud, IBM Cloud, and DigitalOcean.
2. Do some research on your selected cloud provider's website to find the following information:
   a. In AWS, you would likely use ECR to store container images. What is a comparable service on your selected cloud platform that is used to store container images?
   b. In AWS, you could use ECS to run containers natively in the AWS environment. What is a comparable service on your selected cloud platform that is used to host containers?
   c. In AWS, you could use EKS to run containers using an AWS-adapted version of Kubernetes. What is a comparable service on your selected cloud platform that is used to host containers?

Networking Containers

For containers to send and receive traffic on the network, they need to communicate through their host's ports. The information that containers need to find these ports is provided through **port mapping**, which bridges a specific port on the host to a corresponding port on the container. This arrangement enables communication between containers and the larger network, improves security, supports load balancing and scaling, and enables more flexible network configurations for all of a host's containers. Compare this architecture to the way VMs are networked in relation to the host—the VM is connected virtually to the host's network interface at OSI layer 3. But containers are mapped to the hosts' ports instead, at OSI layer 4. This makes a bit more sense if you consider that containers generally are used to run individual applications. And applications are typically run on their host's ports, such as SSH (Secure Shell) using port 22 or a web server using ports 80 and 443. Similar to applications on a host, containers can also use TCP or UDP at layer 4.

Typically, the application running on a container resorts to that application's default ports. For example, suppose you have a container running a MySQL database. The container will default to using port 3306. But suppose you have multiple containers on the same host supporting MySQL. Now you'll need to distribute those containers across different ports to avoid conflicts in communication. Port mapping can be used to spread the containers across a range of acceptable ports rather than all of them trying to use the same default port on the host.

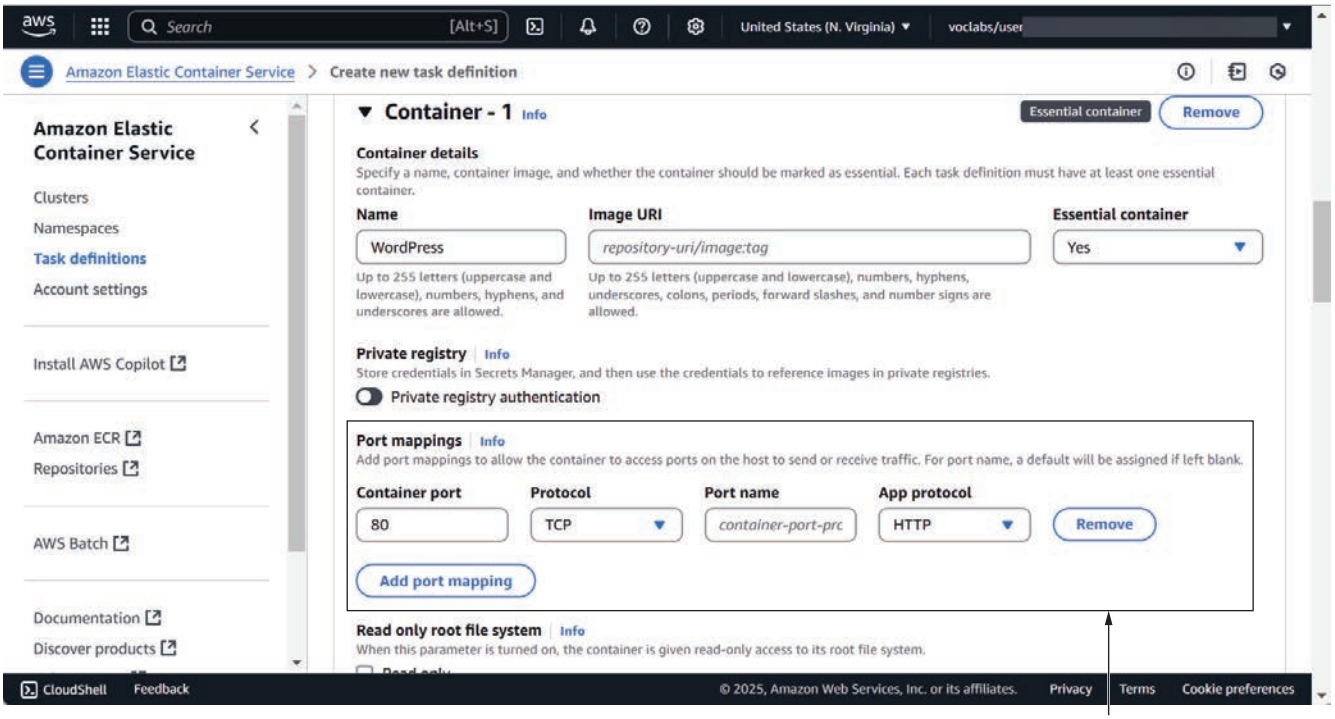
To explore how this information applies in the cloud, consider AWS's ECS, which is used to run containers in AWS. Once you've built or selected a container image, you configure a task definition as a blueprint for running an application on your containers. Figure 2-19 shows where port mapping is set when creating a task definition.

Troubleshoot Applications in Containers

Transitioning to containers doesn't solve all application-related problems. For example, CPU and memory management can prove to be a particular challenge with containers. For example, when originally creating a container image, you can set hard limits on the access a container is granted to its hosts' CPU cycles or memory. If you didn't, the container management system would control CPU and memory resource distribution according to the available resources in the host cluster. Embedding limits within the container

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-19 Map container ports to host ports



Source: Amazon Web Services, Inc.

image instead can help avoid host limitations that could cause a shutdown of the entire container system. However, these hard-coded limits might also restrict an application's performance under high demand.

As with any application infrastructure, container-based applications can suffer from service overload where one or more services within the overall system become inundated with requests and slow or crash the entire system. Consider the following potential solutions for this problem:

- **Rate limiting**—Allows for a constant flow of traffic or requests to the service regardless of the rate of incoming requests or the service's capacity. When more traffic arrives than the rate allows, those excessive requests are delayed and tend to back up until the service can work through earlier requests.
- **Concurrency control**—Applies an upper limit of traffic or requests based on the service's capacity and denies traffic or requests that exceed that threshold, which is called back pressure. Clients detecting back pressure identify the service overload and can adaptively adjust their rate of requests.

While service overload is a common problem regardless of application design, this issue is particularly relevant to the microservice design used in conjunction with containers.

Self-Check

7. Who manages the server behind a serverless compute service?
   a. The end user
   b. The CSP
   c. The cloud customer
   d. The ISP

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8. Which component does a container virtualize?

   a. CPU
   b. Storage device
   c. Memory card
   d. OS

9. Which file contains the information needed to create a container instance?

   a. Secrets file
   b. Volume mount
   c. File path
   d. Container image

☐ Check your answers at the end of this chapter.

Summary

Section 2-1: Virtualization Technologies

- Each VM can have its own OS, and all the VMs on a physical computer share the same hardware resources.
- When configuring VMs inside of a hypervisor, you're given options for the number and size of CPUs and GPUs, amount and type of memory, storage type and space, and access to device drivers.
- Each VM can be configured to network with other machines on the physical network or only to its host and other VMs on the same host.

Section 2-2: VMs in the Cloud

- When deploying VM instances in the cloud, CSPs offer preconfigured instance types where CPU and memory capacity are automatically allocated in a fixed ratio.
- When launching many instances of the same configuration, you can use a template to ensure every instance is configured the same way.
- Billing models commonly offered by CSPs for compute resources include on demand (also called pay-as-you-go), reserved instances, spot instances, and dedicated hosts.
- Factors that directly affect the cost efficiency of virtualized CPU and memory resources for a particular deployment include security considerations, HA requirements, DR strategies, energy savings, and cost considerations such as allocated storage or network placement.
- Guest VMs residing on the same host are part of an affinity group, while an anti-affinity rule will ensure VMs do *not* share the same physical host.

Section 2-3: VM Alternatives

- When running serverless compute, the CSP offers short-term or purpose-driven use of a server without access to the server environment.
- A container is a lightweight, self-contained environment that provides only the essential OS services needed to run an application.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Key Terms

For definitions of key terms, check the Glossary.

affinity
anti-affinity
AZ (availability zone)
clone
cluster
container
dedicated host
DHCP (Dynamic Host Configuration Protocol)
DR (disaster recovery)
egress traffic
elasticity
ephemeral storage
FaaS (Function as a Service)
HA (high availability)
ingress traffic
instance
Kubernetes
managed service
microservice
on demand
pay-as-you-go
persistent storage
placement group
port mapping
region
resiliency
RI (reserved instance)
rightsizing
spot instance
template
vCPU (virtual CPU)
VDI (virtual desktop infrastructure)
vGPU (virtual GPU)
VM (virtual machine)
vNIC (virtual NIC)
volume-based licensing
vRAM (virtual RAM)

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

CPU central processing unit
DHCP Dynamic Host Configuration Protocol
FaaS Function as a Service
GPU graphics processing unit
HDD hard disk drive
IP Internet Protocol
NAT network address translation
NIC network interface controller
OS operating system
RAM random access memory
SSD solid-state drive
USB universal serial bus
vCPU virtual CPU
VDI virtual desktop infrastructure
vNIC virtual NIC

Review Questions

1. You can create many EC2 instances from one _____ .
   a. snapshot
   b. image
   c. container
   d. volume
2. Where should you store a specialized container image that contains sensitive data required by a project development team?
   a. Docker Engine
   b. Kubernetes
   c. Public registry
   d. Private registry
3. Which of the following statements is true?
   a. Managed services offer customers more control over their deployments than do serverless services.
   b. With a serverless service, the customer retains some control over where supporting servers are run and how they're configured.
   c. A cloud customer is fully responsible for the security of their resources in a serverless service and for the security of the service itself.
   d. Managed services function without much input from the customer.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. Which billing model is the best fit for an expensive workload that can tolerate being started and stopped with little notice?
   a. Reserved instance
   b. On-demand instance
   c. Spot instance
   d. Dedicated host
5. Which rule would ensure two VMs are always hosted on different physical machines?
   a. Required affinity rule
   b. Preferred affinity rule
   c. Required anti-affinity rule
   d. Preferred anti-affinity rule
6. Which is largest?
   a. 1 GB
   b. 1 GiB
   c. 1 MB
   d. 1 MiB
7. What resource can provide several containers access to data files over a long period of time?
   a. Image repository
   b. Secrets file
   c. Microservice
   d. Persistent storage
8. Where does a VM instance find its OS files?
   a. Image
   b. Template
   c. Boot volume
   d. Clone
9. Which technique enables a client service to adjust its rate of requests to accommodate service overload?
   a. Hard-coded memory limitations
   b. Concurrency control
   c. Container orchestration
   d. Rate limiting
10. What kind of service do you need to store and access container images?
    a. Persistent instance
    b. FaaS
    c. Hypervisor
    d. Repository

Scenario-Based Questions

Question 2-1

Vicki is a cloud administrator for a medical billing company. She needs to configure cloud VM instances that will process medical transcriptions overnight each night in preparation for billing process actions performed by employees the following day. The transcription processing only takes about two hours and can stop and start repeatedly throughout the night. Primarily, Vicki is searching for ways to perform the processing as cheaply as possible, so long as the data is ready by the start of the business day each day. Which billing model is the best fit for this scenario?

a. Pay-as-you-go instances
b. Spot instances
c. Reserved instances
d. On-demand instances

Question 2-2

Brian is moving his network's backup DHCP server to a VM running Ubuntu Server. The VM will provide a couple of other network services as well, including DNS. Which networking mode should Brian assign to the Ubuntu VM?

a. NAT mode
b. DHCP mode
c. Host-only mode
d. Bridged mode

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Question 2-3

Douglas is analyzing costs for a cloud-hosted application to find ways he can optimize its infrastructure design and reduce expenses. He discovers that the app makes frequent data transfers to a third-party vendor, often sending the same information repeatedly. Which cloud expense can be reduced by implementing a data cache system on the vendor's side of this app's infrastructure?

a. Ingress traffic
b. Affinity
c. Egress traffic
d. Storage

Hands-On Projects

Note 2-10 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Project 2-1: Deploy a VM in AWS

Estimated time: 30 minutes
Objective 1.10: Compare and contrast methods for optimizing workloads using cloud resources.
Resources:

- AWS account (either a standard account or an AWS Academy account)
- Internet access

Description: Each of the major public cloud providers works similarly when deploying VM instances in the cloud. In this project, you will deploy a VM in AWS. In the next two projects, you'll find how that process varies only slightly in Azure and GCP.

AWS VM instances are deployed in the EC2 service. Other than a few minimal choices such as OS image and VM size, most of the typical settings are configured by default. While you can change them if you want to, the default settings work well while you're still learning what you're doing.

Recall that in Project 1-2, you surveyed available AWS account options and created an AWS account. In your AWS console for that same account, complete the following steps:

1. Go to the **EC2** dashboard and click the **Launch instance** button.
2. Give your instance an appropriate name.
3. Choose a free-tier-eligible **Windows Server** AMI. Which AMI did you choose?
4. Choose an instance type, again being careful to choose an option that is free tier eligible. For example, the combination of Microsoft Windows Server 2022 Base AMI and a t3.micro instance type are both free tier eligible and are compatible with one another. Which instance type did you choose?
5. If you're planning to connect to your instance, you'll need a key pair. A key pair acts like a lock and a key. AWS holds the public key, similar to a lock on a door. You save the private key, similar to a key that fits the lock. When you present the private key, it's like using a custom key to unlock the front door to your house. Key pairs are more secure than passwords because the lock and the key must match—this is harder than

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a password for a hacker to crack. Key pairs are also encrypted when stored, so it's harder for a hacker to discover this information.

Create a new key pair, give the key pair a name, and then download the key pair as a .pem file. Be sure to save your key pair in a safe place, as anyone with that file can access your VM instance. Where did you save your key pair file?

CAUTION: Do *not* proceed with the next step until you confirm you have downloaded your key pair. You will not have another opportunity to do so.

6. Scroll through the additional configuration options to familiarize yourself with what options are available, but for now, stick with the default settings. Respond to the following questions:
   a. What traffic is allowed through the security group firewall?
   b. What size root volume will be created?
7. When you're ready, click **Launch instance** to proceed with default settings. When the VM is launched, click **View all instances**.
8. You might need to refresh your list to display your instance in the list. When the VM reaches a running state, **take a screenshot** showing your VM in the instances list; submit this visual with your responses to this project's questions.
9. Explore information about this VM. To do this, on the Instances page, select the check box next to the instance—this generates information in the lower pane. There's a lot of important information on this screen, such as the availability zone, security group, VPC (virtual private cloud), and subnet. You'll learn more about these items in later chapters. Respond to the following questions:
   a. What is your VM's private IP address?
   b. What is its public IP address?

When you're finished with your cloud VM instance, complete the following steps:

10. Return to the list of running instances. Select your instance in the list, click **Instance state**, choose **Terminate (delete) instance**, and then click **Terminate (delete)**. Note that stopping the instance is not sufficient as this is analogous to shutting down a physical server—the VM will continue to exist and could accrue charges. Terminating the instance deletes all related resources to prevent ongoing charges.
11. Wait a few minutes and then refresh the list of running instances. Confirm your instance is removed from the list or that it is listed as terminated.
12. Make sure to properly sign out of your account.

Note 2-11 | It's important to develop good cloud hygiene habits and delete any resources you no longer need. This is an important habit that will help you minimize cloud expenses.

Project 2-2: Deploy a VM in Azure

Estimated time: 30 minutes
Objective 1.10: Compare and contrast methods for optimizing workloads using cloud resources.
Resources:

- Azure account and subscription
- Internet access

Description: Creating a VM instance in Azure works similarly to AWS and other cloud platforms. There are a few significant differences, which you'll learn about in this project.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Recall that in Project 1-3, you surveyed available Azure subscription options and created an Azure subscription for your Microsoft account. Sign into your Azure portal using that same account and subscription, and complete the following steps:

1. Click **Virtual machines** to create and manage VM instances. On the Virtual machines page, click **Create** and then click **Azure virtual machine** to begin the creation process.
2. Check that you're using the correct subscription. An Azure subscription identifies a payment method or some other crediting account. If you signed up for a free trial account with Azure, then you automatically have a Free Trial subscription. In a business environment, subscriptions allow costs to be tracked to specific departments or organizations. Subscriptions can also be transferred so that billing ownership can be reassigned to a different user. If you signed up using Azure for Students, make sure that subscription is selected.
3. Next to Resource group, click **Create new** and give the resource group a name. Resource groups organize various resources with the same life cycle, permissions, and policies. For example, each VM instance can be related to many resources such as network interfaces, storage disks, IP addresses, security groups, and a virtual network. Together, the resources that make the VM function are part of a resource group. Resource groups may contain many VMs and their associated resources. Resource groups also provide a quick and easy way to delete all resources associated with a VM when you're finished with the VM. What is the name of your resource group?
4. Give your VM an appropriate name.
5. Scroll down a little and choose an image. Images are offered from the Marketplace, including several BYOL (bring your own license) options, or you can use your own images. Choose a **Windows Server** image, such as Windows Server 2019 Datacenter. Which image did you choose?
6. Next to Size, click **See all sizes**. Like AWS, Azure offers options for number of vCPUs and allocation of memory, storage disks, storage space, and more. At the time of this writing, the Azure free trial includes 750 hours of Azure B1S General Purpose VMs for Windows Server and Linux. Choose the **B1s** size, and then click **Select**. Note that at the time of this writing, some zones do not support all sizes. You might need to change your zone to access the B1s size.
7. If you're creating this instance to connect to it, you will need a username and password. Specify a username and password and then confirm the password.
8. Next to Public inbound ports, confirm that RDP (3389) is allowed. You'll need this access to connect to your instance if you will be using this instance for Challenge Project 2-1.
9. Click **Next: Disks >** and explore the other configuration options through the other pages while keeping the default settings. Respond to the following questions:
   a. What is the size of the OS disk that will be created?
   b. What is the IP address range of the subnet that will be created?
10. When you're ready, click **Review + create**. If the validation fails, troubleshoot any missing information or configurations. Once the VM passes validation, click **Create**.
11. Return to the Virtual machines page and confirm your VM is listed and is in a running state. **Take a screenshot** showing your VM in the instances list; submit this visual with your responses to this project's questions.
12. Examine available information about your VM. Respond to the following questions:
    a. What is your VM's public IP address?
    b. What is your VM's private IP address?

When you're finished with your cloud VM instance, complete the following steps:

13. Return to the VM's page in Azure. Click **Overview** and then click the name of the resource group you created in Step 3. Delete the resource group.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14. Wait a few minutes for the task to complete. Click **Resource groups** in the navigation pane. Confirm the resource group deletion has completed. If the notification icon still shows activity in the upper-right corner, wait a few more minutes, and refresh the Resource groups list until your resource group is removed from the list.
15. Make sure to properly sign out of your account.

Note 2-12 | It's important to develop good cloud hygiene habits and delete any resources you no longer need. This is an important habit that will help you minimize cloud expenses.

Project 2-3: Deploy a VM in GCP

Estimated time: 30 minutes
Objective 1.10: Compare and contrast methods for optimizing workloads using cloud resources.
Resources:

- GCP account
- Internet access

Description: In GCP, you create VMs using the Compute Engine service. As with AWS and Azure, you can choose CPU, memory, storage, and networking configurations.

Recall that in Project 1-4, you surveyed available GCP account options and created a GCP account. Using that account, sign into Google Cloud and complete the following steps:

1. First check which project you're using. Google projects allow you to organize related resources for easy access. For example, projects can group a set of users with billing, authentication, and resource monitoring settings. Projects can be used to target billing activities so specific departments or programs within an organization can track their GCP expenses separately. When you create a VM in Compute Engine, you must create the VM within a project. When you first create your GCP account, a default project called "My First Project" is created for you. Click the name of the project at the top of the window to display the list of projects in your account.
2. Click **NEW PROJECT**. Give the project an appropriate name, and then click **Create**. What name did you give your project?
3. After the project is created, in the notification, click SELECT PROJECT. Confirm you are working in your new project.
4. Click the Navigation menu icon, and then click **Compute Engine**. There might be a delay while GCP initializes Compute Engine for your new project. If you get an error, watch the notifications list, and try again in a few minutes.
5. Click **CREATE INSTANCE**. Note that you might get a notice requesting that you enable the Compute Engine API. If so, click **ENABLE**.
6. Give the instance an appropriate name.
7. In GCP, you choose a machine type by the number of vCPUs, which then determines the amount of memory the VM will have and gives an initial indication of how much it will cost to run the machine. At the time of this writing, the GCP free tier includes use of the e2-micro machine type within certain limits. Under Machine type, choose **e2-micro**. The e2-micro instance type is only available in a few regions, so if this instance type is not listed as an option, research the current regions where the type is available, and change to that region near the top of the Machine Configuration section. For example, at the time of this

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

writing, the available regions are us-west1, us-central1, and us-east1. You might need to research current free-tier services for GCP.

8. In the left navigation pane, click **OS and storage**.
9. The boot disk holds the OS image for the VM. GCP supports several varieties of Linux and Windows operating systems. Under Operating system and storage, click **CHANGE**. Choose one of the **Ubuntu** operating systems and click **SELECT**. Which version did you choose?
10. Click through the **Data protection** and **Networking** pages and explore the other configuration options. However, leave all default selections in place. Respond to the following questions:
    a. What snapshot schedule will be applied to your VM?
    b. What private IP address range will be used for your VM's network interface?
11. When you're ready, click **CREATE**.
12. Wait for the instance to reach a running state. **Take a screenshot** showing your VM in the instances list; submit this visual with your responses to this project's questions.
13. Click the instance and explore its configuration information. Respond to the following questions:
    a. What is this VM's private IP address?
    b. What is its public IP address?

When you're finished with your cloud VM instance, complete the following steps:

14. On the GCP navigation menu, point to **IAM & Admin**, and then click **Manage Resources**.
15. Select the project you created in Step 2 and delete it. Confirm the project is shut down and scheduled for deletion.
16. Make sure to properly sign out of your account.

Note 2-13 | It's important to develop good cloud hygiene habits and delete any resources you no longer need. This is an important habit that will help you minimize cloud expenses.

Challenge Projects

Note 2-14 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 2-1: Remotely Access a Cloud VM Instance

Estimated time: 20 minutes (+10 minutes for group work, if assigned)
Objective 1.7: Compare and contrast virtualization concepts.
Group work: This project includes enhancements when assigned as a group project.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Resources:

- Internet access
- An account for a public cloud platform, such as AWS, Azure, or GCP

Description: When you create a VM instance in the cloud, there are often reasons why you might need to remotely access that instance, whether via RDP, SSH, or some other platform-specific tool. In this Challenge project, you'll create a new VM and remote into it. Complete the following steps:

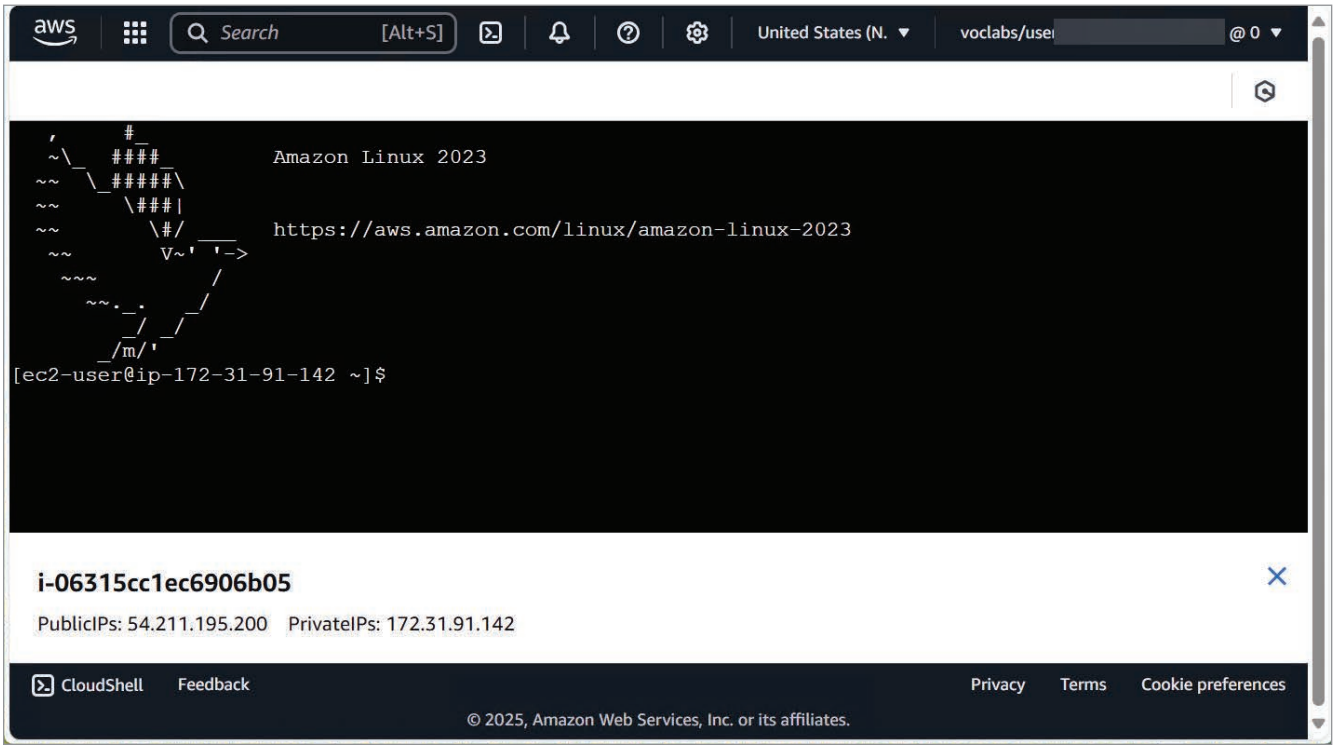
1. If you need to create a project or resource group to manage project resources when you're finished, do so now.
2. Configure and launch a VM instance. Refer to the steps in the relevant Hands-On Project in this chapter if you're using the same cloud platform. Make sure you take note of any username, password, or key pair you assign to the VM as you configure it. You might need to download the key pair, depending on the cloud platform and VM configuration you use.

   CAUTION: In AWS, you have one opportunity during the creation of the EC2 instance to download the key pair. Do not skip this step. You will not be given another opportunity to access the key pair.
3. The remote access method you use partly depends on what options the cloud platform gives you. In your cloud platform, explore the connection options listed for your VM. You might find browser-based options such as AWS's EC2 Instance Connect, RDP-based options such as Azure's Native RDP, or SSH-based options such as GCP's SSH client. If available, the easiest option is a browser-based method.

   If you must use a remote access client with RDP or SSH (such as when connecting to a VM hosted in Azure), the protocol you will use—RDP or SSH—to remotely access your VM instance will partly depend on the VM's OS. For example, if you created a Windows VM, you'll likely prefer to use RDP. If you created a Linux VM, you'll need to use SSH. This decision might also be affected by the OS of the local computer you're establishing the remote connection from. Also, the tool you use on the local computer will be determined by the computer's OS. For example, if you're connecting remotely using RDP and you're working from a macOS computer, you'll need to use the Microsoft Remote Desktop app from the Mac App Store. (Research online for more specific steps for this option.) If you're using a Linux computer, many Linux distros come with an RDP tool to connect to a Windows computer. If you're using a Windows computer, the tools you'll need are built into the OS. In contrast, if you're connecting remotely using SSH, you'll need an SSH client such as OpenSSH or PuTTY. (Again, research online for more specific steps for the option you choose.)
4. With a browser-based connection tool, your next step might be as simple as clicking Connect, accessing a list of connection options, and then clicking the browser-based option from the list. If you're using an RDP or SSH option, follow the on-screen steps, and troubleshoot as needed. You might discover you need to create a new VM with a different OS, a different key pair, or a different username and password in order to troubleshoot issues. Research additional information as needed and make adjustments to your procedures until you obtain a successful remote connection. **Take a screenshot** showing your connection with your cloud VM; submit this visual with your responses to this project's questions. Figure 2-20 shows a successful remote connection to an Amazon Linux VM in AWS, using an AWS Academy student account, from a Windows computer using a browser-based connection tool.
5. When you're ready, terminate the connection.
6. Consider trying again using a different connection tool, a different OS on the VM, or a different cloud platform.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2-20 Remote connection with an Amazon Linux EC2 instance



Source: Amazon Web Services, Inc.

Note 2-15 | If the connection with your VM is not successful, you'll need to do some troubleshooting. For example, make sure your local firewalls (both on your network and on your physical computer) allow RDP or SSH traffic, and make sure your VM instance allows RDP or SSH traffic from the local IP address of your physical network or from all IP addresses (that is, from 0.0.0.0/0). If relevant, make sure you used the correct RDP file.

7. **For group projects:** Communicate with one or more group members to get the information you need to remote into someone else's cloud VM instance. For example, you will need to know the following information for your group member's VM: public IP address and sign-in credentials (such as username and password, or key pair). When sharing sign-in credentials, make sure you are only sharing the sign-in credentials for your cloud VM instance, *not* for your cloud account. Also make sure you have not used this same information for any other account. What might you change to your cloud VM instance to prevent other people from being able to sign into it later, while still allowing your own computer to sign in? Consider testing your theory to find out if it works.
8. When you're ready, delete all resources you created for these projects. If relevant for the cloud platform you're using, you can find more specific information in the final steps of each Hands-On Project.

Note 2-16 | It's important to develop good cloud hygiene habits and delete any resources you no longer need. This is an important habit that will help you minimize cloud expenses.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Solutions to Self-Check Questions

Section 2-1: Virtualization Technologies

1. Which of the following license types do you use when you purchase and install Microsoft Office on your personal computer?

   Answer: d. Client access license

   Explanation: A **CAL (client access license)** is a software license provided per user or per device. A volume-based license is designed for many users under one license. It might be calculated by the number of CPU cores (core-based licensing) or by the number of CPU sockets (socket-based licensing).

2. Your hypervisor can allocate use of the host's GPU to only one VM at a time. Which deployment technique is the hypervisor using?

   Answer: b. Pass-through

   Explanation: With **pass-through** GPU deployment, the GPU can only be used by one VM at a time. With a virtual or shared GPU, the GPU's resources are artificially divided into portions that can be allocated at the same time to multiple VMs. Host-only is a type of VM networking mode.

3. Which network mode requires a VM to receive an IP address from the physical network's DHCP server?

   Answer: c. Bridged

   Explanation: In **bridged mode**, a vNIC obtains its own IP address, default gateway, and subnet mask information from the physical LAN's DHCP (Dynamic Host Configuration Protocol) server. A vNIC in NAT (network address translation) mode obtains IP addressing information from its host, rather than from a server or router on the physical network, and the NAT feature performs address translation on communications between the VM and the physical network. In host-only mode, VMs on one host can exchange data with each other and with their host, but they cannot communicate with any nodes beyond the host. Dynamic refers to the IP address assignment type, which is received from DHCP, in lieu of a static IP address.

Section 2-2: VMs in the Cloud

4. Which configuration selection determines the amount of RAM a cloud VM can use?

   Answer: a. Instance type

   Explanation: The **instance type** determines the instance's virtual hardware resources, including vCPUs, RAM, and storage space (such as storage volumes). The image used to create the VM instance determines its OS. Network bandwidth might be determined by the instance type or might be configured separately.

5. Which configuration parameter can ensure all the VMs for a deployment reside on the same physical host?

   Answer: c. Affinity

   Explanation: Guest VMs residing on the same physical host are part of an **affinity** group. Each partition in a placement group contains multiple instances, and each partition's instances reside on different hosts in different racks than the instances in another partition. A cluster is a group of devices or resources that appear as a single device or resource to the rest of the network. Configuration options for cloning VMs can be defined in a template.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6. Which principle ensures you can consistently access a cloud resource?

Answer: b. HA

Explanation: HA (high availability) is a resource's maximized potential for being available consistently over time and is often referred to as the number of nines identified in the percentage. An instance's network placement determines what traffic is allowed into and out of its network space. DR (disaster recovery) strategies and tools are designed to facilitate an organization's recovery from an adverse event. Multiple host machines working together in a cluster provide failover coverage for guest VMs.

Section 2-3: VM Alternatives

7. Who manages the server behind a serverless compute service?

Answer: b. The CSP

Explanation: Despite the name, servers are still involved with serverless computing. However, the cloud customer doesn't have to bother with configuring or managing the server—the **CSP (cloud service provider)** does this part. Neither the ISP (Internet service provider) nor the cloud customer's end user has access to these cloud servers.

8. Which component does a container virtualize?

Answer: d. OS

Explanation: A container is a self-contained environment that provides the OS services needed to run an application—essentially, this is **OS (operating system)** virtualization. A hypervisor virtualizes CPU, storage, and memory capacity for a VM.

9. Which file contains the information needed to create a container instance?

Answer: d. Container image

Explanation: Once the code and all its variables and dependencies are defined, the information is packaged into a **container image**. This image will serve as a template for all the containers that will run this code. A secrets file might contain authentication information such as passwords or API keys. A volume mount connects a container to persistent storage, which is a storage device that continues to exist even when the container is terminated. A file path can give the container information on how to connect with that storage device.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Chapter 3

Cloud System Design

Chapter 3 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Compare cloud migration types and cloud migration strategies.
2. Explain the benefits and techniques of cloud-native design.
3. Describe the components of system development in the cloud.
4. Apply business continuity planning.

Chapter 3 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 3-1: Cloud Migration
- Migration Types
- Migration Strategies
- Budgeting for Cloud
- Regulations and Compliance
- Project Management

Section 3-2: Cloud-Native Design
- Building with Microservices
- System Integration

Section 3-3: System Development
- DevOps
- Version Control
- Deployment Strategies
- Testing Types

Section 3-4: Planning for Problems
- Common Deployment Issues
- Sizing Issues
- Capacity Planning
- Auto Scaling Types
- Business Continuity Planning
- Disaster Recovery

Chapter 3 Scenario

"I think I get it now!" quips Malik as he walks into Kendra's office. You and Kendra had been discussing an upcoming audit, and you're both more than happy to talk about something else–anything else.

"What's up, Malik?" asks Kendra with an amused and quizzical look on her face.

"VMs. A machine within a machine or, really, lots of virtual machines running on one physical machine. It's like lots of apartments in one apartment building, or lots of stores in a mall, or maybe the way we use a bus system where people ride together instead of driving their own

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

cars. Oh! Maybe it's like a party tray with lots of sandwiches for everyone to share. By the way, I'm hungry."

Kendra laughs as she pushes away from her desk, stands up, and grabs her jacket. "Come on, let's get some lunch."

At the restaurant, as you all settle in to eat your tacos (because it's Tuesday and you always eat tacos on Tuesday), Malik adds a little more detail to his revelation. "So, each VM on a physical server shares resources with other VMs. That makes sense now. But the VM also is its own system, so it can live on just about any server to use shared resources. That makes it easy to move VMs from one system to another."

"Right," you say as you jump in on the conversation. "And we don't even have to use VMs in the cloud—we can use containers instead. They're like travel-size VMs where you take only what you need so it's easier to start, stop, and move around."

"Okay," says Malik. "So, it sounds like this cloud migration thing is going to be easy. We just need to copy all our applications over to VMs or containers in the cloud. Right?" he asks.

"Well, it's not quite that simple," sighs Kendra. "For one thing, VMs and containers have to be configured to communicate with other cloud services. And all the right resources need to *be* there in the cloud, or accessible from the cloud. We also need to think through what's the best design for these systems—some of our applications are old and inefficient. We also have some applications that might work better if run in a different kind of environment, maybe as a standalone service. And other applications really won't work well in the cloud at all. And then there are directory services and databases and other stuff, too. Moving large amounts of data to the cloud—such as our student database—can be really expensive." She puts down her taco. "It makes my stomach churn to think about it all."

"Okay," you say, "remember our mantra—one step at a time." You hold up one finger to emphasize your point. "This process could take months or even years, and we don't have to make all the decisions up front."

"True," says Kendra, picking up her taco again. "We don't even know yet how *much* of our network we'll move to the cloud. Will it be a few applications, or the whole 'kit and caboodle,' or something in between? Let's work on answering a few more questions." She jots down a few notes on a clean napkin:

- How will this migration affect the applications we have and need?
- How can we design our systems to best take advantage of what the cloud has to offer?
- What tools can help us design these systems?
- What problems should we expect?

Section 3-1: Cloud Migration

Certification

2.3 Summarize aspects of cloud migration.

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.2 Compare and contrast aspects of compliance and regulation.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

You might have heard someone joke that "The cloud is nothing special—it's just using someone else's computer." There is a kernel of truth to the statement, which is part of what makes the joke entertaining: Public cloud does, in fact, give you access to hardware owned by the cloud provider. However, using large-scale data centers offered by cloud providers opens realms of possibilities not accessible to most organizations when using their own hardware resources. In this sense, the cloud is much more than simply using someone else's computer.

What the cloud offers that most on-prem deployments do not is nearly unlimited scalability and the ability to distribute resources globally. In particular, this distributed nature of the cloud requires a mindset shift when designing a cloud-hosted system. More than just a collection of technologies, migrating to the cloud can enable cloud-native design; that is, a cloud-native architecture that best takes advantage of what the cloud has to offer, including the highly virtualized nature of the cloud environment. Innovative, cloud-native services, provider-managed services, and software-based system connectivity give architects and developers options that aren't widely available in on-prem environments.

In this first section, you'll explore the process of migrating into the cloud. A cloud migration might shift a minor system from on-prem to the cloud, or it might involve building a large collection of systems that fully take advantage of unique cloud-hosted services. A cloud migration begins with a system-by-system analysis of which systems could be better optimized with a shift to cloud. In most cases, organizations find that a hybrid approach is best—some systems belong in the cloud while others should remain on-prem. Additionally, organizations must decide to what degree a system should be restructured to best incorporate cloud technologies.

Migration Types

The most common cloud migration type involves moving resources from an on-prem data center to the cloud. However, this is not the only possibility. A cloud migration might involve moving resources to a different cloud platform or even back to on-prem, which is a process sometimes called repatriation. However, even when systems are moved back on-prem, today's teams often still expect "cloud-like" tools, which means on-prem technologies are evolving to function more like cloud environments already do. Consider the following migration types:

- **On-premises–to–cloud**—This is the typical scenario considered in the context of cloud migration where resources are moved from an on-prem environment to a cloud environment. Most CSPs (cloud service providers) offer extensive services to help with this process, and much consideration is given to optimizing the new cloud resources to best take advantage of what the cloud has to offer.
- **Cloud–to–on-premises**—At times, companies decide the cloud is not a good fit for a resource or system, and they decide to migrate that resource or system back to on-prem. Sometimes this happens because the cloud did not provide the anticipated cost or performance improvements. Other times, this migration makes sense because the company's own on-prem infrastructure has been upgraded to provide similar benefits as the cloud but in the company's own data center.
- **Cloud–to–cloud**—As CSPs continue to improve their offerings (and the cost packages associated with those services), organizations sometimes decide to move a resource or system from one cloud platform to another. Many of the same processes that support an on-prem–to–cloud migration also can support a cloud–to–cloud migration. Sometimes the target cloud provider offers services specifically designed to support a cloud–to–cloud migration. Many cloud customers try to anticipate a future move to a different cloud platform and make design decisions to prevent **vendor lock-in**, which refers to the ominous expense of changing vendors when you've committed to services that are unique to one vendor and can't be easily adapted to another vendor's platform. Once you commit to a cloud platform, it can be very expensive to change platforms later. This often prevents customers from considering

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

competitors' products. However, anticipatory planning can be particularly challenging when relying on proprietary cloud services, as the configuration of these services typically cannot be transferred between cloud (or other virtualized) platforms.

Compatibility across cloud platforms is rare and typically requires the use of cloud services that are built on open-source solutions and then adapted for several specific platforms, as opposed to choosing similar services built by and optimized for only one cloud platform.

In other words, if you were designing a deployment that you anticipated possibly migrating to a different platform someday, you might choose a cloud service adapted from an open-source solution instead of a more robust and integrated cloud provider-specific service. This way, the configurations and design you build could be mostly transferred to a different service in a different platform at some point in the future. On the other hand, if you anticipated keeping this deployment in the current cloud platform indefinitely, you might instead choose that provider's proprietary service, which is more deeply integrated into their environment and can natively interact with a wider range of their other services.

Migration Strategies

Each organization will take its own unique approach to its cloud migration, depending on the scope of elements being migrated and the company's intent in making the transition. Intentions can range from wanting to save some money or needing to experiment with a new, cloud-hosted technology to overhauling their entire IT infrastructure and anything in between.

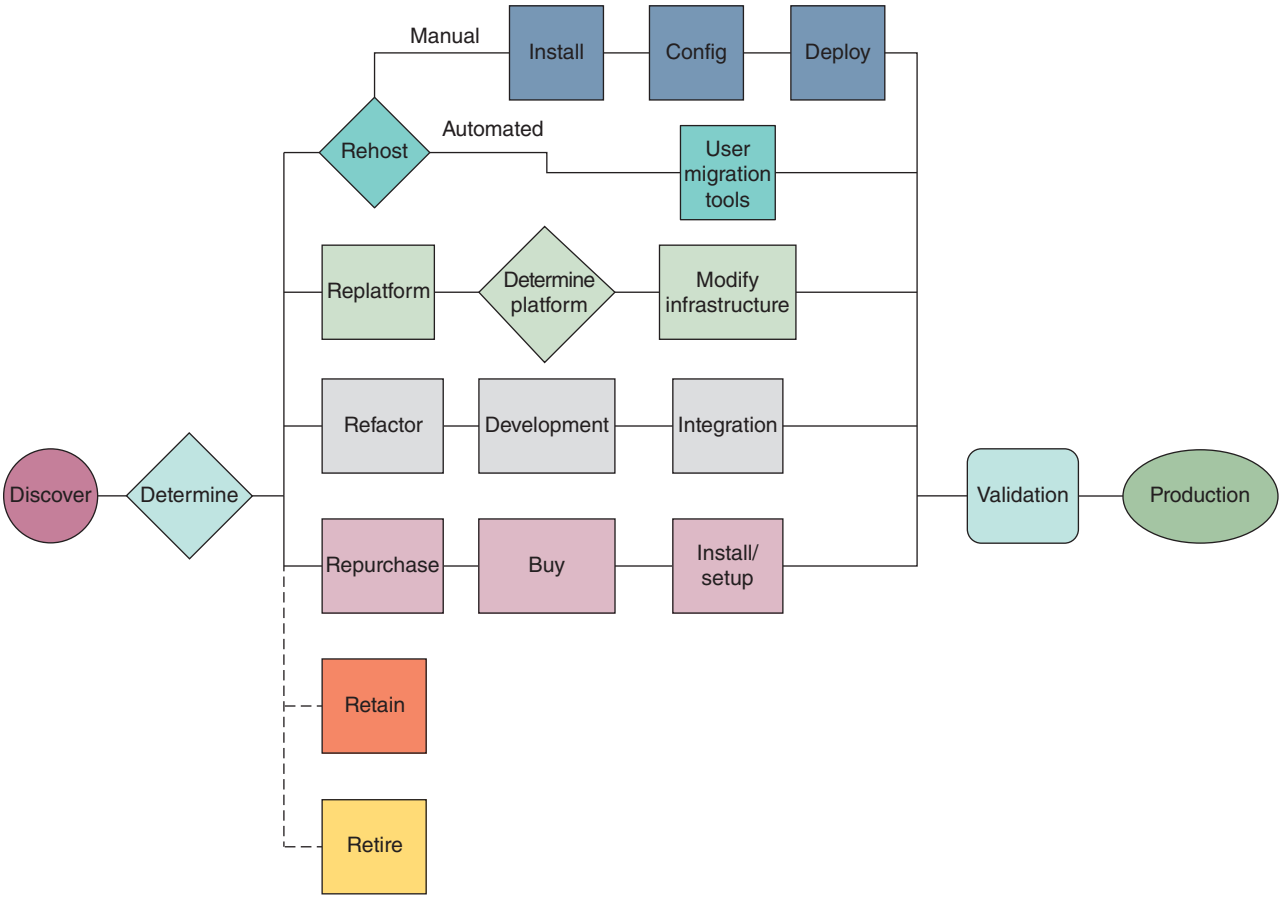
Throughout the industry, some general categories have emerged for application migration strategies that reveal a spectrum of organizational investment in the cloud. These categories have been defined by the research company Gartner, AWS, and others along the following lines:

- **Rehost**—Also called lift-and-shift or forklift, this migration strategy refers to moving an application, server, or data into the cloud as it is. The process is often automated for high efficiency. One disadvantage is that any weaknesses in the system's design or structure are retained. However, the speed of making this transition is appealing and can sometimes save money quickly. Later, the systems can be further adapted or eventually replaced (although realistically, oftentimes this reevaluation never happens).
- **Revise or replatform**—Also called lift-tinker-and-shift, this approach makes some relatively minor changes to the application or data before moving it to the cloud, such as adding a management layer or incorporating automation features.
- **Refactor, rearchitect, or rebuild**—In this approach, the changes are more significant and extensive, such as recoding portions of an application, recreating portions of the system from scratch, or rebuilding the entire system for efficient integration in the cloud environment.
- **Repurchase or replace**—This strategy refers to replacing the product with an existing cloud-native product. This can be a quick, easy, and inexpensive solution, especially considering the increasing variety of options in the cloud marketplace.
- **Retain**—Also called the do-nothing option, this is a low-complexity strategy that means the organization keeps using an application or data as it is, without any changes. This approach can be especially relevant for a system that is highly sensitive to latency caused by communicating over the Internet with cloud-hosted resources and services. The company might revisit the resource later as circumstances and technologies continue to evolve.
- **Retire**—This is also a low-complexity strategy that basically means the organization stops using the application or data.

Figure 3-1 shows how the migration process varies among these strategy options. Study this diagram carefully, as many CSPs and related vendors use similar diagrams to help customers consider their options.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3-1 Cloud migration strategies



Exam Tip | Take some time to study the strengths and weaknesses of each of these migration strategies. The CompTIA Cloud+ exam and other cloud certification exams might ask you to choose one migration strategy that would best fit a given a scenario. To make this decision, you will need to consider factors such as cost of different migration strategies, long-term maintenance expenses and management overhead, the need for greater optimization, compatibility with other systems, and governmental or industry requirements for data security.

Budgeting for Cloud

In choosing which approach to take for a specific resource, consider relevant factors both from a business perspective and from a technical perspective. For example, does the application generate income or does it support ongoing business processes? An app that generates income justifies greater financial investment in adapting it for migration. Initial expenses for developing or updating an app like this might be classified as **CapEx (capital expenditures)**, where the benefits are experienced over a longer period (more than one year). In contrast, the costs of maintaining a simple app that supports basic business functions should be minimized. This is **OpEx (operational expenditures)**, where benefits are used up in about the same period as when the expenses are paid for those benefits.

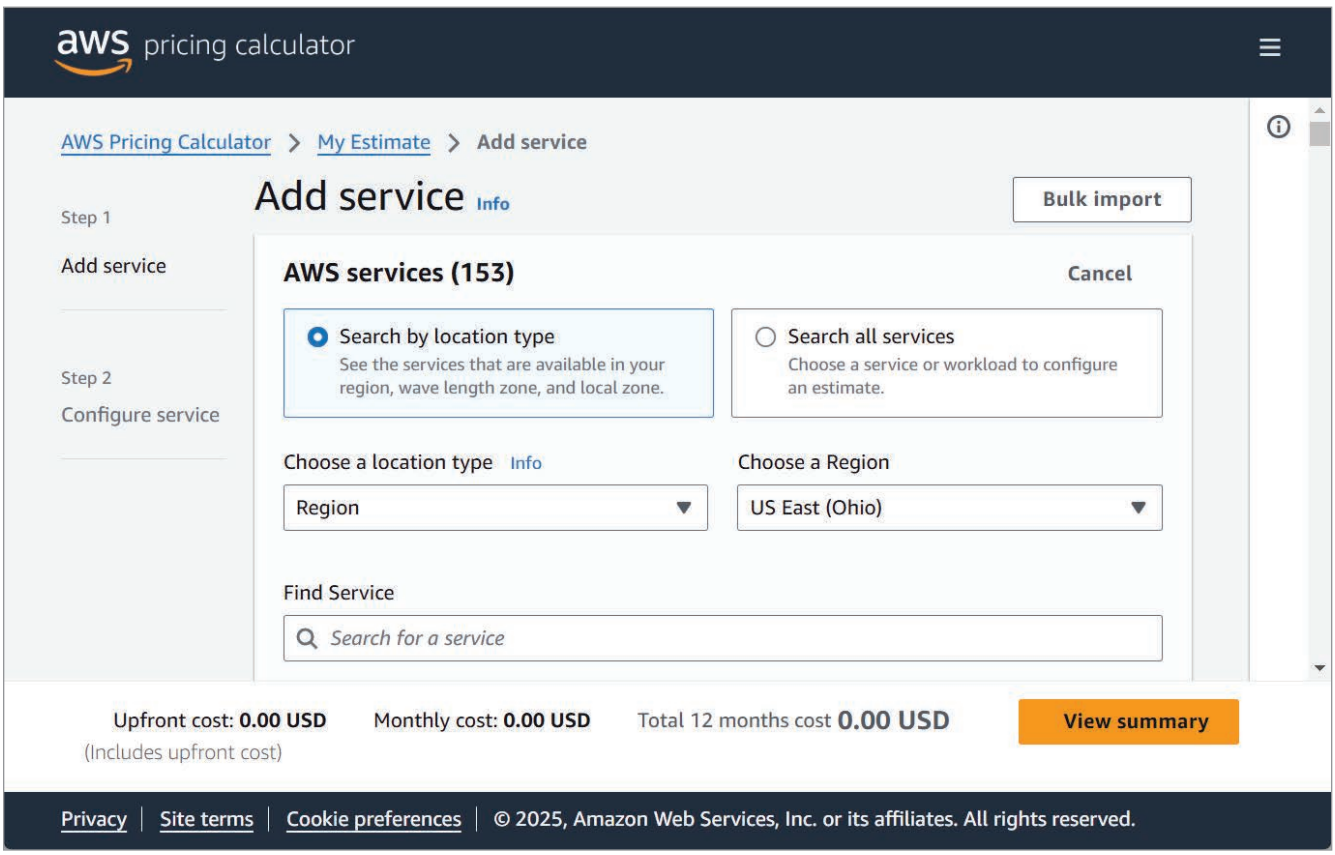
Another factor to consider is technical compatibility. Is the app compatible with a cloud environment? Can it easily be adapted to the cloud, or is it too old or have dependencies too complex to be moved? When the app can't reasonably be adapted, it might need to stay where it is or be replaced entirely.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exam Tip | The CompTIA Cloud+ exam and other cloud certification exams might ask you to differentiate between CapEx and OpEx expense items. To remember the difference, keep in mind the definition of capital. In economics, capital is a type of resource that is considered an investment and is used to generate value or income. Capital can be physical or virtual items, such as equipment, buildings, patents, or proprietary code. Knowing this definition of capital can help you remember that CapEx refers to investment in a long-term benefit. In contrast, OpEx refers to regular expenses for normal operations, which might not bring long-term benefit but ensure the business is able to continue functioning.

When building a budget for the cloud, you can't find an accurate price tag for cloud deployments on a simple chart. You'll need to factor in expenses based on your geographical location relative to your CSP's location, your backup and load-balancing needs, and bandwidth availability, among many other factors. As with an on-prem data center, hidden costs for using the cloud are easy to overlook when projecting your initial budgets. Examples include required adjustments to licensing fees, backup and archive services, and the expense of running cloud services outside of normal business hours for updates, backups, or off-hours access. All these expenses—both obvious and hidden—are collectively called TCO (total cost of ownership or total cost of operations). Many CSPs provide sophisticated TCO calculators to help you anticipate the cloud expenses you'll need to include in your budget. For example, Figure 3-2 shows AWS's pricing calculator. This kind of information will assist you in comparing the projected cloud budget with your existing on-prem costs. However, also keep in mind that the on-prem costs can also include hidden expenses you might not initially consider, such as the power bill to cool your data center or expenses for physical security and insurance coverage.

Figure 3-2 AWS Pricing Calculator



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Applying Concepts 3-1: Calculating Cloud Costs

Estimated time: 15 minutes
Objective 2.5: Given a set of requirements, provision the appropriate cloud resources.
Resource:

- Internet access

Description: All the major CSPs offer calculator tools to help customers explore their options for meeting their business needs while staying within the restrictions of their budgets. In this activity, you will choose a cloud platform, access their pricing calculator, and practice creating an estimate. Complete the following steps:

1. Choose a public cloud platform, such as AWS, Azure, or GCP. Search online to find that platform's pricing calculator service. For example, AWS's pricing calculator is at calculator.aws, and Azure's pricing calculator is at azure.microsoft.com/calculator.
2. Build an estimate that includes a VM server. Keep track of the configuration selections you make, such as the location, operating system, instance size, and payment options.
3. Repeat this process to create a new estimate, this time pricing options for a container service. For example, you might use Fargate in AWS or Container Instances in Azure. Try to make selections that match the first estimate's configurations as closely as possible, such as location, operating system, instance size, and payment options.
4. Compare the prices between your two estimates and respond to the following questions:
   a. Which would be the less expensive approach to hosting an application in this cloud platform, given the configuration options you selected?
   b. Suppose you need to host an application in the cloud. Based on what you know about how VM servers and containers work differently, what other factors do you think would be most relevant in choosing between these two estimates for this scenario?

Regulations and Compliance

During a migration to the cloud, you must consider restrictions that protect sensitive data as it's being moved and once it arrives at its new destination. Some of these regulations affect reporting processes before, during, and after data has changed location. You'll need to ensure that effective cloud security configurations are in place, which requires significant understanding of how cloud security is implemented for the specific cloud platforms you will be using. Additional consideration should be given to the physical location data is being moved to, as compliance requirements vary across geopolitical regions.

Compliance frameworks are used to define standards for cybersecurity resources, configurations, and implementations. These standards have been developed in response to historical failures and lessons learned in those cases. Compliance frameworks define best practices that ensure other organizations don't make the same mistakes. For example, a tech startup in California experienced significant damage to its data center during an earthquake. Because they had failed to implement appropriate disaster recovery measures, they lost sensitive customer data and valuable proprietary data resources. Had they been in compliance with applicable standards, such as backing up their data in the cloud, this loss could have been avoided.

Other compliance requirements apply to how data disposal is handled (such as not recycling paper that has sensitive data printed on it until the documents are thoroughly shredded), partnering with secure vendors (such as choosing vendors that perform regular audits), and implementing effective backup systems (such as using version control systems to minimize the risk of losing sensitive data to accidental deletions).

Countries and states enforce different laws for data protection. For example, some countries require that travelers entering their territory report the password to their phone or laptop to pass customs. These laws might require that the government be able to access stored data. For similar reasons, some types of data cannot be stored at data centers located inside countries with laws against certain kinds of data security. Some types of regulated data are required to be stored only within the geographical borders of an organization's

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

own country or a customer's resident country. This poses a challenge for cloud-based data storage where the physical servers hosting data can be located anywhere in the world regardless of your own location; you'll need to be aware of these issues before signing a contract. In other cases, there are limitations on what software can run on machines hosted in certain countries. An organization should have formal processes for evaluating contracts before signing them and ensure that both legal and IT experts weigh in on the terms of the agreement.

As part of this process, your team will need to research the CSP's industry certifications and audit compliance reports. Some of the most important items to examine include the following:

- **ISO/IEC 27001**—Developed by the ISO (International Organization for Standardization) and the IEC (International Electrotechnical Commission), the ISO/IEC 27001 standard provides an overarching model for organizations to use in keeping information secure. It addresses people, processes, and IT systems through a risk management process; all these components together are referred to as an ISMS (information security management system). An organization can become ISO/IEC 27001 certified by meeting these standards, as determined by an audit conducted by an accredited certification body.

Note 3-1 | The ISO (International Organization for Standardization) is an independent, nongovernmental international organization established in 1947 and is also responsible for having developed the seven-layer OSI (Open Systems Interconnection) model used in networking. Its shortened name, ISO, is derived from a Greek word meaning equal. The organization's founders chose this abbreviation to avoid confusion from having various acronyms for their name, depending on the language used.

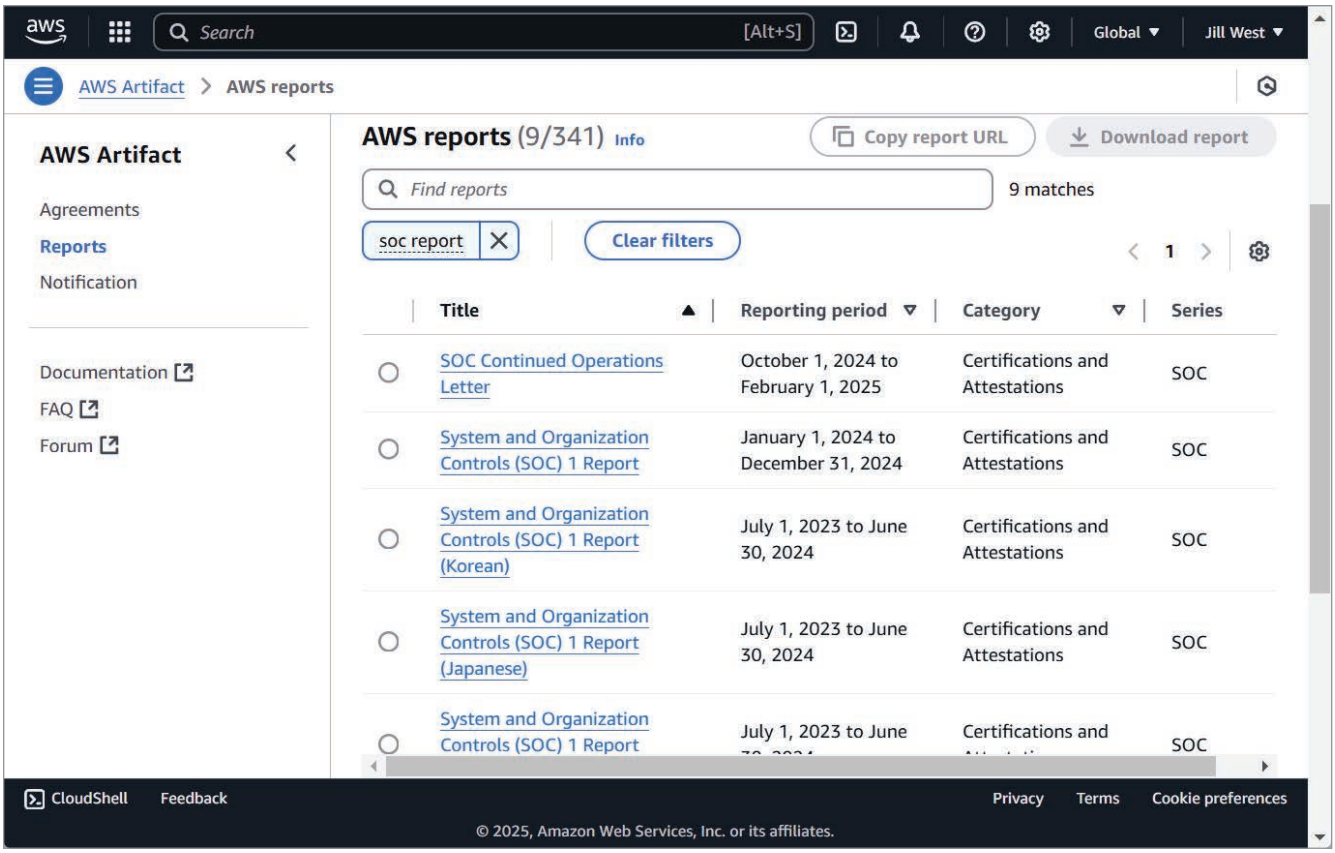
- **SSAE 21 (Statement on Standards for Attestation Engagements No. 21)**—The SSAE, developed by the AICPA (American Institute of Certified Public Accountants) and in effect as of June 15, 2022, is a standard used to determine audit compliance. The SSAE 21 replaced earlier standards such as SSAE 18 and SSAE 16, which replaced an even older standard, SAS 70 (Statement on Auditing Standards No. 70). Additional amendments continue to be made to the SSAE 21 where later releases, such as SSAE 22, replace certain portions of SSAE 21. Instead of resulting in a certification, an SSAE 21 audit results in various **SOC (System and Organization Controls) reports**. Two types of SOC 1 (pronounced "sock one") reports focus on internal financial controls. Most relevant to this discussion, however, are the SOC 2 and SOC 3 reports. Both reports address benchmarks as defined by the organization for information security, availability, processing integrity, confidentiality, and privacy. The SOC 2 report contains proprietary information and often requires a signed NDA (nondisclosure agreement) before release to a customer to restrict the customer from sharing proprietary information. The SOC 3 report is designed for public release.
- **PCI DSS (Payment Card Industry Data Security Standard)**—For all companies that provide credit card services, including accepting, storing, processing, or transmitting credit card data, the PCI DSS sets requirements for keeping that information secure. Compliance standards are determined by the PCI Security Standards Council in gradations of stringency according to annual transaction volume. The standards are enforced by credit card brands, such as Visa and MasterCard.
- **HIPAA (Health Insurance Portability and Accountability Act)**—HIPAA was enacted in 1996 to protect medical information. Among other things, HIPAA sets the stage to establish and continually evolve specific guidelines for data security when that data includes any identifiable information about patients or their medical history. It also imposes stiff penalties and fines for data breaches, whether due to criminal intent or simple negligence.
- **GDPR (General Data Protection Regulation)**—A more recent development, the GDPR became effective in mid-2018 and defines broad protections for any PII (personally identifiable information) as well as standards for how to handle breaches affecting that data. The regulations apply to any organization handling this type of information for EU (European Union) citizens, even if the organization is not based in Europe.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **CCPA (California Consumer Privacy Act)**—Similar in purpose to the GDPR, the CCPA (California Consumer Privacy Act) applies only to medium and large businesses but enforces some standards that exceed those set by the GDPR. In some ways, this law more deeply impacts U.S. business processes, especially due to its broader definition of what information qualifies as private data.

You can gather a great deal of this information by examining the cloud provider's industry certifications and compliance reports. For example, in AWS's Artifact service, you can access extensive documentation on AWS's own efforts toward compliance with a wide variety of industry standards and government regulations, including PCI DSS, HIPAA, GDPR, and many others (Figure 3-3).

Figure 3-3 SOC documentation in AWS Artifact



Source: Amazon Web Services, Inc.

Compliance efforts certainly don't end with the compliance work of the CSP, however—the cloud customer also maintains significant responsibilities for ensuring compliance with any relevant standards. The CSP's documentation addresses the CSP's portion of the shared responsibility, and this documentation can assist the cloud customer in their pursuit of their own compliance certifications.

Meeting compliance requirements can often be facilitated by applying established best practices for security configurations. CSA (Cloud Security Alliance) is a vendor-neutral nonprofit organization that helps establish and maintain IT industry standards and best practices for cloud security. Various CSA best practices and guidelines are relevant to cloud customers, cloud providers, governments, entrepreneurs, or other companies in the assurance industry. Cloud providers can participate in CSA's STAR (Security, Trust, Assurance, and Risk) program to certify their compliance with CSA's definition of cloud security best practices. CSA STAR Level 1 consists of a self-assessment, and Level 2 requires a rigorous third-party independent assessment. At the time of this writing, Level 3 was still in development and will likely incorporate ongoing monitoring and auditing systems for continuous certification.

Compliance is not a one-time endeavor but a continuous process to ensure data privacy and business continuity. A company's cloud-compliance strategy should define its compliance objectives, identify relevant

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

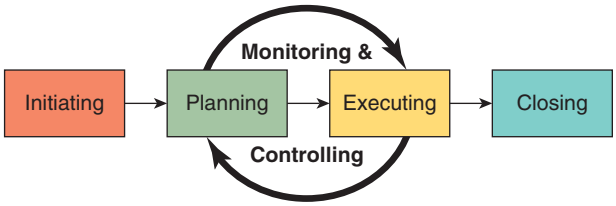
industry standards or government regulations, define a team of personnel and the members' specific job roles in relation to the compliance effort, and implement strategies for monitoring the company's progress toward and continuing conformation to the selected compliance requirements. To assist in this process, the company will need to deploy tools and cloud services, such as log monitoring, access documentation, data encryption tools, and compliance dashboards and reports. In the cloud, you'll find that you can automate many of the processes required for regulatory compliance. Plan accordingly so your cloud deployment takes advantage of cloud-built services that support compliance requirements. Involving compliance experts early in the cloud migration process can help prevent costly mistakes along the way.

Project Management

Now that you've learned how cloud migration works, you might realize that your role in this project relies on more than the technical skills of how to deploy a VM or how to configure a network interface. The management overhead of a cloud migration includes many aspects, such as cultural adjustments, shifting expertise requirements, increased need for vendor and SLA (service-level agreement) management, and potentially more complex governance and compliance requirements. Many IT professionals in the thick of a cloud migration find themselves leaning less on their technical knowledge and more on project management skills.

A **project** is an effort that has a clearly defined beginning and ending. A project involves many processes that are generally grouped into five phases: initiating, planning, executing, monitoring and controlling, and closing (Figure 3-4). **Project management**, then, is the application of specific skills, tools, and techniques to handle these processes in such a way that the desired outcome is achieved.

Figure 3-4 The five project management process groups



There are several popular certifications related to project management, including the following:

- Project+ is an entry-level certification by CompTIA that addresses project management in IT.
- CAPM (Certified Associate in Project Management) or the more advanced PMP (Project Management Professional), both by PMI (Project Management Institute), cover project management across industries.
- CSSBB (Certified Six Sigma Black Belt) by ASQ (American Society for Quality) focuses on improving business processes to increase quality.
- CSM (Certified ScrumMaster) by the Scrum Alliance emphasizes short iterations for incremental improvements.

Note 3-2 | **ITIL (IT Infrastructure Library)** is a collection of publications that defines many kinds of best practices in IT. The ITIL structured framework for ITSM (IT Service Management) guides organizations in choosing projects that are closely aligned with their business goals. ITIL's ITSM framework is useful in project management in that it provides a framework for many of the processes in and surrounding project management objectives. Project management focuses on projects which are, by definition, limited in time and scope. In contrast, ITSM supports the ongoing delivery and management of IT services even beyond the conclusion of each project.

Earning one or more project management certifications adds an attractive layer of expertise to an IT professional's resume. Nearly any undertaking in IT benefits from good project management skills, and cloud computing is an excellent example of why this is true. Let's briefly explore some key project management skills:

- **Communication**—For an industry built upon the need for communication technology, IT professionals sometimes lack in this so-called "soft skill." When asked which one skill a potential employer most wants in a new hire, the ability to communicate well often surfaces at the top of the list. Effective communication is even more critical when managing a project, as many people are involved in various

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

levels and phases of the project. Consistent and thorough communication also involves making good decisions in specific scenarios about communication channels used, such as emails, social media, face-to-face meetings, and online conferencing.

- **Negotiation**—Negotiating with vendors and other business partners creates the opportunity to expand the reach of a limited budget. Knowing how to write good RFPs (requests for proposal) and negotiate favorable contractual terms will help protect your organization not if, but when, something goes wrong during a project.
- **Task and time management**—Deadlines are often the driver for pushing things through and getting people to finish what they start. Time management is the art of learning how to set realistic deadlines that ensure the project will be completed on time while not overly stressing your team. Project management can teach you how to schedule major phases of a project as well as smaller, more incremental steps so the work is well paced, and everyone knows what they should be doing next.
- **Cost and quality management**—In many cases, budgets place nonnegotiable limits on a project. A team that regularly exceeds its budget will not be allowed to continue operations for long. What's more, a budget isn't a number you can set at the beginning of a project and not consider it again until you are near the end. Ongoing expenses and projected costs must be continuously and closely monitored and revised. Like guiding an airplane onto a landing strip, it takes a lot of planning, skill, and practice to learn how to manage costs and budgets for large projects while maintaining high quality standards.
- **Risk management**—Accepting some risks in a project will increase the likelihood that your endeavor is relevant and innovative, while minimizing unhealthy risks will prevent detrimental surprises that undermine confidence in your goals. Effective risk management, to a great degree, requires experience and maturity. Project management training will teach you how to identify the most relevant factors when considering a particular risk and propel you toward greater skill in choosing risks wisely.
- **Leadership**—The value of the ability to inspire and guide others can't be overstated. Whether you're the type of person who loves being out in front of the crowd and energizing people or you prefer to offer a quieter, more subtle form of influence, you can still be a strong leader. In fact, you don't even have to be in charge to be a leader. Anyone on the team can set the pace for focus, creativity, resourcefulness, and productivity. Project management skills help you understand how to balance the details of a project with the big picture, and how to organize communications and activities to help everyone maximize the potential of what they have to offer your team.

Self-Check

1. What physical restriction might make the cloud unreasonable for certain types of workloads?
   - **a.** Latency
   - **b.** Cooling
   - **c.** Power
   - **d.** Documentation
2. Which cloud migration strategy requires the most work to perform?
   - **a.** Retain
   - **b.** Revise
   - **c.** Refactor
   - **d.** Rehost
3. What kind of expense is an electric bill?
   - **a.** STAR
   - **b.** OpEx
   - **c.** CapEx
   - **d.** TCO

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 3-2: Cloud-Native Design

Certification

1.5 Explain the purpose of cloud-native design concepts.

1.10 Compare and contrast methods for optimizing workloads using cloud resources.

5.3 Explain concepts related to integration of systems.

More than just a collection of technologies, migrating to the cloud can enable cloud-native design; that is, a cloud-native architecture that best takes advantage of what the cloud has to offer. How does system design need to evolve when the goal is to integrate natively into the cloud environment? That's the question this section addresses.

The distributed nature of the cloud requires a mindset shift when designing a cloud-hosted system. Resources can be hosted relatively close to where they are used, even when users are located throughout the world. Designing for optimized performance across these distances requires newer technologies that cache data across resources and increase redundancy for high availability.

Cloud-native design optimizes system workflow with the implementation of several techniques:

- Prioritize automation where possible to eliminate manual processes and increase responsiveness to shifting needs and changing conditions.
- Take advantage of cloud-provided managed services to outsource the technical burden and focus instead on business-driven innovation.
- Rely on purpose-built services to support faster development, easier scaling, and greater resilience.
- Incorporate an interdependence of systems to encourage communication while reducing vulnerabilities.

Let's break this down to learn more.

Building with Microservices

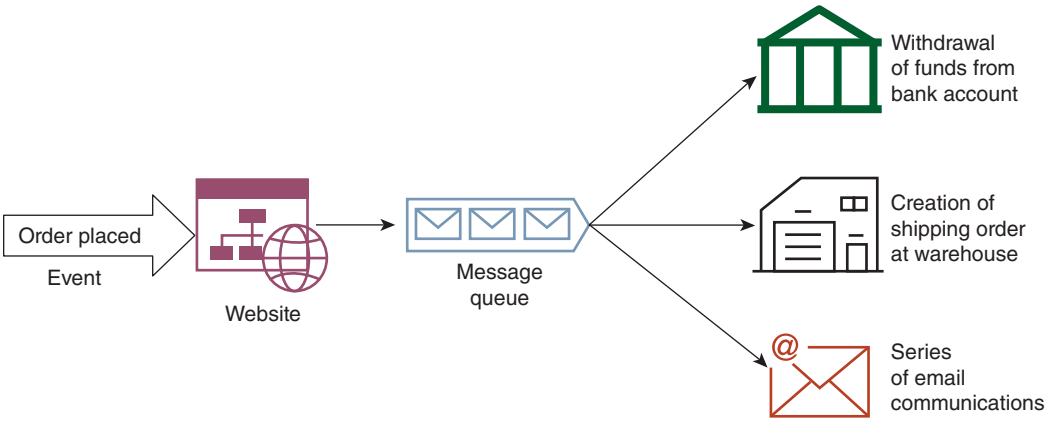
Microservices are fine-grained, purpose-built services that function interdependently as part of a larger system. When creating a microservices-based architecture, system developers take a modular approach to designing a system. Each microservice is intended to do one thing and do it well. The underlying infrastructure for that service is optimized to maximize the service's performance and security. With a microservices architecture, the system owner can also evolve independent components separately from the entire system. In other words, an individual microservice can be updated without changing other microservices in the system.

For microservices to work, they need to communicate. Although there are multiple architectural options to support this communication, microservices often rely on events. An **event** is a detectable change in a cloud service or resource that is being monitored. Basically, something happens, and information about it is captured—this is an event.

Imagine that a customer places an order on a website (Figure 3-5). The creation of this order is an event that can trigger a series of activities in other systems. **Event-driven architecture** allows the communication of events between systems, automating responses to each event. For example, when the customer places an order, this event triggers several additional activities such as a withdrawal of funds from the customer's bank account, the creation of a shipping order at the warehouse, and a series of email communications with the customer (such as messages thanking the customer for their business, informing them when the order is shipped, and even providing a notification when the shipment arrives at their location).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3-5 The website produces a message that is sent to other systems



The website's shopping cart is not directly integrated into all these other systems. Instead, the shopping cart application, which is called the producer, creates an event message that must be delivered to event consumers. There are several options for how to handle this message, but for this scenario, let's say the event message is sent to a messaging queue. Event consumers—such as the payment processing system, the warehouse's shipping system, and the email system—can all collect messages from this queue. This pattern of sending a single message to several consumers is called a **fan-out architecture**.

The consumers might receive the message as soon as it is sent, or they might periodically check the queue for new messages. The key here is to realize that the messaging between systems can be asynchronous; that is, each system works on its own timetable instead of being required to operate in sync with all the other systems.

From a big-picture perspective, the entire architecture is driven by events that occur, messages created by producers, and automated responses performed by consumers. In a cloud-native design, most or all of these event-informed systems are microservices.

With these numerous services running in your cloud, how can you keep track of them all? **Service discovery** is the process of finding and monitoring services and their health throughout your network. Service information is collected and tracked in a service catalog, including information about the service's health status. As needed, users of these services can consult the service catalog to get each service's access information (such as IP address and port number).

One of the greatest benefits of this architecture is that any of these microservices can be replaced by a different microservice that does essentially the same thing without affecting any of the other services. For example, suppose you want to upgrade your shopping cart application on your website. You can replace the entire shopping cart system with a new solution that similarly communicates with events. The change from one shopping cart solution to another doesn't affect any of the other systems in your architecture. This is called a **loosely coupled architecture**, where each system in the infrastructure operates independently of the other systems so that any individual system can be changed or replaced without affecting other systems.

Still, the transition from a monolithic design to a microservices architecture can be challenging both from a technical perspective and in the context of the business mindset. On the technical side, designers must map the flow of data from one component to the next, identifying dependencies where one component relies on activity from another component in order to do its tasks. This work requires intimate knowledge of the original system design as well as the requirements for the final product. Additionally, at an organizational level, costs must be budgeted differently, and development teams must learn to work more collaboratively so updates and interfaces can be shared across services.

Note 3-3 | Although microservices, event-driven architecture, and loosely coupled architecture all seem very similar in concept and all work well together, they're not the same thing. For example, you can have microservices that are not loosely coupled or a monolithic system that is loosely coupled. Keep in mind that, while it's most common for all three of these strategies to be used together, it's not absolutely necessary.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exam Tip | The CompTIA Cloud+ exam might ask you about the benefits of microservices in the context of business needs. For example, microservices can be developed more quickly than monolithic systems, which also makes it easier and cheaper to evolve each microservice as technology continues to change and cloud services mature. Additionally, microservice resources can be scaled more effectively to keep pace with demand, which generally reduces the costs of running these resources. Microservices are easier to maintain, are more resistant to failures, and can boost communication and coordination across diverse teams to more efficiently meet organizational needs. Not every business need can best be met with microservices instead of monolithic systems, especially when legacy software is involved and cannot be replaced. However, the benefits offered by microservices often offset the increased investment required to replace older, monolithic systems.

System Integration

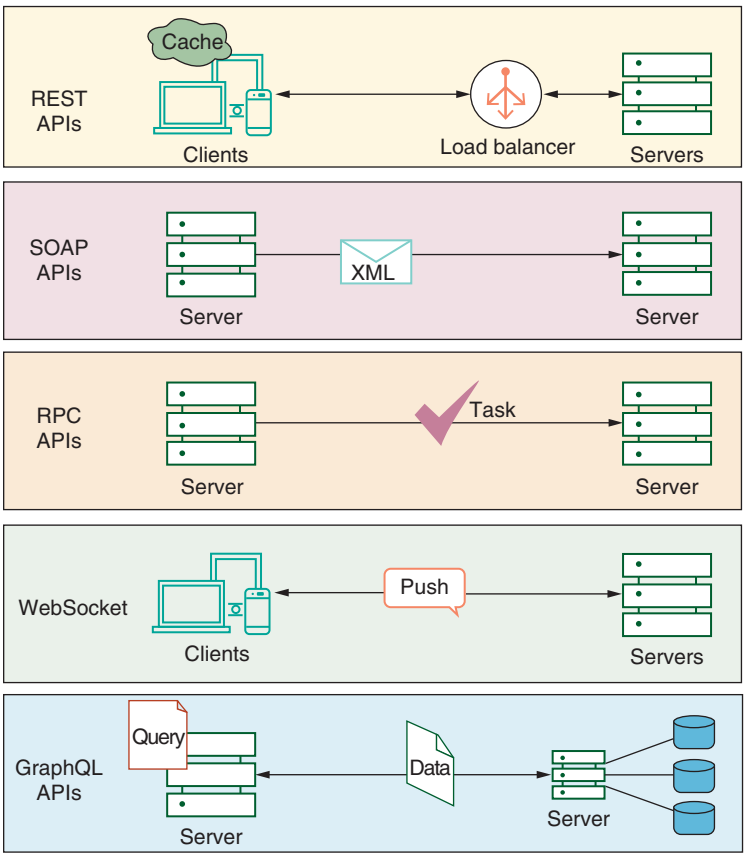
Cloud architecture includes the interfaces between cloud resources so systems can communicate with each other. Most of these communications are handled through **API (application programming interface) calls**, which are messages sent between system components using each system's API. There are several types of API protocols. Some of the most common types are described next and compared in Figure 3-6:

- **REST (representational state transfer)**—A REST API, also called a RESTful API, is popular because it is simple to set up, flexible in design, and easy to scale. REST APIs rely on a client–server structure in which one application or service (the client) receives data from another application or service (the server). REST APIs can be built using any of several programming languages or data formats, although the server requires a consistent format for all requests sent to it. For transport, REST APIs typically use HTTP (Hypertext Transfer Protocol) or HTTPS (HTTP Secure). They can also work over SMTP (Simple Mail Transfer Protocol) or FTP (File Transfer Protocol). Servers don't store information about requests sent to them—that is, the requests are stateless—so any information required to retrieve the data should be provided within the request itself. However, REST APIs can incorporate caching on either the server side or the client side, which improves performance. REST APIs can also tolerate intermediaries, such as a load balancer for performance or an API gateway for security.
- **SOAP (Simple Object Access Protocol)**—A SOAP API is more structured and formalized than a REST API and tends to be a bit slower and more difficult to set up. Whereas REST APIs are typically written with HTML (Hypertext Markup Language), JSON (JavaScript Object Notation), or even plain text, SOAP APIs more often use XML (eXtensible Markup Language). Like REST APIs, SOAP APIs typically rely on HTTP or HTTPS for transport, though they can also use SMTP or other protocols. SOAP does not support caching.
- **RPC (remote procedure call)**—RPC is older than the other API types described here. RPC APIs also use a client–server infrastructure, are written in XML (called XML-RPC) or JSON (called JSON-RPC), and communicate over HTTP. RPC APIs are less structured or standardized than REST or SOAP, and they natively support fewer types of requests, although there are ways around this limitation. As indicated by its name, RPC APIs are designed primarily to perform remote functions on servers. RPC systems can be either stateless or stateful.
- **WebSocket**—WebSocket is intended for interactive two-way communication. Rather than one system requesting information from another system, a web socket allows both the client (typically a browser) and the server (typically a web server) to share information with each other. To do this, a stateful connection is established and maintained until either the client or the server terminates the connection. WebSocket is often the best fit when you need the connection to handle live data (such as an ongoing chat) rather than retrieving existing data (such as website files). Chat applications often use WebSocket so the connection can be established one time, and then it supports an ongoing chat conversation without needing to reestablish the connection.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **GraphQL**—Developed by Facebook in 2012 and released publicly in 2015, GraphQL is both a query language and an API standard, which is considered a successor to REST APIs. Generally, GraphQL is known to return more specific results than REST can do. In other words, with GraphQL, you can expect to get the information you asked for and nothing more or less. GraphQL servers can support many languages, including JavaScript, Python, and PHP (PHP: Hypertext Preprocessor). GraphQL APIs can also retrieve data from multiple sources in a single request. Within the AWS environment, GraphQL is used by the AppSync service to query DynamoDB tables.

Figure 3-6 API protocols



Choosing between these API types is part of the complexity of building apps. Each API type can be a better fit for different jobs. For example, an RPC API is a good choice when you need the system to perform an action, such as take a picture with a remote camera or restart a remote server. REST APIs are better for performing operations on data, such as adding or updating data in a database or retrieving data from the database. These data-manipulation operations are often referred to collectively as CRUD (create, read, update, and delete).

One application of APIs is to support web services. A **web service** is a standardized method of offering network-based communications between software systems. For example, a web service might perform authentication for a website or process payments. Although many people use the terms "web services" and "APIs" interchangeably, web services are actually one way of using APIs. There are many other ways to use APIs, even without a network, such as within a single computer. Often, web services use SOAP APIs to increase security and preserve data integrity, though the other API types can work, too. Generally, the term "web service" is falling out of use, and today's technologies use APIs in other ways to manage system-to-system communications more efficiently.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

4. Which resource is delivered through a fan-out architecture?
   a. Event
   b. Microservice
   c. API call
   d. Web service

5. Which architectural design is responsible for automating responses across a system?
   a. Fan-out architecture
   b. Event-driven architecture
   c. Loosely coupled architecture
   d. Microservices-based architecture

Check your answers at the end of this chapter.

Section 3-3: System Development

Certification

2.2 Given a scenario, implement appropriate deployment strategies.

2.4 Given a scenario, use code to deploy and configure cloud resources.

5.1 Explain source control concepts.

5.2 Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.

5.4 Explain the importance of tools used in DevOps environments.

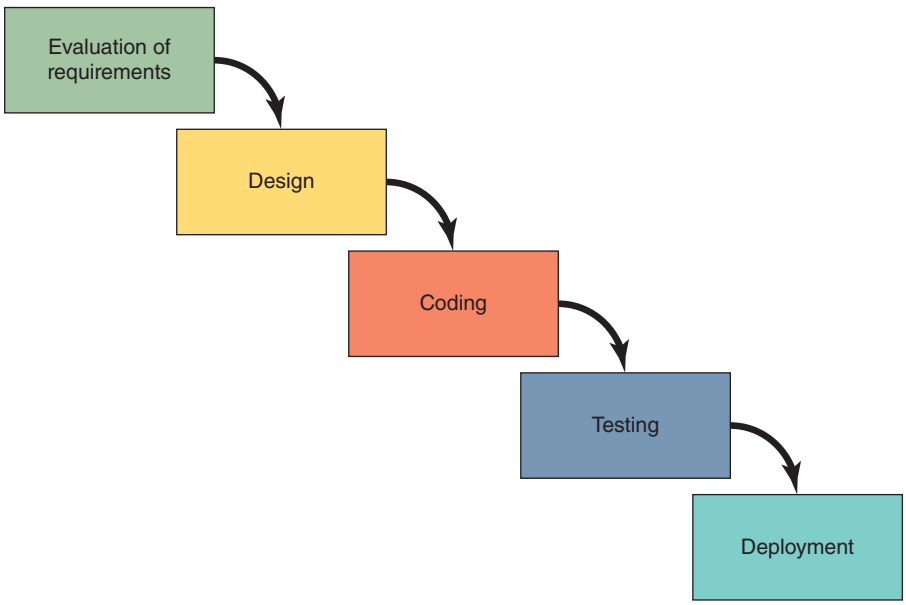
With the cloud-native emphasis on microservices, loosely coupled architectures, and code-built systems, the process of system development in the cloud starts to seem more like software development than building a physical network. A cloud administrator's work often relies on software development principles, techniques, and technologies to design, build, and maintain cloud-native resources. Additionally, cloud admins often need to support services and resources used by software developers. For these reasons, understanding software development processes and tools is critical to a cloud administrator's job. This section will discuss the application of these principles to system development in the cloud.

DevOps

When using a more traditional approach, software development projects rely on siloed teams working in isolation from each other. For example, one or more specialists might focus on designing the UI (user interface) and the overall UX (user experience). These specialists help bridge the gap between the technical design and the business goals of a product. A separate specialist, the software architect, designs the technical architecture of the product, including the selection of databases and other service components of the system. Software developers create the actual code for the product, and software testers ensure everything works as intended. Once produced, yet another team takes over managing the operation of the product. Each of these specialists works separately on their part of the project according to a timeline established by a project manager. As the product progresses from one phase to the next, each specialist or team of specialists takes over the work, having little interaction with other teams before or after their phase of work. This approach is called the waterfall method (Figure 3-7), where work proceeds from one phase to the next and rarely flows backward to revisit previous phases.

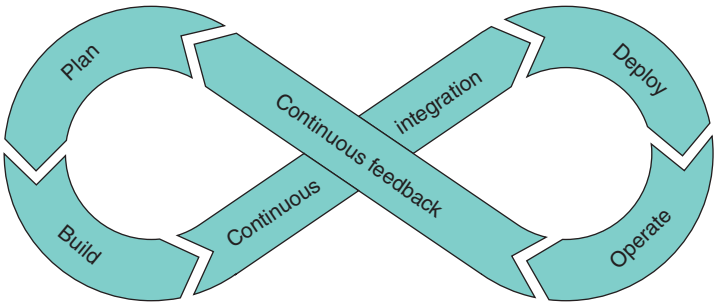
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3-7 Distinct phases of the waterfall method



In contrast, more modern system development often requires more interactive communication and partnership between teams. This culture of reducing the isolation of work responsibilities and job roles in favor of increased collaboration among teams working on a product has come to be known as **DevOps (development and operations)**. Not so much a technical process as a cultural shift, the continuous collaboration between development teams and operations teams brings highly responsive updates to application or software-defined deployments (Figure 3-8). DevOps teams rely heavily on automated processes and aim to release short, frequent improvements to their products.

Figure 3-8 DevOps blends development and operations functions



Version Control

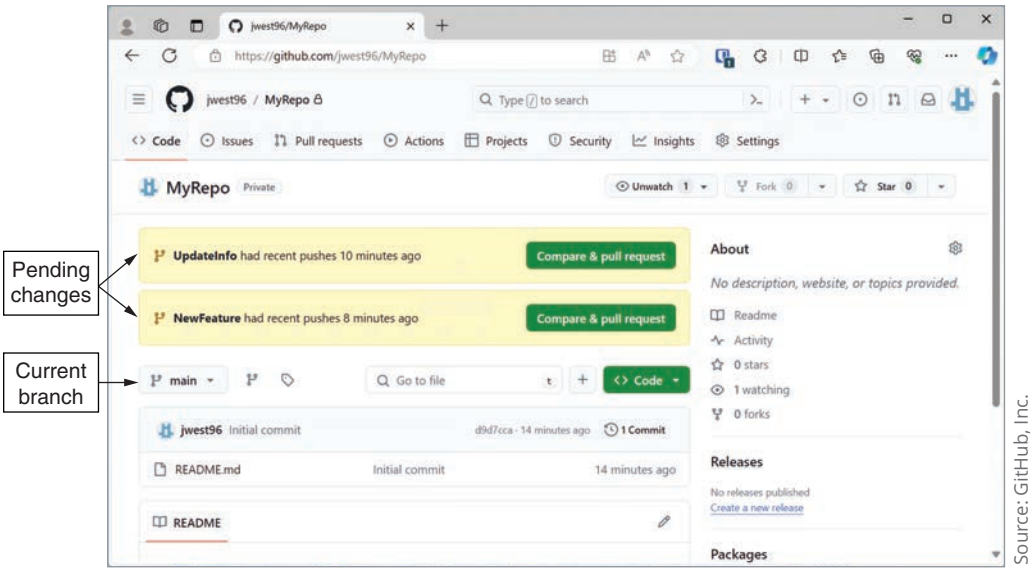
The always-changing, innovative nature of cloud-native systems can thrive in a DevOps environment. Cloud system innovation evolves more easily in response to growing market demands when teams work together, implement extensive automation, and repeatedly release improvements in short bursts.

To do this effectively, teams need robust tools for managing code across diverse teams and job roles. The code underlying cloud deployments undergoes frequent, small changes. Each time changes are saved, a new version of the code is created. These changes and versions must be tracked over time, with code edits published into different environments (such as testing, staging, or production environments) or releases (such as alpha, beta, and stable releases) of the software.

The work of keeping a timeline and record of code changes is typically done with a version control system (sometimes called source control system) such as Git. **Git** is a free and open-source version control tool that can be installed on a developer's local computer. In contrast, **GitHub** is an online service that allows developers to run Git in the cloud (Figure 3-9). At the time of this writing, a basic GitHub subscription is free for individuals and organizations up to a certain threshold of usage levels; more advanced features, such as team collaboration and managed access control, require a paid subscription or are pay-per-use. Both the free and paid subscriptions offer unlimited private and public code repositories.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

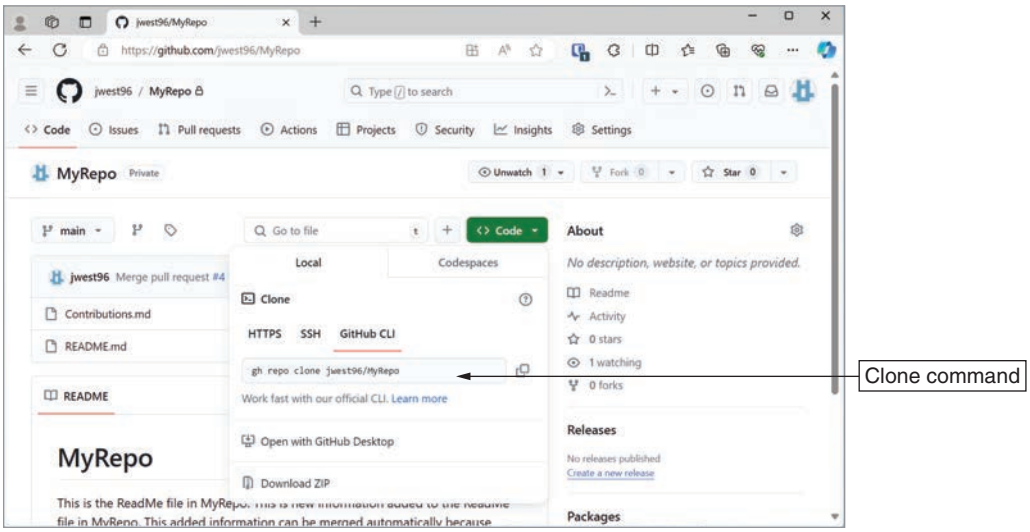
Figure 3-9 A repository in GitHub with recent changes pending review



Most of the work in Git is handled in the context of a project, which contains one or more repositories. To get started, you create a new project folder or access an existing project folder, such as someone else's project that you will be contributing to. In Git, the term repository refers to a .git folder inside a project folder. The repository tracks changes made to project files. Each set of changes collectively results in a new version of the project, which is called a **commit**. You can have many commits over the life cycle of a project.

To work on a Git project, your local computer must maintain a **local repository**. If you're using an online collaboration service such as GitHub, then your project also has a **remote repository**, which is a structured storage space for project files on a remote system. You first clone (that is, download a copy of) the remote repository to your local repository. Figure 3-10 shows the command to clone the sample repository in GitHub to a local computer.

Figure 3-10 Command to clone GitHub repository



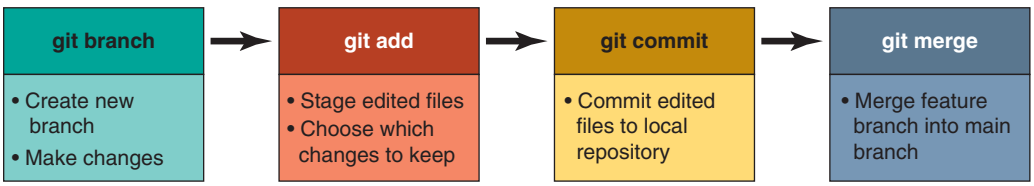
If multiple developers are working on the same project, there can be many local repositories, one on each developer's local computer. After making changes to files in your local repository, you push (or upload) the local repository to the remote repository. As you might guess, all these copies of the project can result in a variety of file versions as each person makes their own edits. Managing these versions and choosing which edits are merged into the main version is the purpose of the commit process.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Editing Code

Let's discuss in a bit more detail the process of editing a project (Figure 3-11). Suppose you need to make some edits to a few project files in your local repository. First, you create a new branch with the command `git branch`. You edit the files as desired, but perhaps you've decided that not all these files should eventually be merged back with the main branch. You use the `git add` command to choose which of these files to add to the staging area. The staging area is a temporary holding area for drafts of edited project files—it holds copies of your edited files until you decide whether to keep them. To list which files are currently in the staging area, use the command `git status`.

Figure 3-11 Progression of git commands



Once you're satisfied with your edits, use the `git commit` command to commit, or save, your edits in the local repository. This is essentially a snapshot of your files, which is a new commit. Notice this is a two-step process: you first add the edited files to the staging area, and then you save the files in a new commit. Both steps must be completed for your edits to be saved in the local repository's new branch and create a new commit (or version) of the files.

When you're finished making changes (which might involve multiple commits), you then need to merge the feature branch back into the main branch. You do this with the `git merge` command. This handling of branches is called branch management.

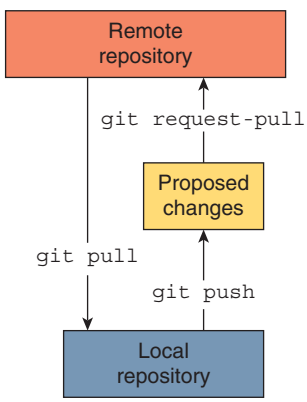
Note 3-4 | The work of branch management includes the provision of branch security. Branch security includes restricting access to branches, preventing unauthorized changes, and preventing the deletion of critical branches. A repository owner should set restrictions on who can merge or push code changes and set requirements for review of those changes. Most of these configurations can be set in the version control tool being used to host the code repositories.

Pull Requests

The steps discussed so far apply to handling the code within your local repository. If you're working with a team and a remote repository, you'll also need to download the remote files and configurations (typically with the `git pull` command or the `git fetch` command) at the beginning of this editing process, and then upload your changes (typically with the `git push` command) when you're finished editing (Figure 3-12).

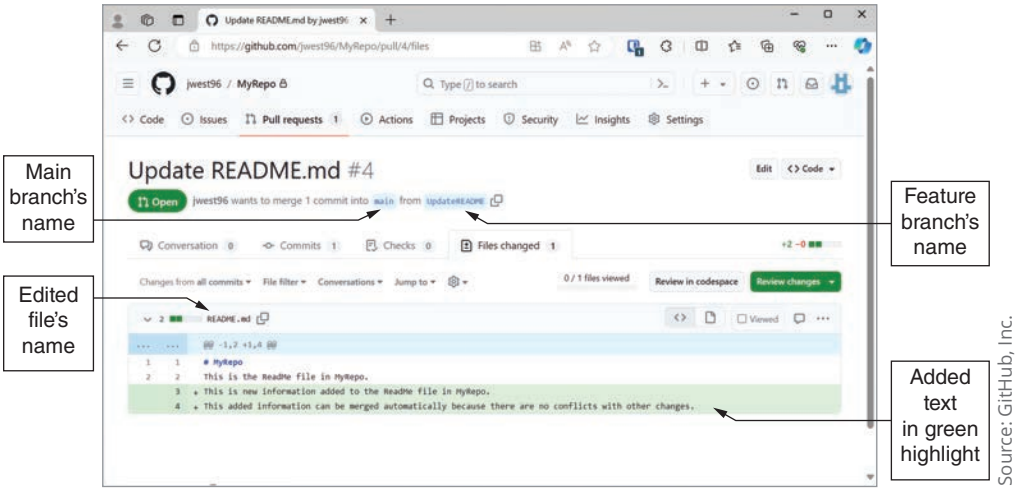
After uploading the files, you then create a pull request with the `git request-pull` command, which lets the repository owner know to review your changes and choose whether to finalize the merge. The remote repository owner then reviews the proposed changes and decides whether to accept the changes. During this code review, additions to the code are highlighted in green (Figure 3-13), and deletions from the code are highlighted in red. Git also identifies whether there are conflicts with the existing code, which might happen if two developers were working on the same portion of the code at the same time. These conflicts must be resolved by a human before the system can complete the merge.

Figure 3-12 Ensure changes are approved before merging with remote repository



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

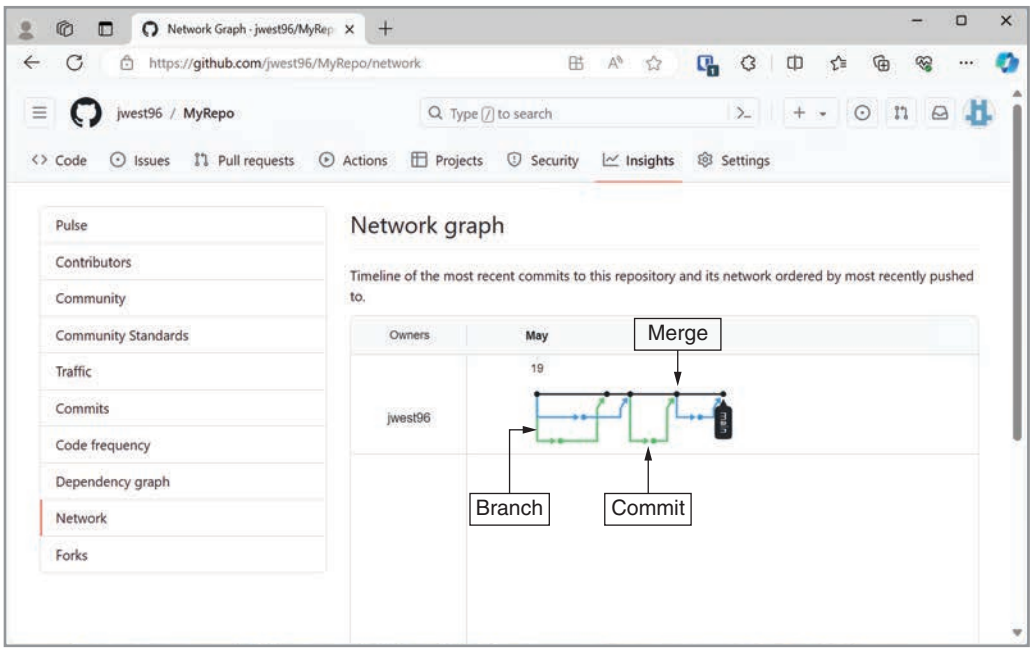
Figure 3-13 Review changes from a pull request before approving a merge



Note 3-5 | The concepts of pull and push don't necessarily correspond to the concepts of download and upload. You have to think about pulling (that is, movement toward) and pushing (that is, movement away from) relative to where you're starting. For example, if you're the owner of the remote repository, a developer on your team might submit a pull request using the `git request-pull` command. This is a request for you to examine the suggested code changes and decide whether to incorporate the new code into your remote repository. The "pull" movement in this case is movement *toward* the remote repository *from the perspective* of that repository, even though this data is being uploaded from the developer's local repository.

Note that many developers these days use a GUI (graphical user interface) platform rather than working with the command line directly, but the functions and overall process are still the same. GUI platforms such as GitHub also offer statistics and charts to visualize activities in the repository. For example, the network graph in Figure 3-14 illustrates recent branches, commits, and merges.

Figure 3-14 Network graph shows a timeline of branches, commits, and merges



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Although multiple steps like this can be helpful in fully processing edits (such as testing and validation) before those edits impact an application, some steps can be automated. For example, when a commit is pushed to the remote repository, you can automate certain tests to run on the edits before those edits are incorporated into the repository's files. These automated workflows can be configured and managed through **GitHub Actions** or with a similar tool called **Jenkins**. Where GitHub Actions functions within the context of the cloud-hosted GitHub platform, Jenkins can be self-hosted on a local computer. Jenkins is also compatible with many other platforms while GitHub Actions is limited to GitHub repositories. As a result of both these differences, Jenkins tends to be more complex to set up while offering many more options for customization.

Deployment Strategies

When code changes are accepted and ready to be released into production, developers must take caution when deciding how to release these changes. Especially when large changes are made, a slow deployment approach can catch problems before they might affect large numbers of users. Consider the following four techniques:

- **Blue-green deployment**—With this technique, one environment enters a passive state and receives updates and follow-up testing while another environment identical in size and configuration handles the active workload. When the updates are complete, all traffic is switched to the updated environment.
- **Canary deployment**—In this scenario, a small number of servers are updated, and a small amount of user traffic, perhaps 1 or 2 percent, is directed to those updated servers as a test run. If all goes well, the updates are then deployed to the rest of the fleet.
- **Rolling deployment**—Also called a rolling update, in this scenario, traffic is directed away from a few servers at a time so they can be updated. As each server is updated, it begins to handle traffic again so the next wave of servers can be updated. Eventually, the entire fleet has received the update.
- **In-place deployment**—With this technique, an update is applied without requiring the creation of any new infrastructure. Instead, the application on each host is stopped, replaced with the new application or an updated version, and then that application is started. The infrastructure itself is not affected.

Testing Types

Performing adequate testing is a critical part of any system changes, both before and after the release of those changes. The following list compares a variety of commonly used testing techniques:

- **Functional testing**—**Functional testing** ensures that predefined system requirements are met. Basically, this process confirms everything works as it should. For example, functional testing can confirm whether users can successfully log into a system when using valid credentials. Functional testing might also check whether appropriate error messages are presented when the user inputs incorrect credentials or whether the "forgot password" feature works as intended. Some approaches to functional testing assess an individual component, which is called component testing or module testing. Further, some kinds of functional testing include initial checks during development, called smoke testing (based on the saying, "Where there's smoke, there's fire," and implies a small problem early on could cause a serious issue later).
- **Performance testing**—**Performance testing** challenges a system to determine its responsiveness and stability under various conditions. It assesses system response times, throughput, resource utilization, and stability. For example, performance testing a website that uses servers distributed in several global regions might give latency and throughput data for requests from different locations.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Load testing**—**Load testing** pushes the limits of a system to ensure it can handle the upper end of its anticipated workload and to find out how well the environment auto scales as user demand increases and decreases. For example, load testing on an e-commerce website might include simulating high volumes of holiday shoppers during a sale. Load testing can be done manually or automatically and often involves both methods. The data collected during load testing informs auto scaling threshold configurations. Carefully plan and monitor this phase of testing to avoid accruing unnecessary costs from auto scaled resources.
- **Regression testing**—When you change one system, this can affect other systems in unexpected ways. **Regression testing** confirms that changes to one system haven't negatively impacted other functions within that system or other systems in the network. This type of testing is not necessarily limited to the system under development—interdependent systems might also be tested to determine if the new system's changes will have unintended effects throughout the network. For example, suppose you update the payment processing system in an e-commerce website's shopping cart. Regression testing ensures that other components of the shopping cart and website continue to work as expected, even after the payment processing system was updated.
- **Usability testing**—During the development process, **usability testing** asks users to test the system while developers observe the users and the system. This is a form of black-box testing and might also be called acceptance testing or UAT (user acceptance testing). The acquired information helps developers identify problems that might cause confusion for users. For example, are instructions clear? Do components function the way users expect? Does the system perform well when nontechnical users, who don't have knowledge of the background design and processes of the system, interact with it? Phases of usability testing might include alpha testing or beta testing.
- **Security testing**—Security testing includes **vulnerability testing**, which checks for application liabilities or gaps in protection of sensitive data, and **pen (penetration) testing**, which attempts to take advantage of vulnerabilities. For example, suppose you want to test your cloud deployment for its ability to detect and resist a ransomware attack. Testing should confirm that unusual encryption activities in your accounts raise alarms or trigger other automated responses. However, some kinds of security testing in the cloud require special permission from the CSP. Before performing any security testing in the cloud, carefully study the CSP's guidelines and restrictions on testing and get any necessary approvals from the CSP. Unauthorized security testing could result in loss of use from the provider or severe financial penalties.

Exam Tip | The CompTIA Cloud+ exam is known to ask questions requiring you to differentiate between testing techniques, when each is appropriate to use for a specific objective, or what issue a particular test type might reveal.

Testing in the cloud often requires a different set of tools from what is used for testing in a physical environment, especially for load testing that is designed to ensure auto scaling. Many CSPs offer testing tools optimized for their platforms.

Testing might also require help from users who are familiar with the resources being evaluated. For example, an accounting application is best understood by the accountants on staff and so can best be evaluated and challenged by a tech-savvy accountant. Coordinate testing efforts so the appropriate team members are available to process feedback as it comes in from these users.

As you design your testing plan, anticipate how to analyze the results that you'll obtain. The information you gather can be compared to any standards defined by your change management documentation and the SLA provided by your CSP. You'll need to refer to this document occasionally during the course of any service subscription you maintain, so be sure you thoroughly understand the commitments made by the vendor and your options for recourse should performance fall short of guaranteed minimums or maximums.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

6. Which command performs the final merge of a commit into the main branch of code?

   a. `git add`
   b. `git branch`
   c. `git commit`
   d. `git merge`

7. What deployment type does not require replacing infrastructure?

   a. Blue-green
   b. Rolling
   c. In-place
   d. Canary

8. What kind of testing would detect unintended changes to an interdependent system?

   a. Usability testing
   b. Regression testing
   c. Vulnerability testing
   d. Performance testing

☐ Check your answers at the end of this chapter.

Section 3-4: Planning for Problems

Certification

1.2	Explain concepts related to service availability.
1.8	Summarize cost considerations related to cloud usage.
3.2	Given a scenario, configure appropriate scaling approaches.
4.3	Given a scenario, implement identity and access management.
6.1	Given a scenario, troubleshoot deployment issues.

In later chapters, you'll learn more about types of problems you might encounter in the cloud with various types of cloud services, and you'll learn about some of the tools available to you to address these issues. This section discusses capacity-related issues and steps you can take to ensure your cloud services will continue functioning even when something goes wrong.

Common Deployment Issues

Despite the best-laid plans, things still go wrong in the cloud. Knowing more about what to expect and how to handle the more common issues will prepare you to manage problems quickly and minimize disruption of services. Common issues when deploying cloud resources include the following:

- **Resource contention**—Resource contention refers to the demand many workloads make on limited resources, such as processing power, memory, or network bandwidth. On-prem, it's your responsibility

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to provide sufficient hardware and network support for the workloads you're running. In the cloud, you rely on the CSP to provide sufficient underlying hardware resources so that other customers' workloads don't interfere with your own ability to receive reliable and dependable service. If your CSP is not providing this level of service, you might find yourself needing to migrate to a different CSP. That said, most often, the problem of resource contention results from misconfigured resources. If a cloud-hosted resource is not functioning at the desired level, evaluate its configuration for ways to improve its performance, such as increasing its allocated memory, storage, network bandwidth, or processing power. Benchmark testing before, during, and after migration can help minimize the likelihood of resource contention being an issue.

- **Template misconfiguration**—When creating or migrating servers, template misconfigurations can have serious implications for security, in addition to concerns about functionality and connectivity. Leaving remote access channels such as SSH (Secure Shell) or RDP (Remote Desktop Protocol) open to the Internet makes servers vulnerable to attack, as does missing software updates or leaving wide-open permissions. There are many tools available to help monitor and audit server configurations, such as AWS's Config and CloudTrail services. Server images should be scanned when first created and again any time a change is made. Regular configuration audits after migration are also essential. Image management services, such as AWS's EC2 Image Builder, can help with building, testing, and updating images used to deploy servers. Later in this course, you'll learn about other methods of enforcing resource configurations using automation tools.
- **CSP or ISP outage**—Wide-reaching CSP outages, which could affect a single service or even the entire platform, are somewhat rare but, when they happen, they can deeply disrupt your organization's ability to do business. Similarly, while a CSP outage can affect your customers directly, an ISP (Internet service provider) outage can interfere with your own ability to reach your cloud services. Larger companies likely have redundancies, such as multiple ISP connections to their premises and multiple cloud-based resources to perform critical functions, and they possibly even have cyber-insurance to cover unavoidable losses. Smaller businesses, however, are more vulnerable to outages. Redundancies across platforms and across regions within a platform can reduce the negative impact of an outage, even a full outage. Additionally, data backups should be stored in a location separate from your primary CSP (such as in another platform or on-prem).
- **Platform integration issue**—While a hybrid or multi-cloud can help reduce risks related to a CSP outage or vendor lock-in, integrating services across multiple platforms also introduces compatibility issues. Many CSPs and other vendors offer products that help address these issues. The reliance of cloud services on APIs helps open communication between platforms, and new tools are being developed to take advantage of this interactive layer. Cross-platform staff training is also a crucial component of an organization's ability to take full advantage of the features and tools offered by various cloud platforms.
- **Licensing outage**—Many kinds of licenses, such as subscription licenses for Microsoft 365 or pay-as-you-go licenses for AWS services, require ongoing payments to maintain the license. If the bill doesn't get paid, the licenses expire. This might seem like a minor thing, and yet it can cause major problems. SLM (software license management) can help minimize license expenses by eliminating unneeded licenses and renewing licenses before expiration, thereby avoiding fees or legal fines. Keep license documents and receipts in a central repository (not your email inbox). Set reminders or appointments in your calendar or schedule email reminders to yourself to renew or update licenses periodically.
- **Time synchronization issue**—Due to irregularities in Earth's rate of rotation, occasionally an extra second must be added to UTC (Coordinated Universal Time), which is called a leap second. These leap seconds can offset the accuracy of network and system clocks. To increase synchronization accuracy across cloud-based resources, many CSPs offer a time service. For example, AWS's Amazon Time Sync Service operates by default on all EC2 instances and connects through the internal IP address 169.254.169.123. Time services rely on NTP (Network Time Protocol) to sync clocks across a network and smooth out these leap seconds in a process called leap smearing.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

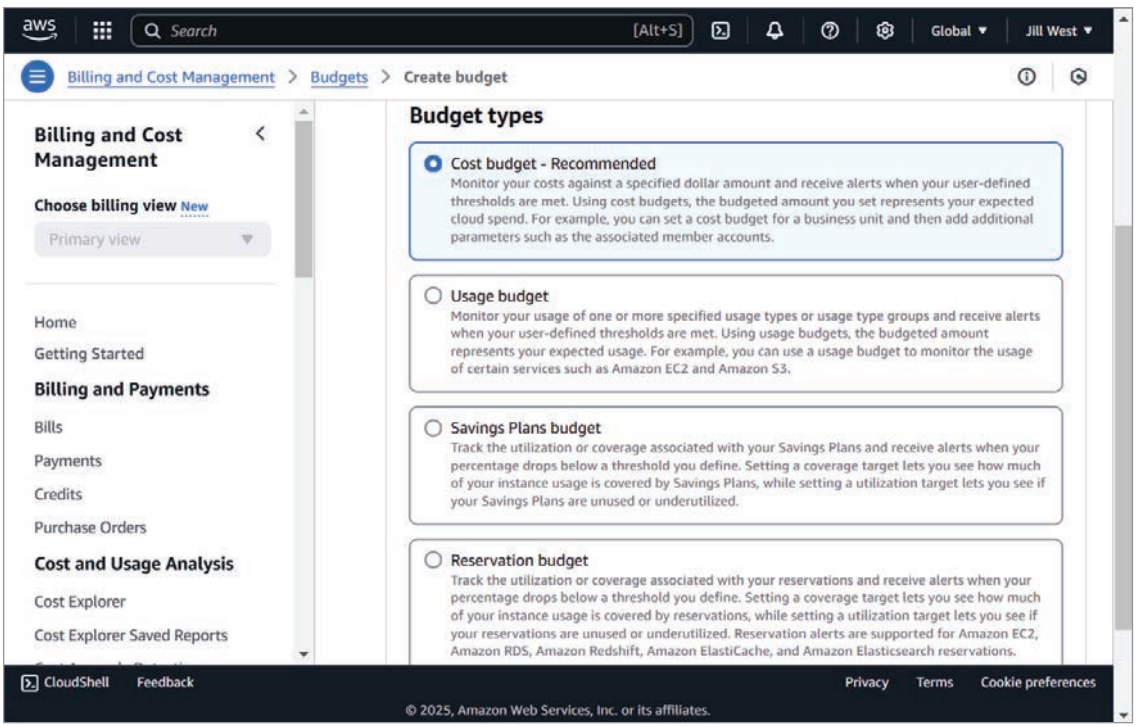
Note 3-6 | The abbreviation UTC for universal time was agreed upon as a compromise between English and French authorities. English speakers wanted the abbreviation CUT (coordinated universal time), and French speakers wanted TUC (*temps universel coordonné*).

Sizing Issues

While the cloud itself might be capable of supporting almost unlimited capacity for compute, storage, and networking resources, your budget might not be so flexible. Additionally, there are other limits on how far a cloud service can expand to support increased demand. Limitations on cloud capacity might be imposed by any of the following sources:

- **Service quotas and throttles**—As you've been completing projects already in this course, you might have noticed resource limits defined on various services. For example, EC2 customers are allowed to run up to a certain number of vCPUs per instance family in each region. These vCPUs can be spread across instances of any size within each family so long as the total running vCPUs at any one time does not exceed the maximum for that instance family. Often, you can request an increase to these limits by contacting the CSP. With SaaS products, such as Salesforce, you'll also run into limits on API requests. Recall that APIs allow for standards-based communication between different services, and a product such as Salesforce might be integrated via API with other custom applications. In that case, an API request limit throttles communications with the provider. This way, they're not overwhelmed by one customer's API activities and can maintain optimal performance for all customers.
- **Account owner caps**—As the account owner, you can cap usage and expenditures within your account across users. Some cloud platforms will allow you to configure budget restrictions according to service, while others will only allow you to set overall spending limits. Figure 3-15 shows configuration options for a cost budget in AWS. Budget requirements might also dictate a limit on license purchases, which restricts the number of users who can use software or other services. As the number of users varies, the budget and resource allocations must be increased to keep pace with user demand.

Figure 3-15 AWS offers four budget types: cost budget, usage budget, savings plans budget, and reservation budget



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

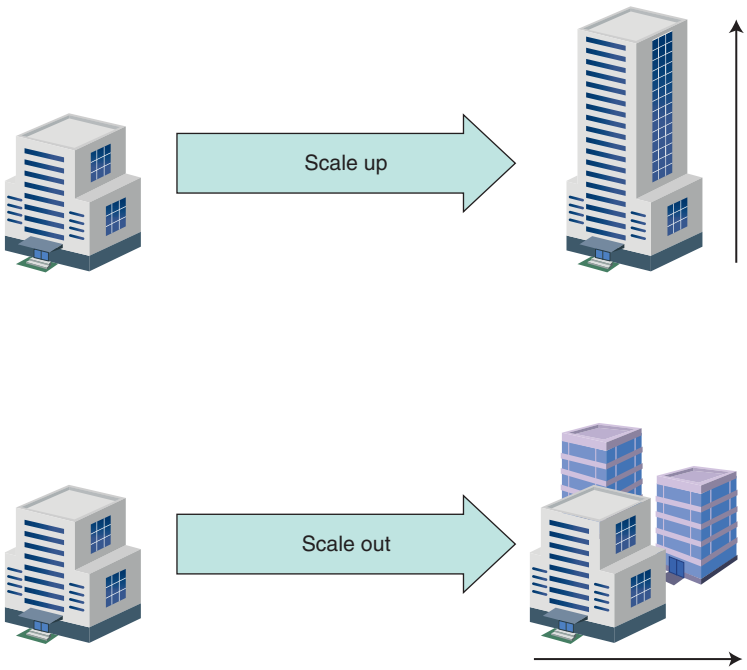
- **Technology limitations**—While many limitations are set due to the cloud consumer's budgetary restrictions or the CSP's need to balance resources among customers, many other limitations are simply due to physics. For example, bandwidth limitations increase latency. Throwing money at this problem won't change the speed of light, the geographical distance between data centers, or the amount of data that can be carried across a cable at any given time. Similarly, IP address limitations require careful planning of IP address spaces. Also, choosing a particular VM type limits the capability of the hardware (virtual and physical) supporting those instances.

Capacity Planning

Although the cloud offers nearly unlimited possibilities for experimentation and expansion, usable cloud capacity does have its limitations. This reality requires planning and prioritizing on your part. You can then avoid problems such as VM sprawl (where forgotten instances are left running) and misappropriated or underutilized resources (such as unassigned public IP addresses or storage volumes that aren't holding much data). The following resource balancing techniques can help you better plan for cloud capacity allocations:

- Refer to your SLAs. Whether you're working under a commitment to your organization or to your customers, plan for your cloud's capacity to exceed expectations defined by the SLA. This can be accomplished partially through business continuity planning and high-availability configurations. You can also moderately overprovision resources for optimal performance and schedule auto scaling in response to increasing demand.
- Balance reserved services, which provide stable performance, with on-demand services, which start in response to increasing demand. For example, recall that AWS, Azure, and GCP all allow you to reserve VM instances, meaning you can commit to pay for an instance over a one- or three-year period at a lower rate than on-demand instances that run only as needed. This arrangement can save you as much as 70 percent of the cost of an instance and is especially useful with workloads that require a predictable minimum resource demand. However, subscribing to more reserved instances than you need during that time will reduce or eliminate the potential savings. In some cases, you're also locking yourself into a specific level of technological sophistication (such as a processor's generation) or resource utilization (such as the number of processors). Subscribing to reserved instances reduces your ability to adapt to new technologies or changing workload needs during that time frame. It also locks you into that price point even if cloud service costs continue to fall.
- Establish utilization baselines and refer to these standards when determining initial resource allocation as well as when analyzing performance trends over time. These trends can help you forecast future resource capacity needs so you can fine-tune resource sizing. You can upsize—or increase the capacity of—resources in anticipation of increasing demand. Alternatively, you might find that shifting trends indicate a need to downsize, or decrease, resource capacity. Upsizing and downsizing might occur in any of the following ways (Figure 3-16):
  - Scale up (called **vertical scaling**) by increasing components of a resource, such as vCPUs, allocated to an existing instance or container.

Figure 3-16 Scaling up increases the size of an existing resource while scaling out increases the number of resource instances



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- Scale down by reducing components allocated per instance or container.
- Scale out (called **horizontal scaling**) by increasing the number of running instances or the number of available containers.
- Scale in by reducing the number of running instances or available containers.

Due to the convenience of scalability in the cloud, it's easy to fall into a pattern of asymmetrical provisioning of resources. You scale up or out to improve slow or under-provisioned resources, but then you forget to scale down or in when resources are over-provisioned. Elastic services (those that scale up/down or out/in automatically) can help with this problem. However, elasticity is sometimes charged as a premium feature. Habitually configure checkpoint alarms or alerts that will identify unused or underused resources. Budget alarms or spending limits can also help prevent unexpected charges, either from unplanned expansions where resources are auto scaling beyond budgetary allowances or from resources that are running without being used. Some CSPs also offer features such as auto shutdown for VMs. This can prevent ongoing charges for instances no longer in use.

Exam Tip | Vertical and horizontal scaling is a topic commonly covered by questions on the CompTIA Cloud+ exam. Be sure you thoroughly understand these concepts, the differences between them, and when to use one versus another.

Auto Scaling Types

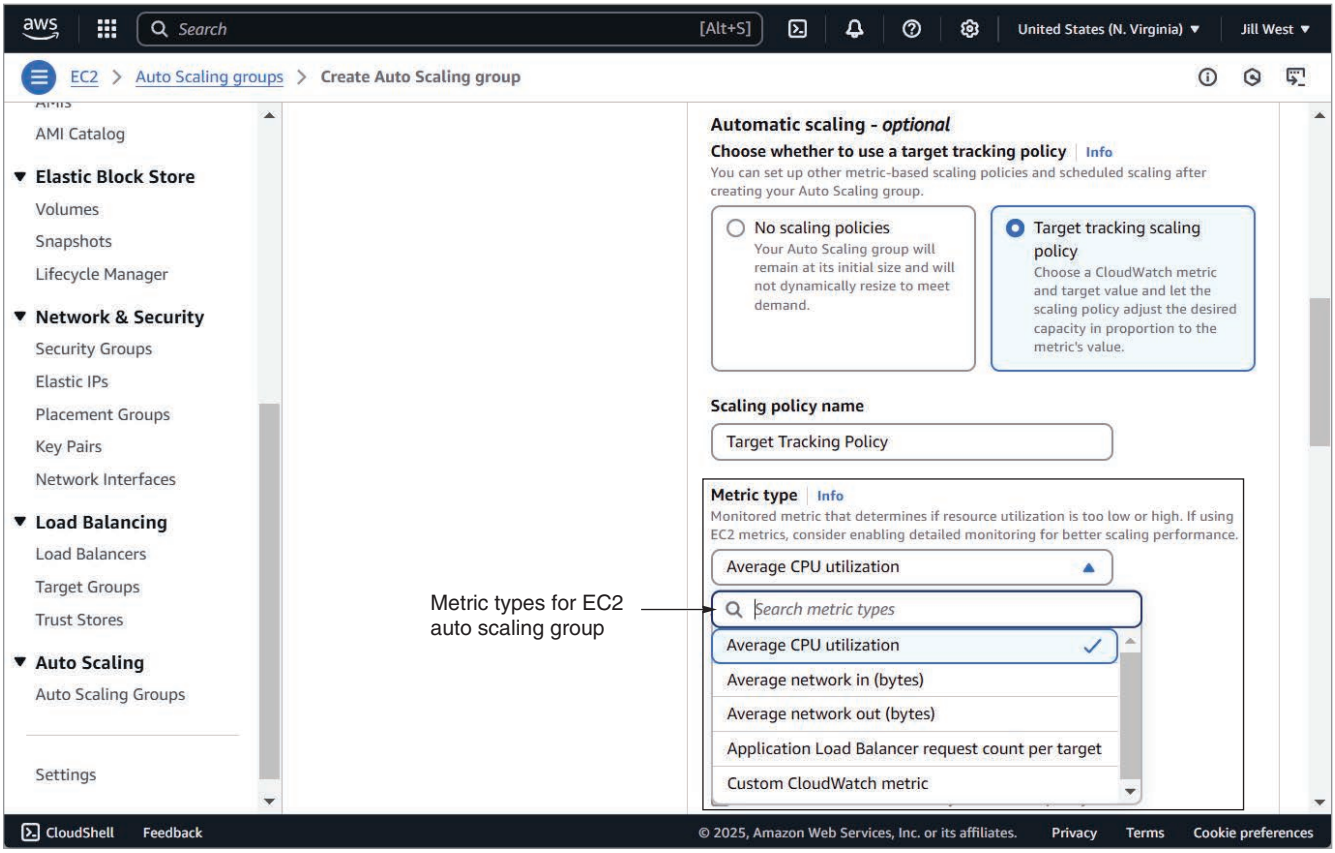
One of the great benefits of the cloud is the ability to scale resources to match fluctuating demand, and much of this can be done automatically. Auto scaling can be configured to occur automatically in response to specific triggers, on a scheduled basis, or you can scale resources manually. Common triggers for auto scaling include the following:

- **Schedule**—If demand fluctuates on a predictable schedule, you can configure auto scaling to match that schedule.
- **Load**—Most auto scaling is set to meet fluctuating demand based on the detected load being placed on the resources in the cluster. For example, CPU utilization is a commonly monitored metric used to determine when an auto scaling group of VMs should scale in response to changing demand. When the load on the existing instances' CPUs exceeds a specific threshold, the auto scaling group automatically adds more instances. When the load again decreases, instances are removed from the auto scaling group. This is a form of dynamic scaling. Figure 3-17 shows other metrics available in AWS for load-based auto scaling.
- **Trending**—Over time, patterns of fluctuations can possibly be detected and used to forecast auto scaling needs. This predictive auto scaling analyzes historical usage data to determine patterns, and then auto scaling can be performed shortly before the need occurs. This decreases latency and improves the user experience.
- **Event**—Event-based auto scaling is triggered when a specific event occurs. Sometimes, things happen within a system that necessitate the increase of resources in anticipation of increased need. For example, suppose an instance group processes images that have been uploaded to a storage bucket. The addition of images to the bucket is an event that can then trigger an auto scaling action, which results in the creation of the instances that will process those images.

In deciding when to use each of these approaches to auto scale, consider the needs of the system and its users, common utilization patterns, and anticipated growth. One caveat—manual scaling should almost never be performed on a production system. As you learn more about automation in the cloud, you'll find that manual configuration changes of all types should be avoided in production environments, including scaling. Instead, scaling should nearly always be performed automatically.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3-17 Metric types available for a metric-based scaling policy in an EC2 Auto Scaling Group



Source: Amazon Web Services, Inc.

AI in IT | One of the many benefits of designing cloud systems with microservices is that much of the automation in these systems can be powered by AI (artificial intelligence) for improved performance, security, and cost optimization. AI can manage auto scaling processes, self-healing mechanisms, and dynamic load balancing across resources in a cluster. Some AI systems can even perform predictive analytics on cloud traffic, anticipating bottlenecks, traffic spikes, and system failures to enable proactive measures that prevent or reduce the impact of these issues. Some cloud platforms, such as AWS, are incorporating AI services that can analyze application requirements and suggest design changes that would further optimize system performance.

Business Continuity Planning

The worst disaster is the one you didn't prepare for. A disaster could be anything from a localized disruption, such as a break-in or building fire, to a regional crisis, such as a flood, hurricane, or act of terrorism. Additionally, these events might be natural disasters, such as a forest fire, or human-made disasters, such as arson. While you hope that none of these things will interfere with your business activities, chances are, you'll have to face one or more significant crises at some point in your career.

At the same time, it doesn't take a disaster to interfere with your IT systems' abilities to function and meet the demands of your business. Even a small failure, if not compensated for, can "break" the system and cost your company money or even cost you your job. To protect your company's IT resources—and to protect your employment—do the work ahead of time to prepare for these eventualities.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

BC (business continuity) refers to a company's ability to weather a failure, crisis, or disaster of some kind while maintaining continuity of operations, especially for critical services to customers and income-generating activities. Part of the BC planning process is developing a BIA (business impact analysis) document. This document evaluates how different departments and components of an organization's business would, if lost, affect the company's bottom line. Corporate guidelines will help in the decision-making process to determine what services should be prioritized for BC protection and how much money should be invested in protecting those resources. These terms might be spelled out in an OLA (operational-level agreement), which is a document that supports the standards defined by the SLA. Although simple redundancies and replication will account for many of the lower-grade incidents, such as a failed server or the proverbial backhoe cutting your connection to the Internet, more extensive planning is required in preparation for larger-scale crises.

A BCP (business continuity plan) contains documentation to support these standards and might include the following details:

- Contact names and phone numbers for emergency coordinators who will execute the response in case of disaster, as well as roles and responsibilities of other staff
- Details on which data and servers are being backed up, how frequently backups occur, where off-site backups are kept, and (most importantly) how backed-up data can be recovered in full
- Details on network topology, redundancy, and agreements with national service carriers, in case local or regional vendors fall prey to the same disaster
- Regular strategies for testing the BCP
- A plan for managing the crisis, including regular communications with employees and customers, taking into consideration the possibility that regular communication modes (such as phone lines) might be unavailable

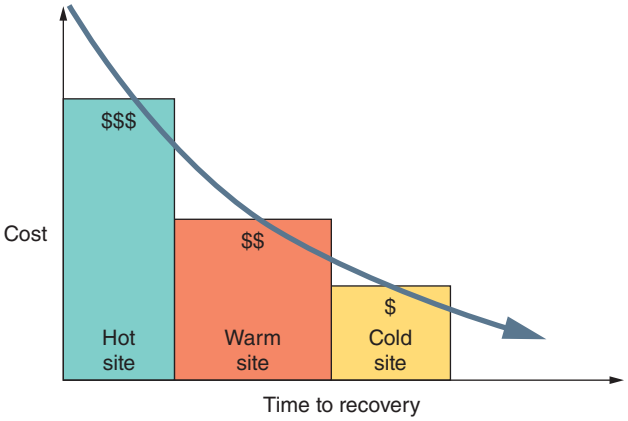
Organizations employ many strategies as part of their BCP, such as the following:

- **Roles and responsibilities**—Assigning roles and responsibilities to specific people ensures that everyone knows what is expected of them and how to perform those tasks. For example, the team leader is expected to serve as the final decision maker during an incident while the BC planner is responsible for developing and maintaining plan documentation. These assignments can also help identify gaps in coverage of tasks or gaps in skills or knowledge needed by everyone.
- **Call trees**—Recall that part of BC planning involves keeping a list of contact names and information for emergency coordinators. A more formal structure for this information is a call tree, which defines who should be contacted, in what order, and how to contact them. Automated call tree systems can perform these calls, texts, or emails automatically.
- **Incident types and categories**—Although you can't anticipate every possible event, you can develop categories of incident possibilities to inform your plans and preparations. Categories might be organized by the type of response needed, the severity of the event, or the resources affected.
- **Training**—Regular training ensures affected people know what to do and how to do it whenever an incident occurs for which they must respond. This training doesn't necessarily apply only to IT personnel. For example, an emergency security alert system for instructors at a school requires training to ensure faculty know how to use the system and under what circumstances. A localized incident, such as a medical emergency for a student in a classroom, might require a different response from affected faculty and security personnel than a more global incident, such as someone bringing an unauthorized weapon onto campus.
- **Backup services**—A company might team up with partners or hire third parties to help with organizing backup resources. A backup service can manage data backups, or a business partner might arrange a sharing agreement where each partner could provide backup equipment or other IT services should the other partner suffer a severe loss. These strategies hinge on redundancy, such as duplicate or triplicate connectivity to the Internet and other off-site resources, or backup and restore policies to compensate for loss of availability of other services, such as communications or power. Some forms of redundancy include alternate sites, such as company-owned edge sites where services are located at remote

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

offices or in data centers closer to customers to reduce latency. Some edge sites also could serve as backups if needed. Outsourcing to third-party services or sites is also an option. Backup locations host backup hardware, data, and services in preparation for use during an emergency.

Figure 3-18 The most expensive option also provides the fastest recovery



Different types of backup storage offer various options for ease and timeliness of restoration. The following three categories provide a general description of the options (Figure 3-18):

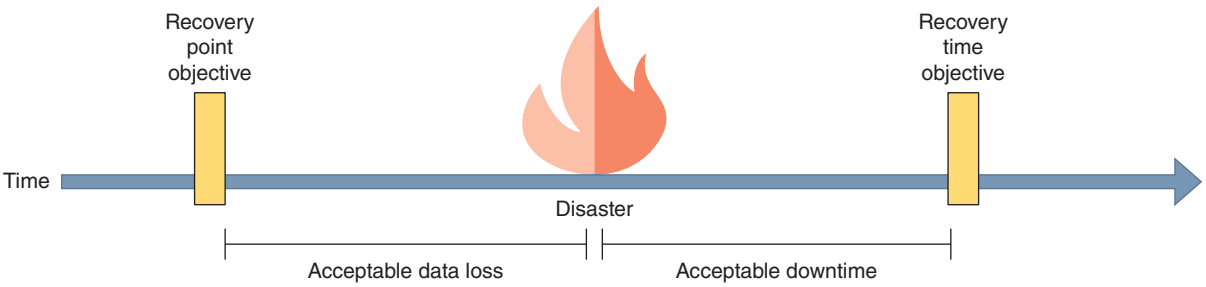
- **Cold site**—Computers, devices, and connectivity necessary to rebuild a replacement network exist, but they are not appropriately configured, updated, or connected. Therefore, restoring functionality from a cold site could take a long time.
- **Warm site**—Computers, devices, and connectivity necessary to rebuild a replacement network exist, with some pieces appropriately configured, updated, or connected. Recovery using a warm site can take hours or days, compared with the weeks a cold site might require. Maintaining a warm site costs more than maintaining a cold site, but not as much as maintaining a hot site.
- **Hot site**—Computers, devices, and connectivity necessary to rebuild a replacement network exist, and all are appropriately configured, updated, and connected to match your active network's current state. For example, you might use server mirroring to maintain identical copies of your servers at two locations. In a hot-site contingency plan, both locations would also contain identical connectivity devices and configurations and thus would be able to stand in for the other at a moment's notice. As you can imagine, hot sites are expensive and potentially time-consuming to maintain. For organizations that cannot tolerate downtime, however, hot sites provide the best disaster recovery option.

These backup sites are designed primarily as backups for on-premises data centers. However, in the cloud, you will also find the concept of hot, warm, and cold backups. The "temperature" is determined by how quickly the backup can be engaged in the event a primary resource fails. To understand these cloud-hosted backup services, let's first discuss some metrics used in disaster recovery.

Disaster Recovery

"Business continuity" and "disaster recovery" are two related but distinct terms. While BC refers to an organization's ability to continue doing business *during* a crisis, DR (disaster recovery) refers to the process of getting back to normal *after* the crisis is over. Typically, DR is considered to be a subset of BC planning. Two concepts commonly used to measure disaster recovery effectiveness are RPO (recovery point objective) and RTO (recovery time objective). Consider the diagram in Figure 3-19.

Figure 3-19 RPO defines how much data loss is acceptable, whereas RTO defines how much downtime is acceptable



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The RPO shows at what point in the past data will be recovered from. Data that was created or changed since that point will be lost because data backups were not copying data in real time. For example, relying on a full backup that's created only once a week will result in multiple days' worth of lost data if the disaster happens later in the week. However, if you're also keeping partial backups each day, your RPO would lose less than a day's worth of data regardless of when a disaster occurs during the week.

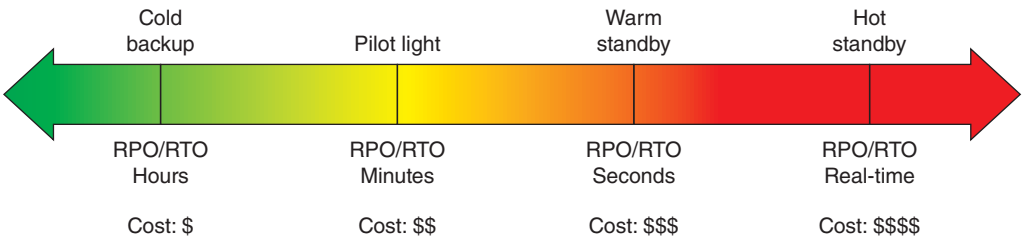
In contrast, the RTO shows at what point in the future full functionality will be restored, less any lost data. If you're creating a full backup each month and a partial backup each day of all changes since the last full backup, you can apply the full backup and the latest partial backup and you're done. If, however, your daily backup only contains changes since the previous day's daily backup, then you'll have to apply every daily backup since the last full backup, which will take more time. While smaller backups take up less space and take less time to create, they require more time to recover and therefore result in more recovery time needed after a loss.

RPOs and RTOs in the cloud have the potential to reach near-zero numbers. Cloud DR options open many new avenues of preparedness with much lower costs due to the pay-as-you-go nature of public cloud services. DRaaS (Disaster Recovery as a Service) can greatly reduce RTO because it doesn't take nearly as long to start new VMs from existing images as it does to reinstall OS images onto physical servers. RPO also benefits from the cloud due to inexpensive data replication and storage options. DRaaS is available natively in most major cloud platforms and is also widely available from third-party providers.

When planning for your cloud RTO, however, also consider bandwidth or ISP limitations and costs for transferring files from backup locations back into your main data center.

Figure 3-20 shows the spectrum of recovery tiers available in AWS to illustrate the relationship among RPO, RTO, and cost. These tiers are described as follows:

Figure 3-20 Various tiers of recovery services



- Cold backup services are relatively inexpensive. Perhaps you're storing regular snapshots of VMs and databases, but none of these resources is running in duplicate. This means you'll need hours to recover from a loss.
- AWS developed a new DR approach called pilot light that allows you to design the backup resources you need and leave them in a stopped state. This way, you're not charged as much as what you would pay for running services. You might keep only the database running so it's regularly updated, and you might periodically update server images, but these are not left running indefinitely. Recovery requires the amount of time needed to start these replacement services, which might be measured in minutes or even seconds.
- Warm standby means copies are designed and created, and a scaled-down version of the environment is left running for immediate failover. This arrangement might buy you enough time to start the remaining resources for full backup services if needed.
- Hot standby provides real-time failover with no significant data loss. For example, site mirroring copies the current version of your website to a duplicate web server cluster. A load balancer or DNS (Domain Name System) server can transfer traffic over to the mirrored site instantly if the original web service becomes overloaded or goes down completely. Ideally, you would use an active-active, multisite configuration so that both sites are working and productive at all times. However, each site is large enough to handle the entire workload if the other site fails.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Deciding which of these DR tiers is most appropriate for your cloud services will be determined by corporate guidelines and by existing SLAs for disaster recovery scenarios, which will define expectations for service during a crisis. Make sure users understand the limitations of your DR arrangements, especially according to what is defined in the relevant SLAs. Consider the CSP's guidelines for DR planning, and incorporate other techniques as needed. Whatever BCP or DRP (disaster recovery plan) your organization develops, be sure to test it thoroughly two or three times a year, or more often if required by relevant compliance regulations. These tests might use any of the following techniques:

- **Tabletop exercise**—Involves a team of representatives from each department reviewing the plan together and searching for gaps
- **Structured walk-through**—Requires each department to walk through the plan in detail, usually with a specific disaster scenario in mind
- **Disaster simulation testing**—More involved and requires the presence and participation of all personnel, including business partners and vendors

Various models identify core phases of incident response. Some of these systems consist of four or five phases while others list six or even seven phases. Furthermore, the phases listed vary slightly from model to model. A common distinction between these models is whether preparation before the incident is included as a distinct phase.

Preparation tasks include planning for possible incidents and gathering the resources you would need to handle those issues. For example, a BCP tends to be lengthy and detailed in its written preparations and plans. However, this unwieldy document might not provide the most usable guidance in a moment of crisis. Instead, a playbook provides succinct, organized information that is easy to access and follow during the stress and urgency of an emergency. You might create a different playbook for each major type of incident, such as for a cybersecurity breach, a terror attack, a weather crisis, or a customer or employee protest.

As the playbook for each scenario is developed, it should address all of the following incident response phases, which begin when an incident occurs:

1. **Identification**—This phase begins when an incident is identified, such as a security breach. Initial identification of the incident includes defining the scope of the incident and activating a response. An automated response might include failover to backup systems where services are transferred to backups if their primary systems suffer problems.
2. **Investigation**—The investigation phase involves determining the initial trigger of the incident, tracing the effects of the incident, and evaluating the details of the incident, such as listing items that were stolen, data that was compromised, services that are down, and so on.
3. **Containment, eradication, and recovery**—The focus of this phase is to minimize the impact of the incident and restore services as soon as possible. Affected systems are isolated, services are restored to their primary systems (a process called a failback), and evidence is gathered. When acquiring and managing evidence, be sure to maintain a legally defensible chain of custody. All collected data must be carefully processed and tracked so it does not leave official hands at any point in the forensics process. Typically, documentation used to track **chain of custody** describes exactly what the evidence is, when it was collected, who collected it, its condition, and how it was secured. If at any point in the process you have custody of evidence, be sure to sign off on a chain-of-custody document and obtain a signature from the next person in line when you hand over custody of the evidence.
4. **Post-incident and lessons learned**—The saying "hindsight is 20/20" refers to the fact that reviewing what happened provides a clearer perspective than trying to foresee what *might* happen in the future. The post-incident phase, sometimes called the "lessons learned phase," takes advantage of this insight. After the incident is handled and services are restored, this phase requires a review of processes to determine how the incident happened and, most importantly, how to prevent similar problems in the future. Part of this process includes a **root cause analysis**, which is a formal

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

procedure that identifies all the contributing factors that allowed the incident to happen. Some root cause analysis techniques include the following:

- **Five whys**—Dig deeper into each possible explanation by repeatedly asking why; this helps determine if the suspected cause is instead a symptom of a deeper issue.
- **Cause map**—List what happened and, for each item, identify its cause, which might also have a cause, which might have a deeper cause. Connect all these items in a map to identify the root cause of the incident.

These discussions should also include reviews of the incident response procedures and evaluation of any needed changes to these processes.

Self-Check

9. Adding more RAM to a web server is what kind of capacity adjustment?

a. Scaling in
b. Vertical scaling
c. Scaling down
d. Horizontal scaling

10. Which kind of backup site provides the lowest RTO?

a. Hot site
b. Pilot site
c. Cold site
d. Warm site

11. What kind of DR test involves the most people?

a. Structured walk-through
b. Tabletop exercise
c. Disaster simulation testing
d. Playbook

Check your answers at the end of this chapter.

Summary

Section 3-1: Cloud Migration

- A cloud migration might involve moving resources from an on-prem data center to the cloud, from one cloud platform to a different cloud platform, or even from the cloud back to on-prem.
- Each organization will take its own unique approach to its cloud migration, depending on the scope of elements being migrated and the company's intent in making the transition.
- In choosing which approach to take for a specific resource, consider relevant factors both from a business perspective and from a technical perspective.
- During a migration to the cloud, you must consider restrictions that protect sensitive data as it's being moved and once it arrives at its new destination, including the physical location data is being moved to, as compliance requirements vary across geopolitical regions.
- Many IT professionals in the thick of a cloud migration find themselves leaning less on their technical knowledge and more on project management skills.

Section 3-2: Cloud-Native Design

- Microservices are fine-grained, purpose-built services that function interdependently as part of a larger system.
- Cloud architecture includes the interfaces between cloud resources so systems can communicate with each other.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 3-3: System Development

- The culture of reducing the isolation of work responsibilities and job roles in favor of increased collaboration among teams working on a product has come to be known as DevOps.
- The work of keeping a timeline and record of code changes is typically done with a version control system (sometimes called source control system) such as Git.
- Especially when large changes are made, a slow deployment approach can catch problems before they might affect large numbers of users, such as the following deployment strategies: blue-green deployment, canary deployment, rolling deployment, and in-place deployment.
- Performing adequate testing is a critical part of any system changes, both before and after the release of those changes; commonly used testing techniques include functional testing, performance testing, load testing, regression testing, usability testing, and security testing.

Section 3-4: Planning for Problems

- Common issues when deploying cloud resources include resource contention, template misconfiguration, CSP or ISP outage, platform integration issue, licensing outage, and time synchronization issue.
- Limitations on cloud capacity might be imposed by service quotas and throttles, account owner caps, or technology limitations.
- Although the cloud offers nearly unlimited possibilities for experimentation and expansion, usable cloud capacity does have its limitations, which requires attention to avoid problems such as VM sprawl and misappropriated or underutilized resources.
- Common triggers for auto scaling include predictable schedules, loads on various components, trending patterns, and events.
- Business continuity refers to a company's ability to weather a failure, crisis, or disaster of some kind while maintaining continuity of operations, especially for critical services to customers and income-generating activities.
- Two concepts commonly used to measure disaster recovery effectiveness are RPO and RTO.

Key Terms

For definitions of key terms, check the Glossary.

API (application programming interface) call
BC (business continuity)
CapEx (capital expenditures)
chain of custody
commit
DevOps (development and operations)
event
event-driven architecture
fan-out architecture
functional testing
GDPR (General Data Protection Regulation)
Git
GitHub
GitHub Actions
GraphQL
HIPAA (Health Insurance Portability and Accountability Act)
horizontal scaling
ITIL (IT Infrastructure Library)
Jenkins
load testing
local repository
loosely coupled architecture
OpEx (operational expenditures)
PCI DSS (Payment Card Industry Data Security Standard)
pen (penetration) testing
performance testing
project
project management
regression testing
remote repository
REST (representational state transfer)
root cause analysis
RPC (remote procedure call)
service discovery
SOAP (Simple Object Access Protocol)
SOC (System and Organization Controls) report
usability testing
vendor lock-in
vertical scaling
vulnerability testing
web service
WebSocket

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CRUD create, read, update, delete
DR disaster recovery
DSS Data Security Standard
GDPR General Data Protection Regulation
HTTP Hypertext Transfer Protocol
ISO International Organization for Standardization
ITIL Information Technology Infrastructure Library
NTP Network Time Protocol
PCI Payment Card Industry
RDP Remote Desktop Protocol
REST representational state transfer
RPC remote procedure call
RPO recovery point objective
RTO recovery time objective
SOAP Simple Object Access Protocol
SOC 2 System and Organization Controls 2
SSH Secure Shell
STAR Security, Trust, Assurance, and Risk

Review Questions

1. What kind of expense is the purchase of a new firewall device?
   a. CapEx
   b. OpEx
   c. TCO
   d. PCI DSS
2. What type of migration would support the transition of an application from an on-prem VM server to a cloud-hosted container instance if no code changes are required for the application?
   a. Revise
   b. Repurchase
   c. Rehost
   d. Rearchitect
3. You've just deployed an update to an HR application that is widely used throughout your company. Which type of testing should you conduct to ensure your scheduling and shift planning software has not been adversely affected by the HR application update?
   a. Functional testing
   b. Regression testing
   c. Beta testing
   d. Usability testing
4. You work as a software developer and need to release a small update to an application during a set maintenance window where short periods of downtime are acceptable, and no changes will be made to the underlying infrastructure. Which deployment strategy is the best fit for this scenario?
   a. Canary deployment
   b. Blue-green deployment
   c. In-place deployment
   d. Rolling deployment
5. During which incident response phase are backup systems activated?
   a. Identification
   b. Recovery
   c. Post-incident
   d. Investigation
6. Which resource might you use to quickly determine how to respond to a physical security breach?
   a. Chain of custody
   b. Call tree
   c. Playbook
   d. DRaaS
7. Which auto scaling policy would be the best fit for a workload that fluctuates according to a complex pattern consistently over time?
   a. Event
   b. Load
   c. Schedule
   d. Trending
8. Which tool is the best fit to set up a complex automation workflow for software in an on-prem environment?
   a. Jenkins
   b. GitHub
   c. Git
   d. GitHub Actions
9. Your team has decided to update the database technology used to handle your customer data. What architecture design allows you to replace one system in the application without disrupting the functionality of the rest of the application?
   a. Service discovery
   b. Loosely coupled architecture
   c. Event-driven architecture
   d. Fan-out architecture
10. Which auto scaling technique would be the best fit for a server cluster that needs to scale according to a complicated but predictable pattern?
    a. Load
    b. Trending
    c. Schedule
    d. Event

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Scenario-Based Questions

Question 3-1

Corin works in the IT department of a law firm specializing in real estate closings. The firm is already using G Suite (web-based Google apps) productivity software for office staff, and it recently acquired a company-wide license for Zoom, an online video conferencing app. The next, more audacious target is their billing software. The company has been using an antiquated program for more than 20 years—longer than most of their staff has been employed. One accountant has been with the firm for nearly 30 years and remembers when the firm first started using this billing software. She says everyone hated it even then. Corin has been asked to make a recommendation as to how to proceed with this phase of the migration. Given these conditions, what is the best approach to migrating the billing functions to the cloud?

a. Rehost
b. Replatform
c. Retire
d. Replace

Question 3-2

You're developing an application designed to assist an HR (human resources) team with handling employee data. The application's API must support stateless requests and requires caching to manage performance. Which API type would be the best fit for this application?

a. XML-RPC
b. REST
c. JSON-RPC
d. SOAP

Question 3-3

Mykel and his team recently completed a major migration of several business-critical applications to the public cloud, and they're performing some final tests before going live. While taking a short break at a nearby coffee shop, Mykel decides to remote into one of the primary servers from his personal laptop so he can make additional configuration changes. As he's completing the changes, it occurs to him that this particular server is supposed to be locked down to provide access only from his computer back at his desk. He pokes around to find out how much access he has with this user account and then returns to the office to address the problem. What kind of testing did Mykel use to identify this issue?

a. Vulnerability testing
b. Penetration testing
c. Performance testing
d. Load testing

Question 3-4

Ramona's company is planning to expand internationally through a business acquisition next year. Through this acquisition, the company hopes to attract several larger international accounts that will require higher levels of security compliance than what her company has tackled in the past. Due to this anticipated increase in security needs, Ramona has been put in charge of increasing the company's cybersecurity posture. Which audit-based framework will ensure the cybersecurity improvements of Ramona's company meet the demands of their expanding international market?

a. ISO/IEC 27001
b. SSAE 21
c. PCI DSS
d. GDPR

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hands-On Projects

Note 3-7 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Project 3-1: Create an Auto Scaling Group in AWS

Estimated time: 20 minutes
Objective 3.2: Given a scenario, configure appropriate scaling approaches.
Resources:

- AWS account (either a standard account or an AWS Academy account)
- Internet access

Description: Major public cloud platforms offer a variety of options for auto scaling resources. In this project, you'll create an AWS ASG (Auto Scaling group) to manage the number of running EC2 instances in your group. To do this, you'll first need a launch template, which will determine the configuration of each of the instances in your ASG.

Note 3-8 | During this project, you'll create an auto scaling group containing two EC2 instances. Plan to complete this project in one sitting so your instances do not run longer than necessary. This will minimize the impact of these instances on the use of your credits or free-trial limitations.

In your AWS console, complete the following steps:

1. Go to the **EC2** dashboard and click **Launch Templates** in the left pane. Click **Create launch template**.
2. Give the launch template a name, such as MyTemplate. Check the box for Auto Scaling guidance, which identifies fields that are required of the launch template for it to be used in an ASG. What did you name your launch template?
3. Choose the **Amazon Linux** AMI. Select the t3.micro instance type. Under Security groups, select an existing launch-wizard security group. When you're ready, click the **Create launch template** button.
4. Return to the **EC2** dashboard. Scroll to the bottom of the left pane and click **Auto Scaling Groups**. Click **Create Auto Scaling group**.
5. Give the ASG a name, such as My-ASG. Select the launch template you just created. Click **Next**.
6. Notice you have the option to override the launch template settings. Choose an Availability Zone, leave all other default selections on this page, and click **Next**.
7. Notice you have the option to integrate your ASG with other services, such as a load balancer. You'll learn more about load balancing in the cloud in later chapters. You also have the option to set up health checks, which would monitor instances within your ASG and replace them if they experience detected problems. Leave all default selections on this page and click **Next**.
8. The group size options allow you to determine the desired capacity of your ASG. For desired capacity, enter **2**. This will run two instances in your ASG under normal circumstances.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9. The scaling options allow you to determine the minimum and maximum capacity of your ASG. Leave **1** as the min desired capacity, and enter **3** as the max desired capacity. Because you want automatic scaling, check the option for **Target tracking scaling policy**. The default metric type to determine when the ASG scales out or in is Average CPU utilization. What are two other metrics you could use instead?
10. Leave the metric type as Average CPU utilization. The instance maintenance policy allows you to choose a replacement behavior when an instance in your ASG needs to be replaced. What are two of the options for the instance maintenance policy?
11. Leave all other default selections on this page and click **Next**. Click **Next** on the following two pages to reach the Review page. When you're ready, click the **Create Auto Scaling group** button.
12. Once your ASG is fully created, click the name of your ASG in the list to go to the resource page. How many instances are running in your ASG?
13. Click the **Instance management** tab on your ASG's page. **Take a screenshot** showing the list of running instances; submit this visual with your responses to this project's questions.
14. Deleting an instance directly will cause the ASG to react to maintain the desired capacity. Go to the **EC2** dashboard, and then click **Instances** in the left pane. You should have two running instances in this list. Select one of the instances in the list, click **Instance state**, and click **Terminate (delete) instance**. Confirm that you want to terminate (delete) the instance.
15. Refresh your list of instances every 20–30 seconds for two or three minutes. You should notice the terminated instance shut down, but then the ASG automatically responds. What does the ASG do to maintain the desired capacity?
16. The only way to permanently delete all of an ASG's instances is to delete the ASG itself. Return to the list of ASGs, select your ASG, click **Actions**, and click **Delete**. Confirm that you want to delete the ASG. Verify the ASG and its instances are all deleted.
17. Go to the list of launch templates and delete your launch template. Make sure to properly sign out of your account.

Note 3-9 | It's important to develop good cloud hygiene habits and delete any resources you no longer need. This is an important habit that will help you minimize cloud expenses.

Project 3-2: Create a Virtual Machine Scale Set in Azure

Estimated time: 20 minutes
Objective 3.2: Given a scenario, configure appropriate scaling approaches.
Resources:

- Azure account and subscription
- Internet access

Description: In this project, you'll create an Azure VMSS (virtual machine scale set) to manage the number of running VM instances in your group.

Note 3-10 | During this project, you'll create a virtual machine scale set containing two VM instances. Plan to complete this project in one sitting so your instances do not run longer than necessary. This will minimize the impact of these instances on the use of your credits or free-trial limitations.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In your Azure portal, complete the following steps:

1. In the search box, enter **scale sets**, and then click **Virtual machine scale sets** in the search results.
2. Click **Create virtual machine scale set**.
3. Confirm the correct subscription is selected. Create a new resource group. What did you name your resource group?
4. Give your VMSS a name. In the Orchestration section, select **Uniform**. Note that student subscriptions are not allowed to choose autoscaling for the scaling mode; therefore, this project will use the manual scaling mode. Make sure the instance count is set to **2**.
5. In the Instance details section, choose an Ubuntu server image. Make sure the size is set to **Standard_B1s**. Leave all other default selections and click **Review + create**. Once the validation is passed, click **Create**.
6. Click **Download private key and create resource**. You will not need the key pair for this project.
7. Return to the list of VMSS instances and wait until your deployment succeeds. Click the name of your VMSS to go to the resource page, and locate a list of instances in your VMSS. How many instances are running in your VMSS?
8. **Take a screenshot** showing the list of running instances; submit this visual with your responses to this project's questions.
9. Select one of the instances, click **Delete**, and confirm the deletion. Refresh your instances list every 20–30 seconds for two or three minutes. Does the VMSS add a new instance to replace the deleted instance?
10. The VMSS manual scale setting was automatically adjusted to match your number of running instances. You can reset the manual scale. In the VMSS navigation menu, click **Availability + scale**, and then click **Scaling**. On the Configure tab, the Manual scale setting shows an instance count of 1. Change the instance count to **2** (type 2 into the field rather than using the slider, as the slider will jump from 1 to 4). Click **Save**.
11. Return to the list of instances in your VMSS. How many instances are running now?
12. Currently, there is no auto scaling configured for your VMSS. Return to the Scaling page and select **Custom autoscale**. Scroll down and select **Scale based on a metric**. In the warning message that appears, click **Add a rule**.
13. Under Criteria, click to display the Metric name list. By default, Percentage CPU is selected. What are two other metrics available in this list?
14. Close the Scale rule blade without making any changes. Discard the scaling changes.
15. Click **Resource groups** in the navigation pane, and then click the resource group you created in Step 3. Click **Delete resource group** and confirm the deletion.
16. Wait to confirm the resource group deletion has completed. If the notification icon still shows activity in the upper-right corner, wait a few more minutes, and then refresh the Resource groups list until your resource group is removed from the list.
17. Make sure to properly sign out of your account.

Note 3-11 | It's important to develop good cloud hygiene habits and delete any resources you no longer need. This is an important habit that will help you minimize cloud expenses.

Project 3-3: Create a Managed Instance Group in GCP

Estimated time: 20 minutes
Objective 3.2: Given a scenario, configure appropriate scaling approaches.
Resources:

- GCP account
- Internet access

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Description: In this project, you'll create a GCP MIG (managed instance group) to manage the number of running VM instances in your group. To do this, you'll first need an instance template, which will determine the configuration of each of the instances in your MIG.

Note 3-12 | During this project, you'll create a managed instance group containing two VM instances. Plan to complete this project in one sitting so your instances do not run longer than necessary. This will minimize the impact of these instances on the use of your credits or free-trial limitations.

In your GCP console, complete the following steps:

1. Create a new project and give the project an appropriate name. What did you name your project?
2. After the project is created, select the project and confirm you are working in your new project.
3. In the search box, enter **instance templates**, and then click **Instance templates** in the search results. Click **CREATE INSTANCE TEMPLATE**. You might need to enable the Compute Engine API to proceed. Wait until the API is enabled.
4. Give the instance template a name. For the location, make sure **Regional** is selected. For machine type, select **e2-micro**. When you're ready, click **CREATE**.
5. In the left pane, scroll down and click **Instance groups**. Click **CREATE INSTANCE GROUP**. Give your MIG a name and select the instance template you just created. Leave the location as **Single zone**.
6. Change the minimum number of instances to **2** and change the maximum number of instances to **4**.
7. Notice the default autoscaling signal is CPU utilization. What are two other signal types you could use instead?
8. Leave the autoscaling signal as CPU utilization 60% and leave all other default selections on this page. Notice you also have the option to set an autoscaling schedule and to set scale-in controls. When you're ready, click **CREATE**.
9. Once your MIG is fully created, click the name of the MIG in the list to go to the resource page. How many instances are running in your MIG?
10. Scroll to the VM instances section of the resource page. **Take a screenshot** showing the list of running instances; submit this visual with your responses to this project's questions.
11. Deleting an instance will cause the MIG to react to maintain the desired capacity. Select one of the instances in the list and delete the instance. What does the MIG do to maintain the desired capacity?
12. The only way to permanently delete all of an MIG's instances is to delete the MIG itself. At the top of the page, click **DELETE GROUP** and confirm the deletion. Verify that the MIG and its instances are all deleted.
13. Go to the list of projects and shut down the project you created in Step 1. Confirm the project is shut down and scheduled for deletion.
14. Make sure to properly sign out of your account.

Note 3-13 | It's important to develop good cloud hygiene habits and delete any resources you no longer need. This is an important habit that will help you minimize cloud expenses.

Challenge Projects

Note 3-14 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 3-1: Manage a GitHub Repository

Estimated time: 30 minutes (+15 minutes for group work, if assigned)
Objective 5.1: Explain source control concepts.
Group work: This project includes enhancements when assigned as a group project.
Resource:

- Internet access

Description: Increasingly, cloud technicians are expected to understand how code development works, how to use related tools, and even how to use these tools in managing cloud deployments. In this project, you will create a free GitHub account and practice managing a repository.

Complete the following steps:

1. In your browser, go to **github.com** and create a free account.
2. When you complete the account creation process, you should arrive at your Dashboard.
3. Click the **Create repository** button.
4. Give the repository a name, such as my-cloud-repository. What name did you choose?
5. Select **Private** to keep your repository private.
6. Keep all other default settings and click **Create repository**.

You've created a private repository, and now you're ready to create a file in your repository. Complete the following steps:

7. In the Quick setup area, click **creating a new file**.
8. In the file path, enter a name for your file. Most repositories should contain a README file, so this would be a good file name to start with. What file name did you choose?
9. In the file content area, enter some text. When you're finished, click the **Commit changes** button.
10. In the Commit changes dialog box, replace the default commit message with your own text. The message should be brief and explain the changes you're making. The Extended description section gives you an opportunity to add more detailed information. When you're ready, click the **Commit changes** button.

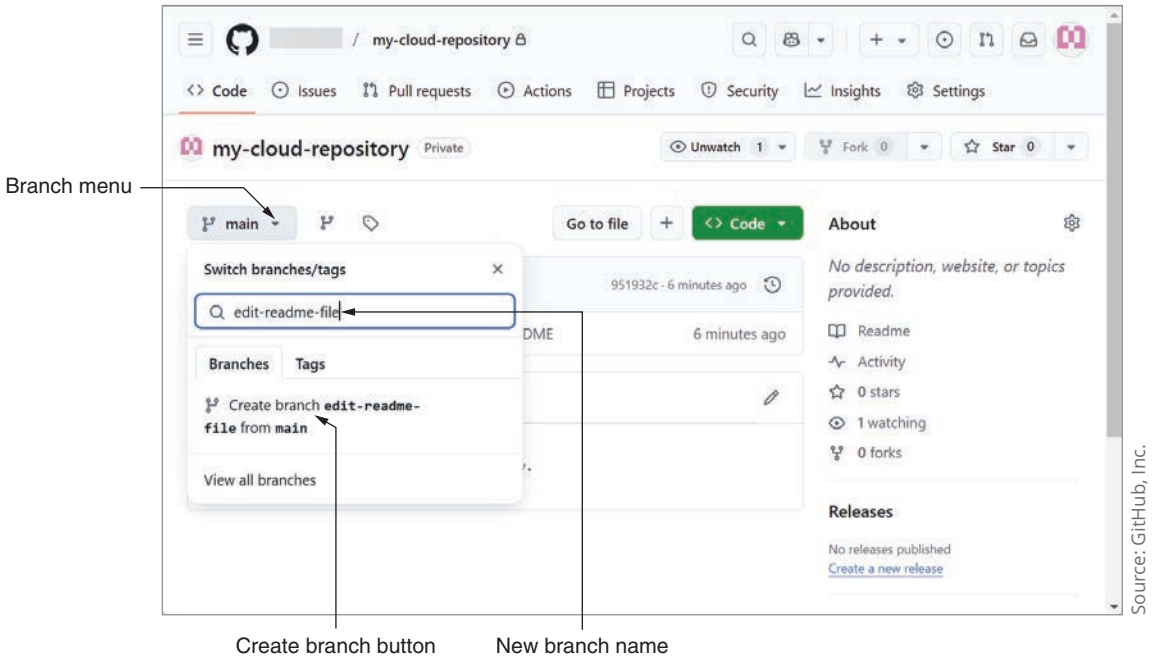
The change you just made—creating a new file—was completed within the main branch. Going forward, all changes should be made in an alternate branch and then submitted for approval. Complete the following steps:

11. Click the **Code** tab to list the contents of your repository. **Take a screenshot** of your repository showing your file name and contents; submit this visual with your responses to this project's questions.
12. Click the dropdown arrow next to the main branch (Figure 3-21). Enter a name for your new branch, such as edit-readme-file. What name did you give your new branch?
13. Click the **Create branch** button to create your new branch.
14. Make sure your new branch is selected in the branch picker. Then click the **Edit file** button on your file.
15. Add more text to the file. For example, you might describe what you are learning about GitHub as you practice with your first repository. When you're ready, click the **Commit changes** button.
16. In the Commit changes dialog box, replace the default commit message with your own text. The message should be brief and explain the changes you're making. The Extended description section gives you an opportunity to add more detailed information. When you're ready, click the **Commit changes** button.

This time, your changes are committed in the alternative branch, not in the main branch. For this reason, you need to create a pull request so the changes can be reviewed before being merged with the main branch. Complete the following steps:

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3-21 Create a new branch



17. Click the **Pull requests** tab. There should be a notification at the top of this tab saying your new branch recently received a push. Click the **Compare & pull request** button.
18. You can keep the default title. Add some more information in the Add a description area. When you're ready, click the **Create pull request** button.

Notice that GitHub evaluates the changes and detects that the changes can be merged automatically with no conflicts. If the changes would cause conflicts with other changes to the main branch, that information would appear on this page and the human admin would need to review the changes in order to merge the pull request. You also have the option here to add other reviewers or assignees to evaluate the pull request, and you can add comments for any additional information that should be tracked along with the pull request. Complete the following steps:

19. Click the **Merge pull request** button. Leave the default text for the Commit message, and then click the **Confirm merge** button.
20. Because no other changes are pending or needed for this branch, you can delete the branch. Click the **Delete branch** button.
21. **For group projects:** Each group member should share their repository with another group member. To do this, on the repository's page, click **Settings**, and then click **Collaborators** in the left pane. Click **Add people**, and then enter your team member's GitHub email address. Accept the invitation sent to you. Go into the repository that is shared with you and follow the earlier steps to create a branch, edit the file, commit your changes, and create a pull request. However, do not merge the pull request in your group member's repository. When your group member's pull request arrives for your repository, access the pull request and click the **Commits** tab to view the changes. Return to the **Conversation** tab and decide whether to merge the changes or to close the pull request without merging. What were the proposed changes, and what was your decision on how to handle them?
22. Return to the **Code** tab. **Take a screenshot** of your repository showing your file name and updated contents; submit this visual with your responses to this project's questions.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Solutions to Self-Check Questions

Section 3-1: Cloud Migration

1. What physical restriction might make the cloud unreasonable for certain types of workloads?

   Answer: a. Latency

   Explanation: Latency is the delay between the transmission of a signal and its receipt, and this physical reality might make certain cloud-hosted services unreasonably slow. Hidden expenses to be considered in choosing a cloud migration strategy include the power bill to cool your data center or expenses for physical security and insurance coverage. The CSP's documentation addresses the CSP's portion of shared responsibility, and this documentation can assist the cloud customer in their pursuit of their own compliance certifications.

2. Which cloud migration strategy requires the most work to perform?

   Answer: c. Refactor

   Explanation: To **refactor** a system requires the most significant changes, such as recoding portions of an application, re-creating portions of the system from scratch, or rebuilding the entire system for efficient integration in the cloud environment. To retain a system is a low-complexity strategy that means the organization keeps using an application or data as it is, without any changes. A revise approach makes some relatively minor changes to the application or data before moving it to the cloud. The rehost strategy refers to moving the application, server, or data into the cloud as it is without making any significant changes.

3. What kind of expense is an electric bill?

   Answer: b. OpEx

   Explanation: The benefits of **OpEx (operational expenditures)**, such as an electric bill, are experienced in about the same period as when the expenses are paid, such as on a monthly or quarterly basis. Cloud providers can participate in CSA's STAR (Security, Trust, Assurance, and Risk) program to certify their compliance with CSA's definition of cloud security best practices. The benefits of CapEx (capital expenditures), such as the purchase of a physical server, are experienced over a long period (more than one year). All expenses—both obvious and hidden—are collectively called TCO (total cost of ownership or total cost of operations).

Section 3-2: Cloud-Native Design

4. Which resource is delivered through a fan-out architecture?

   Answer: c. API call

   Explanation: An **API call** is a message sent between system components; if the message is sent from one producer to several consumers, this is a fan-out architecture. API calls might be generated in response to events. Events are often performed by microservices. A web service is a standardized method of offering network-based communications between software systems.

5. Which architectural design is responsible for automating responses across a system?

   Answer: b. Event-driven architecture

   Explanation: An **event-driven architecture** allows the communication of events between systems, automating responses to each event. The pattern of sending a single message to several consumers is called a fan-out architecture. In a loosely coupled architecture, each system in the infrastructure operates independently of the other systems. A microservices-based architecture gives system developers a modular approach to designing a system.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 3-3: System Development

6. Which command performs the final merge of a commit into the main branch of code?

Answer: d. `git merge`

Explanation: When you're finished making changes (which might involve multiple commits), you then need to merge the feature branch back into the main branch with the **`git merge`** command. Use the `git add` command to choose which files to add to the staging area. Create a new branch with the command `git branch`. Use the `git commit` command to commit, or save, your edits in the local repository.

7. What deployment type does not require replacing infrastructure?

Answer: c. In-place

Explanation: With an **in-place deployment**, the application on each host is stopped, replaced with the new application or updated version, and then that application is started—the infrastructure itself is not affected. A blue-green deployment shifts the workload to the new, updated environment. A canary deployment is similar but shifts only portions of the workload at a time to the new environment. A rolling deployment does not require a separate environment, but it does replace portions of the underlying infrastructure at a time with the new infrastructure hosting the new application version.

8. What kind of testing would detect unintended changes to an interdependent system?

Answer: b. Regression testing

Explanation: Regression testing confirms that changes to one system haven't negatively impacted other functions within that system or other systems in the network. Usability testing asks users to test the system while developers observe the users and the system. Vulnerability testing searches for application liabilities or gaps in protection of sensitive data. Performance testing challenges a system to determine its responsiveness and stability under various conditions.

Section 3-4: Planning for Problems

9. Adding more RAM to a web server is what kind of capacity adjustment?

Answer: b. Vertical scaling

Explanation: Scaling up, called **vertical scaling**, increases resources such as vCPUs or IOPS allocated to an existing instance or storage container. Horizontal scaling increases the number of running resources, such as VM instances or containers. Scaling in and scaling down both reduce capacity rather than increase capacity.

10. Which kind of backup site provides the fastest RTO?

Answer: a. Hot site

Explanation: A **hot site** contains identical resources to the original site and can take over functionality nearly instantaneously. Restoring functionality at a cold site could take a long time, even weeks. Recovery using a warm site can take hours or days. A pilot light tier in the cloud allows you to create backup resources but leave them in a stopped state, which is cheaper than leaving them in a running state; pilot light recovery generally takes minutes or seconds.

11. What kind of DR test involves the most people?

Answer: c. Disaster simulation testing

Explanation: Disaster simulation testing requires the presence and participation of all personnel, including business partners and vendors. A structured walk-through requires each department to walk through the plan in detail, but does not need to involve every member of the department. A tabletop exercise involves a team of representatives from each department. A playbook provides succinct, organized information that is easy to access and follow in stressful situations.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Chapter 4

Cloud Networking

Chapter 4 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Explain how common networking concepts apply to cloud networking.
2. Manage IP address spaces in the cloud.
3. Configure basic AWS networking resources.
4. Configure basic Azure networking resources.
5. Configure basic GCP networking resources.

Chapter 4 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 4-1: Networking Concepts in the Coud
- Networking Concepts
- From OSI Model to Cloud Stack
- Software-Defined Networking in the Cloud

Section 4-2: IP Address Spaces
- IP Addressing
- Subnetting
- Cloud Network Interfaces

Section 4-3: Networking in AWS
- Regions in AWS
- Availability Zones in AWS
- VPCs in AWS
- Subnets in AWS
- Gateways and Route Tables in AWS

Section 4-4: Networking in Azure
- Regions and Availability Zones in Azure
- VNets and Subnets in Azure
- Route Tables in Azure

Section 4-5: Networking in GCP
- Regions and Zones in GCP
- VPCs and Subnets in GCP
- Routes in GCP

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 4 Scenario "What won't work?" Kendra asks as you realize she's standing beside your desk.

"Huh?" You drag your thoughts away from the gadget you were fiddling with and attempt to focus on what Kendra is saying.

"You just muttered something about, 'This is not going to work!' Is everything okay?"

"Oh, I must have been talking to myself. Sorry about that." You turn your attention back to the access point device on your desk. "I was trying to configure this access point so any preapproved computer or smartphone that connects to it will automatically authenticate to the network *and* recognize this network printer. But something's not working, and the printer keeps showing up as an unknown device."

"Ah, networking," she replies. "My favorite topic. Need some help?"

"No, I'll mess with it again after I take a break. I'm not even sure what I want it to do is possible." You slide the access point to a side table and turn to Kendra. "I do have a question, though. I've been thinking…"

"That's a dangerous activity!" Kendra says. "What's on your mind?"

"Networking. Except in the cloud." As Kendra settles into a chair, you continue to muse aloud. "How do devices in the cloud talk to each other? I mean, here in our own data center, we have routers and switches and cables and access points. But in the cloud, we won't be able to get to any of those devices. Do we remote into the cloud provider's devices so we can configure them? How does that work?"

At that moment, Malik finishes a phone call, hears the word "cloud," and wheels his chair to your makeshift meeting space to join the conversation. "This I've gotta hear," he says as he reaches to his desk to grab his cup of coffee.

Kendra props one foot on a nearby box and starts explaining. "No, we can't access the cloud provider's equipment directly. Remember how we talked about VMs that run on top of a virtualization layer? Those VMs are configured separately from the host hardware underneath the hypervisor. It's the same idea with the cloud. The network functions are all managed at a software layer through a cloud console of some kind. Like … if you want a new subnet, you go into the console and create one. Tell it what IP address range to use, maybe configure some routes in a route table, and then spin up VMs inside your virtual network. All of the networking configuration is done at the virtualization level."

"That sounds easy," says Malik. "But then when I think about it, it also sounds really complicated. How does it all work?"

"Let's do some more research," suggests Kendra. She grabs a sticky notepad and pen from your desk and scribbles the following questions:

- Where do networking functions fit in the big picture of cloud computing?
- How are cloud networks created and configured?
- What IP addresses do cloud resources use?
- What kinds of default connections are included in a cloud network?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 4-1: Networking Concepts in the Cloud

Certification

1.3 Explain cloud networking concepts.

1.7 Compare and contrast virtualization concepts.

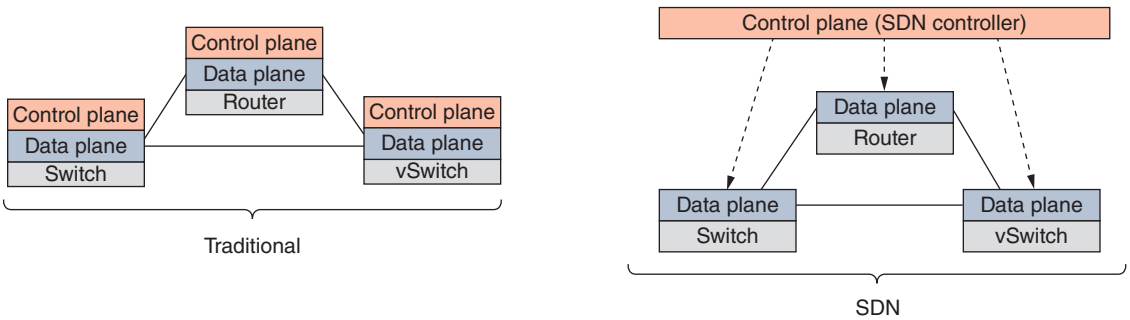
2.5 Given a set of requirements, provision the appropriate cloud resources.

Networking in the cloud works similarly to networking in a physical data center, except at more of a logical level rather than a physical one. In the cloud, you do not need to rely directly on the configurations of physical hardware. Due to the virtualization layer provided by the cloud platform, you have freedom to build the components you need in the ways you need them without many of the restrictions of working with physical infrastructure. To a great degree, if you can imagine it, you can build it. This chapter discusses how networking concepts work in the cloud, and it covers networking configurations in AWS, Azure, and GCP.

Networking Concepts

The IT industry has been shifting toward more virtualized, abstracted networking for some time. For example, **SDN (software-defined networking)** is a centralized approach to networking that removes most of the decision-making power from network devices and instead handles that responsibility at a software level. SDN technologies separate the control plane, or management layer, of the network from the data plane where transmissions on the network actually traverse physical devices. This separation allows for centralized control of network devices from an SDN controller. Figure 4-1 illustrates how this works at a conceptual level.

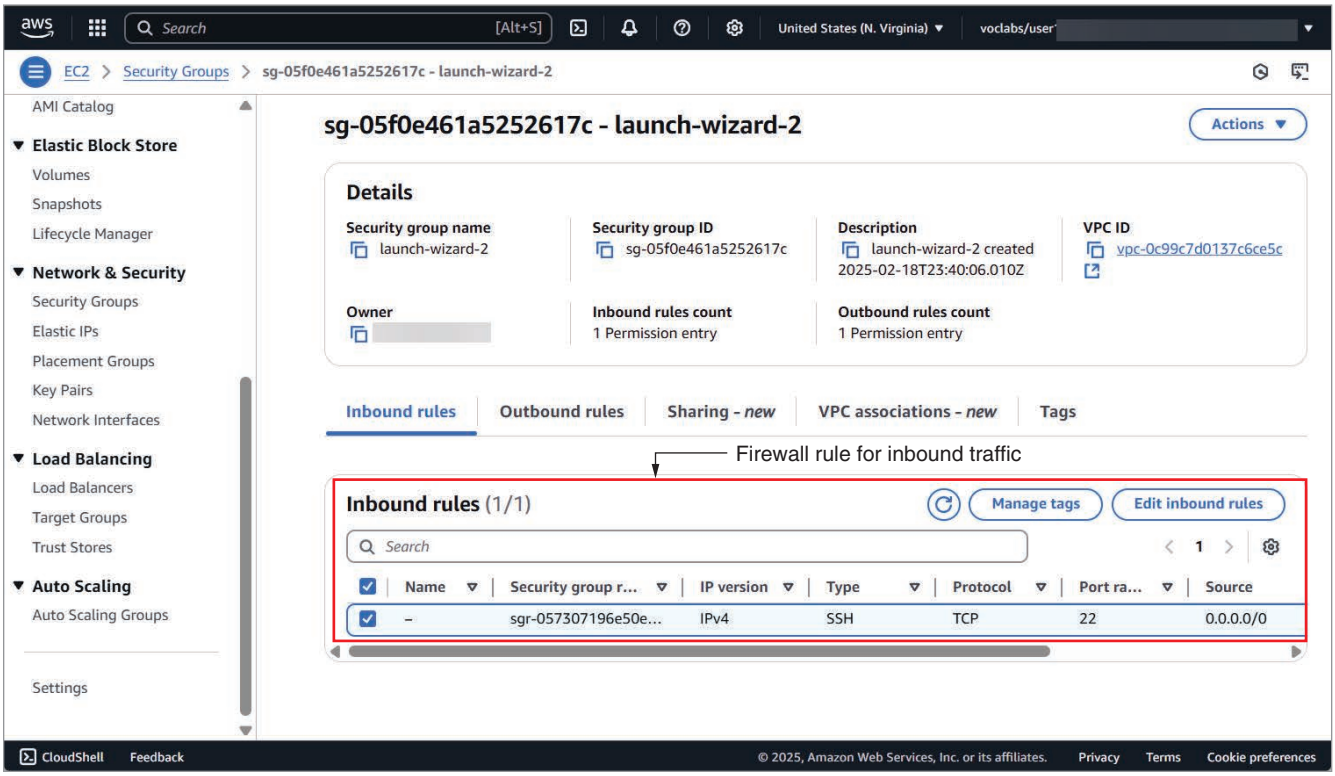
Figure 4-1 Individual control planes in a traditional network versus a centralized control plane in an SDN network



NFV (network functions virtualization) takes this idea a step further by virtualizing network devices and services, much like VMs are virtualized computers. Most of the technologies are the same but abstracted to the virtual platform. For example, cloud networks still use firewall technologies. Similar security features apply firewall rules, such as an AWS security group (Figure 4-2) that can allow inbound SSH traffic to manage a Linux instance. However, these firewalls are virtual appliances or logical configurations running on a virtual, software-defined network. Messages are directed through invisible hardware according to rules enforced by tags added to message headers.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-2 Inbound rule in an AWS security group



Source: Amazon Web Services, Inc.

When your data lives in the cloud, accessing your data and securing your data must be discussed simultaneously. As a result, this course's chapters on cloud networking might at times appear more like chapters on cloud security. Nearly every networking decision you make for your cloud will need to consider the security implications of each option available to you.

Note 4-1 | What you already know about physical networking will serve you well when working with cloud-based networks. You'll recognize familiar network components such as routers, switches, firewalls, load balancers, authentication servers, and DNS (Domain Name System or Domain Name Service) and DHCP (Dynamic Host Configuration Protocol) servers.

This course focuses on how these components work when used in the cloud, and it won't cover that content as if it's the first time you've encountered it. If you've not yet completed a networking course, you might need to do some additional research on some networking concepts to help you understand the big picture of what's being discussed.

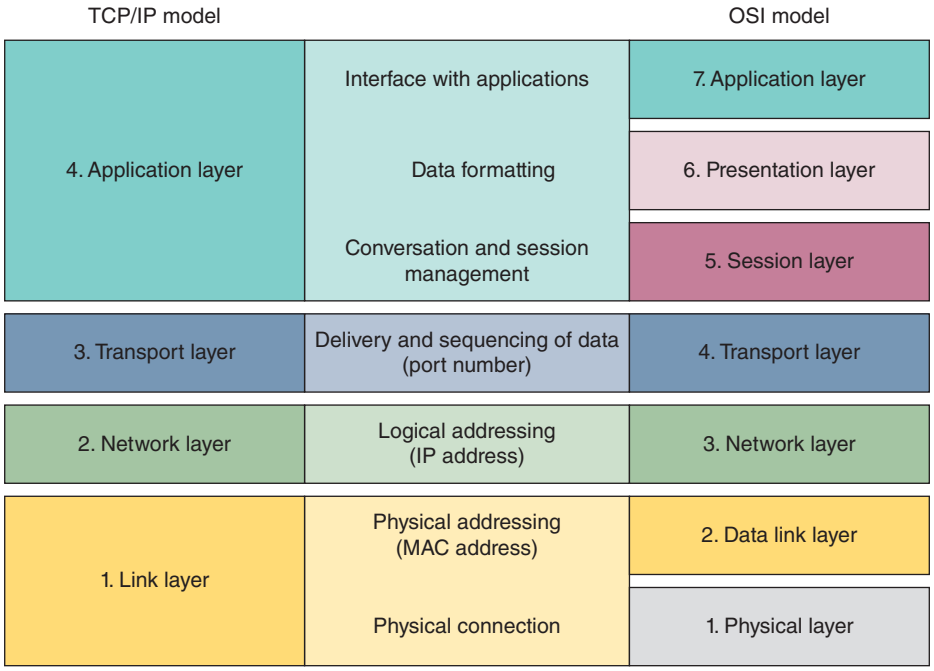
Some ways to find this information include searching online for helpful articles and videos, which is often the way IT professionals keep their own knowledge up to date after completing their formal education. You might also buy a networking textbook—even a used one can be a handy resource. If you have a Cengage Unlimited subscription, you can bookmark the *CompTIA Network+ Guide to Networks*, 10th edition, for quick reference throughout the next few chapters.

From OSI Model to Cloud Stack

If you've spent time in networking, you're familiar with the OSI model and its cousin the TCP/IP model. Figure 4-3 illustrates the OSI and TCP/IP layers as a brief refresher. Both models present theoretical representations of what happens during network communication.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-3 The OSI and TCP/IP models of networking communications

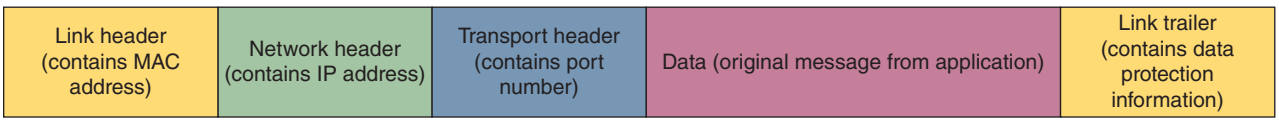


Note 4-2 | The names of the TCP/IP layers are not as tightly standardized in the industry as are the names of the OSI layers. Many organizations and other sources take the liberty of altering the names of the TCP/IP layers to make them easier to understand or more informative in a certain context. For example, the lowest TCP/IP layer might be called "network," "network access," or "host to network." The second layer is often called "internet" even though it does not necessarily refer to the public Internet. The third layer might be called "host-to-host," and the top layer might be called "process application." To minimize confusion between the two models and to simplify references to each layer, this text uses the layer names as listed in Figure 4-3.

Traditional Network Model

These traditional network models are helpful in troubleshooting network problems and in understanding how networking protocols relate to one another because different protocols function at different layers. At each layer, information is added to the message that various devices along the way will need. For example, the port number at the transport layer is added at the front of the message's payload in a section called a header. At the network layer, another header is added to the front of the first header; it contains IP addressing information for the source device and the destination device. At the link layer, yet another header is added to provide the MAC address for the next device on the path that will receive the message. This process of adding headers to carry various layers of information is called encapsulation. In Figure 4-4, three headers containing addressing information are added to a message. The transport layer first adds a header to the message, followed by the network layer, and finally the link layer's header and trailer on the outermost edges.

Figure 4-4 Headers on a message payload



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To get the message from the source to the destination, the message must traverse several other devices. Initially, the message crosses the local network's router to reach the Internet. The message then encounters many routers and other devices along the way. When it reaches its destination network, the message crosses that network's router and probably a firewall, followed by a switch and possibly several other devices. Each device reads only deeply enough into the message's headers to know where to send the message next (or, in the case of a firewall, to check the message content for security issues).

Cloud Stack Model

Cloud computing adds a new level of networking, called an overlay network, on top of this underlying physical infrastructure, called the underlay network. An **overlay network** is a virtual network that uses its own logical configurations to direct traffic on the underlying physical infrastructure. For example, a VLAN (virtual LAN) is a relatively simple overlay network that directs traffic according to its own rules. SDN is another type of overlay network. Cloud networking also relies on overlay networking so customers can configure their own networking rules in their cloud environments without needing to configure underlying physical network devices.

Cloud networking brings some creative modifications to the traditional networking models. Still, the overall structure and logic remain somewhat the same in that each layer builds on another to create a stack of functions. In the cloud stack, hardware remains at the bottom of the stack and applications reside at the top. Let's explore how the cloud stack compares to traditional network layers.

Various versions of the cloud stack model list five to seven layers performing specific functions that are different from those described in traditional network models for physical data centers. With the OSI model, layers are defined by where information is found within a message's headers. But cloud layers are better defined by who has access to the configurations of that layer. While the layers are not standardized, in general, these layers can be described as listed in Table 4-1.

Table 4-1 Cloud stack theoretical model

Layer	Name	Description	Responsibility and access
1	Physical layer	Contains the physical infrastructure, including cables, servers, networking devices, cooling systems, and power systems.	In a public cloud, the physical layer is the CSP's responsibility and is not accessible by the cloud consumer, who typically does not know where these components are geographically located.
2	Virtualization layer	Abstracts virtual hardware from the physical hardware through virtualization software. This might be a hypervisor, such as VMware's ESXi, or a cloud platform, such as OpenStack, or a combination of these tools.	In a public cloud, the virtualization layer is the CSP's responsibility.
3	Network or management layer	Provides APIs that allow IaaS customers to create and manage virtual networks. This layer is where server resources, access control, many costs, cybersecurity, and other services are managed.	From a public cloud customer's perspective, options here include the major CSPs, such as AWS, Azure, and GCP. Depending on the service used, this layer is controlled either by the CSP or the customer.

(continued)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 4-1 Cloud stack theoretical model *(continued)*

Layer	Name	Description	Responsibility and access
4	Image or OS layer	Provides instances or platform-native services that support features at the application layer. Workloads, data storage, and data processing happen at this layer.	PaaS customers can manipulate various OS environments for development, testing, and hosting. The customer can set some infrastructure preferences (such as connection bandwidth or dedicated processing power), but the CSP controls most aspects of the underlying infrastructure. Alternatively, cloud customers can run their own underlying virtual infrastructure (such as running a hypervisor on VM instances), and then top that layer with a self-managed image or OS layer.
5	Application layer	Hosts many traditionally lower-layer functions such as virtual load balancing and firewalls. Also presents applications to SaaS users.	Cloud customers can set some configurations, such as access control or app settings, but the CSP manages the entire underlying infrastructure. Alternatively, cloud customers can build the application layer over their own virtual infrastructure, which runs on cloud IaaS services.

It's helpful to think about cloud computing in these layers. As with the OSI model in networking, these cloud stack layers can help you organize information in your mind as you're learning it and help with troubleshooting later, when something goes wrong. This chapter—and the next few chapters—focus primarily on the network layer where IaaS configurations are applied to achieve networking and security services.

Software-Defined Networking in the Cloud

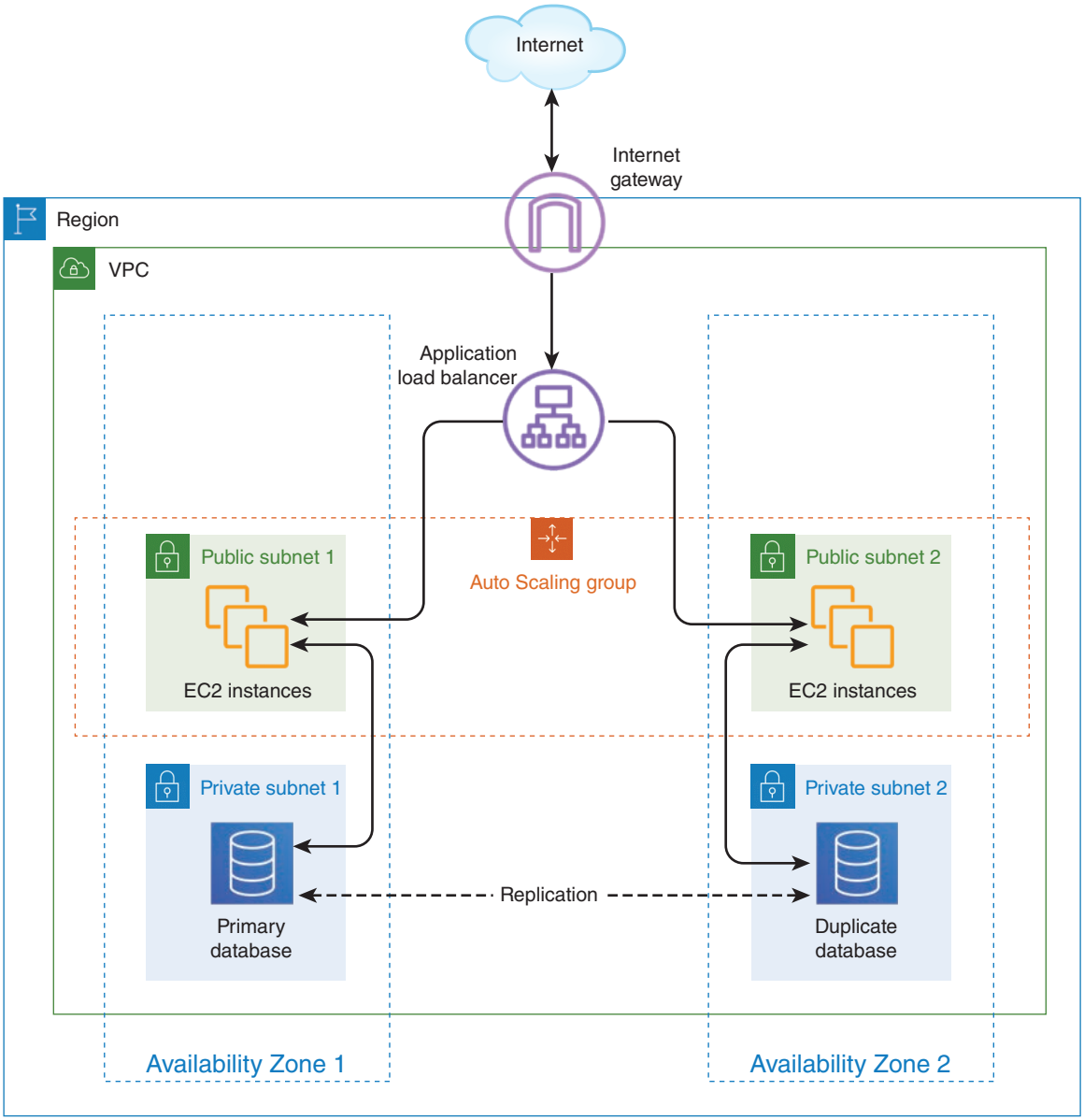
As you've learned, SDN separates the control plane where decision making occurs from the data plane where data bits are transferred from device to device. This is the basis on which the public cloud works: The underlying physical infrastructure that supports the cloud is not accessible to the cloud customer. In the public cloud, you as the customer are not responsible for ensuring there are enough physical devices to guarantee performance and availability—the cloud provider is. In fact, you don't configure physical routers and switches at all. Instead, these functions (routing and subnetting) are configured through the cloud platform's SDN-backed interface. For example, you set static routes in a **route table**, which is a list of routes used to determine where traffic is routed. You also assign IP address ranges to your subnets.

The cloud customer configures networking infrastructure, such as routing and switching, at a software-defined level using tools and services built into the cloud platform. For example, when you add routes to a route table, your routes are automatically propagated to many physical routers so traffic is properly handled. Additionally, the cloud provider must ensure there are a sufficient number of physical devices supporting its cloud platform to ensure HA (high availability) of these virtual services.

What steps should you take, then, to ensure availability for your routes and subnetting in the cloud? Just like with on-prem networks, you need to avoid any SPoF (single point of failure). For example, Figure 4-5 illustrates a deployment in AWS with duplicate web servers and a load balancer to distribute traffic between both servers. Notice you do not need to duplicate the load balancer because it's inherently redundant due to AWS's design for its load balancing service.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-5 Redundant EC2 instances and databases



You might configure the servers to auto-scale in response to increasing demand or to automatically compensate for a failed server instance. You might also configure database replicas in multiple regions. For each of these redundancies, you'll need to make sure IP address configurations will carry over to replacement resources as they're needed. In that case, duplicating IP address configurations across redundant servers or databases is a form of HA for switching. Similarly, if you duplicate a subnet and copy route tables to redundant subnets, this is a form of HA for routing.

AI in IT | With the increasing complexity of network configurations, especially in the cloud, cloud customers and CSPs are adopting AI-powered technologies to increase automation, optimize performance, and ensure network security. AI can analyze vast amounts of network data, automatically adjust configurations of network devices and services, and identify indications of performance issues followed

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

by automated responses. AI systems can also be given access to anonymized network data from many sources to better inform ML-driven decision making. Much of this capability is being built into existing or new networking tools. For example, Cisco offers many AI-native networking solutions, such as Cisco Meraki to manage network deployments and Cisco Catalyst Center to secure and automate network operations.

Self-Check

1. Which theoretical plane moves traffic from switch port to switch port?
   - **a.** SDN plane
   - **b.** Data plane
   - **c.** Control plane
   - **d.** Management plane
2. At which cloud layer are user access accounts managed?
   - **a.** Virtualization layer
   - **b.** Image layer
   - **c.** Physical layer
   - **d.** Network layer
3. Who configures physical devices hosting the cloud?
   - **a.** The IaaS user
   - **b.** The SaaS user
   - **c.** The CSP
   - **d.** The PaaS user

Check your answers at the end of this chapter.

Section 4-2: IP Address Spaces

Certification

1.3 Explain cloud networking concepts.

2.3 Summarize aspects of cloud migration.

6.2 Given a scenario, troubleshoot network issues.

At the network layer of the TCP/IP or OSI model, IP addresses are used to identify devices across networks. This system is also used in the cloud to some degree. Many cloud resources can be given one or both of two kinds of IP addresses: public and private. **Public IP addresses** can provide access to a resource from the Internet. **Private IP addresses** are only accessible from within the cloud-based network where that resource is located.

Most of the time, private IP addresses are automatically assigned to a resource if it needs one to communicate. Some cloud platforms—such as AWS—also assign a public IP address to many cloud resources by default. However, these public IP addresses often are dynamic, meaning the address could change at any time without warning, especially when the resource is power cycled (shut down and restarted). Assigning a static public IP address to a resource typically incurs charges.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As you might guess, there is much more to the process of managing network connections in the cloud than choosing between public and private IP addresses. This section explains IP addressing, subnetting, segmentation, and interface management in the cloud.

IP Addressing

In an on-prem network, the process of assigning private IP addresses is normally managed by a DHCP server. In the cloud, IP address assignment is more automatic. Private IP addresses are assigned chronologically from specific IP address ranges that are defined when you create the virtual network and its subnets without requiring a customer-managed DHCP server. The IETF (Internet Engineering Task Force) published **RFC (Request for Comments) 1918** in February 1996 to specify best practices for address allocation on private networks. RFC 1918, which you can read at https://datatracker.ietf.org/doc/html/rfc1918, identifies three blocks of IP address space reserved by IANA (Internet Assigned Numbers Authority) for use on private networks (Table 4-2). This includes private networks hosted in the cloud.

Table 4-2 RFC 1918 address blocks

Start address	End address	Prefix	Subnet mask
10.0.0.0	10.255.255.255	10.x.x.x/8	255.0.0.0
172.16.0.0	172.31.255.255	172.16.x.x/12	255.240.0.0
192.168.0.0	192.168.255.255	192.168.x.x/16	255.255.0.0

Note 4-3 | The addresses in Table 4-2 are written using IPv4 (IP version 4). IPv4 has been around for a long time. When it was first defined, no one realized how many devices would eventually be connected to the Internet. The numbering system used by IPv4 is not expansive enough to support the demand for IP addresses today. A newer version, IPv6, was developed to solve this problem. Although IPv6 has been around for several years, it has not yet been widely adopted due to the lengthy addresses it uses, the updates required to other network protocols to support it, and the way it changes the rules used to direct traffic across a network. Still, IPv6 offers increased security and greater efficiency for network traffic. Here's an example of an IPv6 link-local address (that is, an address used on a local network, not across the Internet). Notice how it's formatted differently from an IPv4 address. This particular IPv6 address begins with FE80, which indicates it only works on a local network:

```
FE80::64D2:BD2E:FA62:B911
```

IPv6 is supported to some degree by many CSPs, but its use is still considered optional in most cases.

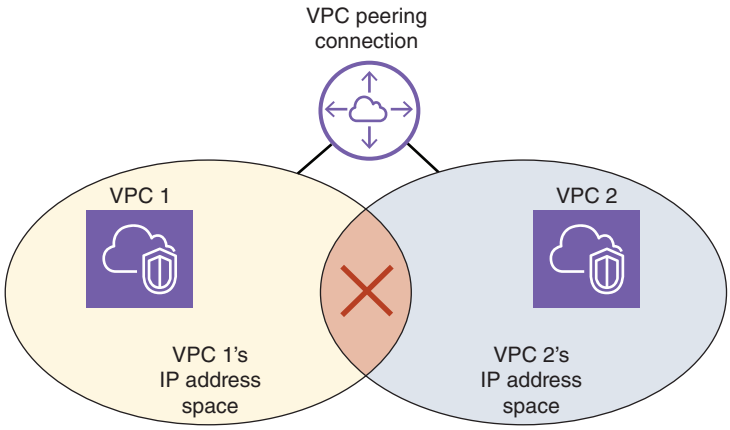
When planning a network space, whether that's on-prem or in the cloud, you need to consider the IP addressing used for that space. It's important to think about not just the network's current needs but also its future needs. Considering that the cloud offers nearly unlimited opportunities for expansion, future planning is especially important. As your company grows, **scope exhaustion**—that is, running out of available IP addresses in a defined CIDR (classless interdomain routing) block—can be just as much a problem in the cloud as it is on-prem. Make sure to define CIDR blocks that can accommodate future growth within each subnet. You also need to consider which CIDR blocks might need to be able to communicate directly with each other, which requires avoiding any overlap in their IP address spaces.

In the cloud, you choose IP address spaces for subnets and possibly for virtual networks (depending on the cloud platform used). Although there is a natural sense of limitation in not wanting to position too many resources within a single subnet, there are times when you might want to connect subnets or networks to each other, such as with a configuration called **peering** (Figure 4-6). This is another form of growth in the cloud

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

that your IP address spaces should be designed to anticipate. When you peer virtual networks, you have to ensure that each of the peered networks uses a different IP address space than all the others it is peered with. The potential for network overlapping particularly becomes a problem if you've used the same or similar default CIDR block for every virtual network in your cloud deployment—peered networks are not allowed to overlap IP address spaces.

Figure 4-6 Peered VPCs cannot have overlapping CIDR blocks



Thoughtfully planning IP address spaces across your organization, even if some resources reside in very different portions of your overall cloud deployment, can make future expansions and reorganizations much easier to implement.

Subnetting

IPv4 address blocks, such as 192.168.0.0/16, are written using **CIDR notation**, which is a shorthand method for identifying network and host bits in an IP address. This way, subnet masks, such as 255.255.255.0, don't have to be written out. The number after the slash, such as /16 or /24, indicates how many bits in the IP address indicate the network portion, which is the network ID. For example, 192.168.4.1 is written in binary as

`11000000.10101000.00000100.00000001`

Notice that each set of bits between the decimals contains 8 bits and is called an octet. If this IP address is part of a /24 CIDR block of addresses, then the first 24 bits are used to identify the network, and here are written in bold:

`11000000.10101000.00000100.``00000001`

Every host on this network will have the same first 24 bits and share the same first three octets: 192.168.4. The last octet will be different for each host. Only the last eight bits can be used to identify different hosts, and so are called host bits or the host ID. However, the number of bits committed to identifying the network can be adjusted. For example, the IP address 192.168.4.1/25 has 25 network bits:

`11000000.10101000.00000100.0``0000001`

Because more of the bits are used to identify the network, fewer of the bits can be used to identify the hosts. This means that you have more networks and fewer hosts on each of these networks. The process of calculating address spaces in this manner is called **subnetting**.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Subnetting helps manage IP address spaces in the cloud. It's important that you're comfortable with choosing subnets and identifying subnet-related problems. Most of the time, you'll find that cloud-hosted IPv4 subnets are listed in CIDR blocks, such as 192.168.0.0/24. For ease of use, most configurations will default to classful subnets, where the network bits take up complete octets with no splits in the middle of an octet:

- Class A subnets use the first octet for the network ID, such as **10**.0.0.0/8.
- Class B subnets use the first and second octets for the network ID, such as **172.16**.0.0/16.
- Class C subnets use the first, second, and third octets for the network ID, such as **192.168.100**.0/24.

Note 4-4 | If you'd like a more detailed explanation of how to calculate subnets, refer to this book's Appendix E, "Calculating Subnets."

Practice Subnetting

In adjusting subnet settings, recall that increasing the number after the slash (/) effectively increases the number of subnets you have—and decreases the number of hosts on each subnet. This is accomplished by increasing the number of IP address bits used for the network ID and decreasing the number of bits used for the host ID. Consider the following example where you calculate the number of bits to borrow for a network's subnets.

Suppose you're given a network ID of 192.168.89.0/24 for an entire network, and you want to create eight smaller subnets. The 192.168.89 portion of the network ID represents the 24 bits (**11000000 10101000 01011001**) dedicated to the network portion of the IP addresses used in this network. The .0 portion at the end of the CIDR block represents the eight bits (00000000) that will identify individual hosts within the network. To create eight smaller subnets, you need to use some of the bits from the host portion to instead identify these different subnets. How many bits should become part of the network ID?

Considering that each bit represents two possible numbers—1 or 0—think through your options:

- 192.168.89.0/25 would adopt one of the host bits for the network portion (**11000000 10101000 01011001 ?**0000000), giving you two subnets:
  - 192.168.89.0 (**11000000 10101000 01011001 0**0000000)
  - 192.168.89.128 (**11000000 10101000 01011001 1**0000000)

The scenario requires eight subnets, so this is not enough. Let's try again:

- 192.168.89.0/26 adopts an additional host bit for the network portion, yielding four subnets ($2^2 = 4$) with the following possible combinations of bit values:
  - 192.168.89.0 (**11000000 10101000 01011001 00**000000)
  - 192.168.89.64 (**11000000 10101000 01011001 01**0000000)
  - 192.168.89.128 (**11000000 10101000 01011001 10**0000000)
  - 192.168.89.192 (**11000000 10101000 01011001 11**0000000)

That's still not enough, so take one more bit from the host portion:

- 192.168.89.0/27 adopts three host bits for the network portion, yielding eight possible subnets ($2^3 = 8$) with the following possible values:
  - 192.168.89.0 (**11000000 10101000 01011001 000**00000)
  - 192.168.89.32 (**11000000 10101000 01011001 001**00000)
  - 192.168.89.64 (**11000000 10101000 01011001 010**00000)
  - 192.168.89.96 (**11000000 10101000 01011001 011**00000)
  - 192.168.89.128 (**11000000 10101000 01011001 100**00000)
  - 192.168.89.160 (**11000000 10101000 01011001 101**00000)
  - 192.168.89.192 (**11000000 10101000 01011001 110**00000)
  - 192.168.89.224 (**11000000 10101000 01011001 111**00000)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The scenario requires eight subnets. Borrowing three bits has accomplished this goal. The range for each subnet begins at the subnet address and ends at the address immediately preceding the next subnet address. For example, the range of IP addresses for the first subnet in this example is 192.168.89.0 through 192.168.89.31.

Within each subnet's IP address range are reserved addresses that cannot be used for hosts in the subnet. In the typical on-prem data center, these reserved addresses include the first address (192.168.89.0) for the network ID and the last address (192.168.89.31) for broadcast traffic. Other reserved addresses might include the router, a printer, or some other network service. Cloud-based subnets also require reserved IP addresses. For example, AWS reserves five addresses in each subnet, as listed in Table 4-3.

Table 4-3 Reserved IP addresses on each AWS subnet

Location within range	Example	Purpose
First address	192.168.89.0	Network address
Second address	192.168.89.1	Reserved for virtual router
Third address	192.168.89.2	Reserved for DNS server or other service
Fourth address	192.168.89.3	Reserved by AWS for future use
Last address	192.168.89.31	Reserved by AWS even though broadcasting is not supported

Note 4-5 | A quick way to determine how many subnets you create with each borrowed bit is to skip-count by powers of two (such as $2 \times 2 \times 2$, which is 2^3) while holding up an additional finger each time you count up. At the end, count how many fingers you're holding up, and that's how many bits you need to borrow to get the number you counted to.

For example, imagine that your network uses a /24 CIDR block, and you want eight subnets. You already have 24 bits allocated to the network ID. How many bits do you need to borrow from the host ID to get eight subnets? Say "*two*" and hold up one finger. Say "...times *two* is four" and hold up a second finger. Say "...times *two* is eight" and hold up a third finger. You reached your goal of eight (which is 2^3), so stop counting and notice that you're holding up three fingers. This means you need to borrow three bits from the host ID to get eight subnets.

Now try this yourself. Say you have a /20 network and you want 32 subnets. How many bits do you need to borrow from the host ID? Lift another finger each time you say "two": *Two* times *two* is four, times *two* is eight, times *two* is 16, times *two* is 32. How many fingers are you holding up? That's the number of additional bits you need to borrow from the host ID portion.

If you're familiar with subnetting calculations from the CompTIA Network+ exam or from a networking course, you'll remember using formulas or subnet tables to calculate subnets. While these tools are necessary for extensive, detailed calculations, subnetting in the cloud doesn't always have to be quite so precise. In fact, you'll generally want to include significant room for growth. At the same time, however, you'll find situations where your subnets should be carefully sized, identified, and/or mapped across multiple networks.

Consider carefully which address ranges are used for various segments of a network or for different networks that will be connected, as connected address ranges should not overlap. In these cases, like most networking professionals, you can use a subnet calculator to help eliminate potential calculation errors. For your convenience, Table 4-4 details a CIDR chart, which lists a breakdown of network bits and host bits for each CIDR block from /16 through /28. These are the block sizes most commonly available for use in a public cloud. Remember to allow for the number of host addresses the CSP will reserve in each subnet in addition to the number of host addresses needed for your purposes.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 4-4 CIDR chart

CIDR	# of host bits	# of IP addresses	Subnet mask
/16	16	65,536	255.255.0.0
/17	15	32,768	255.255.128.0
/18	14	16,384	255.255.192.0
/19	13	8,192	255.255.224.0
/20	12	4,096	255.255.240.0
/21	11	2,048	255.255.248.0
/22	10	1,024	255.255.252.0
/23	9	512	255.255.254.0
/24	8	256	255.255.255.0
/25	7	128	255.255.255.128
/26	6	64	255.255.255.192
/27	5	32	255.255.255.224
/28	4	16	255.255.255.240

Exam Tip | While you're not likely to be required to calculate detailed subnets on the CompTIA Cloud+ exam, you should still review the steps involved in these calculations. You might be given a scenario that requires you to analyze appropriate subnet sizing or positioning when considering specific factors. For example, you should be able to identify smaller CIDR blocks within a larger CIDR block. You should also know how to identify overlapping CIDR blocks, which can cause problems with IP address spaces. And you should be able to determine whether a specific IP address falls within a given IP address range.

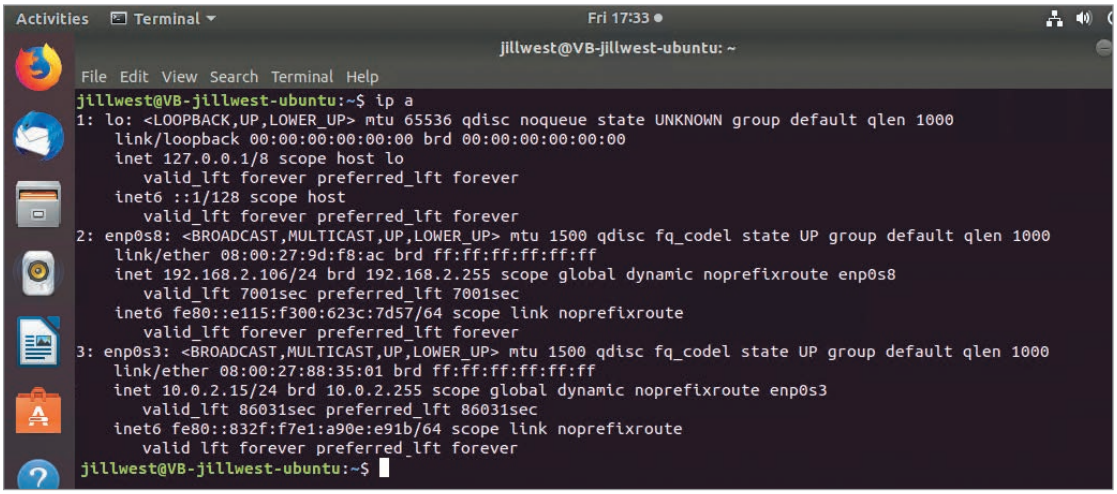
Cloud Network Interfaces

When you plug an Ethernet cable into the RJ-45 port on your computer's motherboard, you probably don't think about that port as being anything other than an integral part of the computer. In networking, however, each interface on a computer or other network device is, in a way, its own entity. Each interface has its own address(es) and can be configured with different rules.

This reality is more obvious on a Linux computer than on a Windows computer. When you run `ipconfig` on a Windows computer, you get a list of all interfaces on the computer and configuration information about each. However, when you run `ifconfig` (or the more modern `ip` command) on a Linux computer, the output identifies each interface by a name. Examples include eth0 for Ethernet 0, wlan0 for wireless LAN 0, or enp0s3 for a network interface on an Ubuntu machine. For example, in Figure 4-7, an Ubuntu VM lists three network interfaces: a loopback interface, a connection to the host machine's Ethernet interface, and a wireless interface to the local network.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-7 Output of the `ip a` command on a Linux VM



Source: Canonical Group Limited

Note 4-6 | The `ifconfig` command has been deprecated, or retired, in some Linux distros. Instead, use the `ip` command, which is part of the iproute2 suite of tools, to display and configure settings on network interfaces in Linux.

Interfaces are also separate entities, or resources, in cloud networking. When you create a VM, you also create a vNIC (virtual network interface card) with its own ID number and configuration settings. This vNIC provides a connection between the VM and its subnet in a virtual network. You can create and delete vNICs separately from VMs, and you can move vNICs between VMs or subnets. Keep in mind that you must power down a VM before attaching or detaching any of its vNICs.

Applying Concepts 4-1: Identify Network Interfaces

Estimated time: 10 minutes

Objective 1.3: Explain cloud networking concepts.

Resource:

- Access to a local computer's CLI

Description: Commands such as `ipconfig`, `ifconfig`, and `ip` can provide valuable information about network interfaces on a computer. To explore how to use these commands on your local computer, complete the following steps:

1. Open a CLI on your local computer, such as Terminal, PowerShell, or Command Prompt.
2. Enter the appropriate command for your OS. For example, on Windows, enter `ipconfig`. On a Mac or Linux system, enter `ip a` or `ifconfig`. Note that one or both these utilities might need to be installed on your system to use it. If necessary, search online for more specific instructions to install the required utility on your system.
3. Respond to the following questions:
   a. What is your local computer's private IP address?
   b. What is the subnet mask for this IP address?
   c. How would this subnet mask be written in slash notation?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In the cloud, an assigned public IP address is treated as a separate resource and sometimes incurs fees. In Azure, IP address *assignments* are also treated as a separate resource called an IP configuration. This IP configuration identifies public and private IPv4 and IPv6 address assignments to each vNIC, whether those assignments are dynamic or static.

Self-Check

4. What kind of address is 8.8.8.8?
   a. RFC 1918 address
   b. Private IP address
   c. IPv6 address
   d. Public IP address

5. If a VM is assigned the address 172.16.18.10/24, what is its network ID?
   a. 172.16
   b. 172.16.18
   c. 192.168.18
   d. 192.168

6. When you delete a vNIC in the cloud, what else is also deleted?
   a. The VM instance
   b. The VM instance's storage drive
   c. The subnet the vNIC is part of
   d. Nothing—the vNIC is a separate entity

Check your answers at the end of this chapter.

Section 4-3: Networking in AWS

Certification

1.2 Explain concepts related to service availability.

1.3 Explain cloud networking concepts.

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.3 Given a scenario, implement identity and access management.

6.1 Given a scenario, troubleshoot deployment issues.

6.2 Given a scenario, troubleshoot network issues.

The cloud has necessitated the development of many new segmentation concepts. Most notable is the concept of the **VPC (virtual private cloud)**. When you create a network in a public cloud, you're reserving your own space within that larger cloud environment to configure as you choose. This is your VPC (or in Azure, a **VNet**, short for virtual network). It's essentially a software-defined portion of a larger, cloud-based network.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

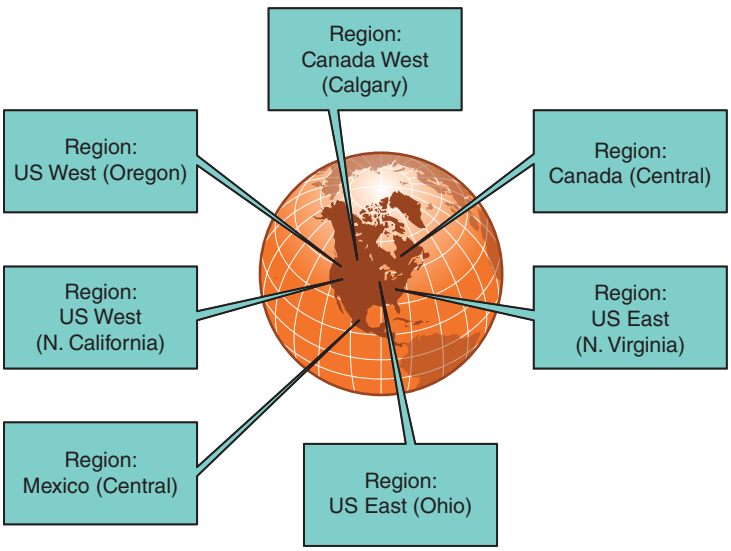
Inside a VPC or VNet, you can host servers, some storage or database resources, and nearly anything else that has a virtual network interface. Each CSP conceptualizes the VPC a little differently, allocating various segmentation and configuration options, but the overall idea is the same. To get a good idea of how this works, you will first learn about VPCs in AWS. Then you'll discover some of the differences with how VPCs function in Azure and GCP.

When you first create an account, AWS creates a default VPC for you. When you create a VM instance in AWS, the instance is automatically configured to work within your default VPC unless you change this setting. When adding new resources, you can create a new VPC, or you place resources in a nondefault VPC that you've already created. Within each VPC, you can configure subnets, route tables, network gateways, and network connections, as well as many other options. To understand what all this means, this section discusses the networking components used to segment and connect cloud resources in AWS, beginning with the largest segment type: regions.

Regions in AWS

AWS offers services in more than 30 geographical regions around the world, with many more slated for release in the next few years. Each region is named for the geographic location of a group of AWS's physical data centers in that area and has an identity code, such as us-east-2, which is US East (Ohio), or eu-west-3, which is Europe (Paris). Figure 4-8 points out the approximate locations of AWS's public North American regions. With few exceptions, cloud consumers typically can create resources in any region. You might choose one region or another, for example, in order to place web servers near customers or to comply with data residency restrictions.

Figure 4-8 AWS public regions in North America



When you create an AWS account, you're assigned a default region. You can later choose a different region near you in which to create your AWS resources. You might instead choose a region farther away from you according to the following factors:

- **Cost estimates of the resources you intend to use**—Costs can vary considerably among regions.
- **Inherent latencies for a specific region**—Some regions perform better than others.
- **Relevant latencies**—Placing cloud resources closer to their users reduces the time needed to communicate with those resources.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

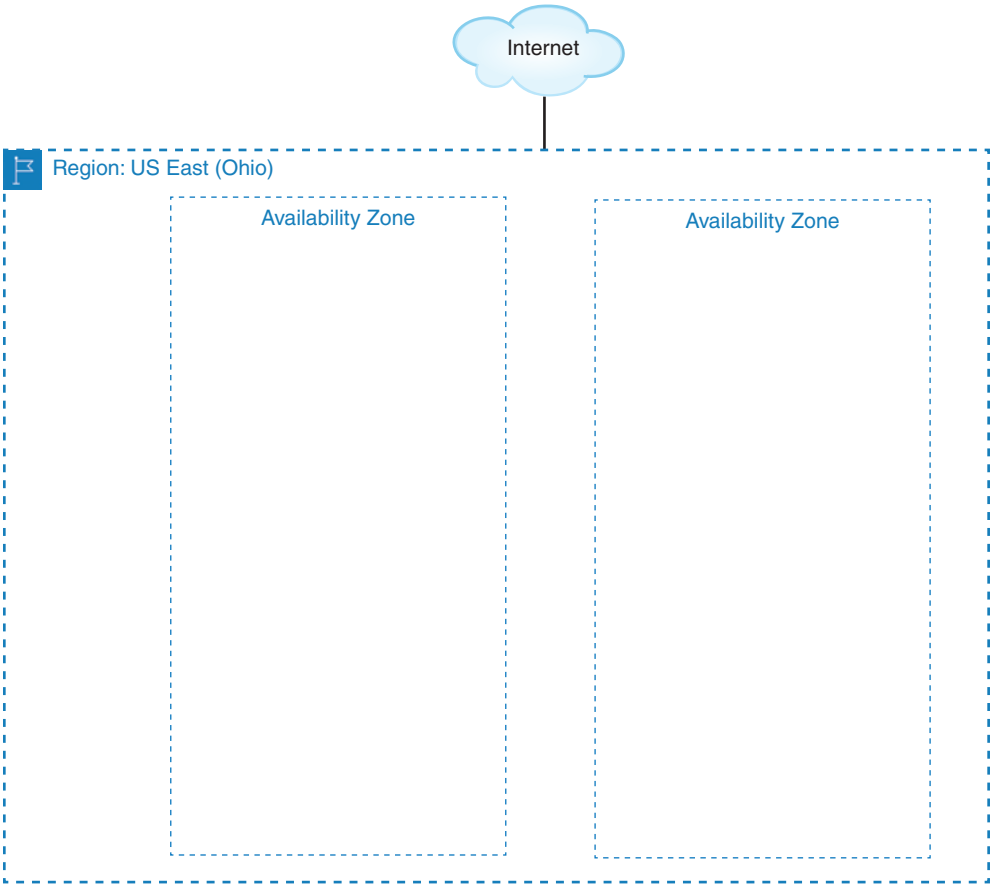
- **Compliance requirements**—Some data is required to be stored within certain geopolitical boundaries.
- **Available features or services in a region**—Not all services are available to the same degree in all regions.
- **Resource availability**—Duplicating resources across regions can increase availability in case a widespread outage affects an entire region.

Cost estimates are also increased when needing to transfer data between regions. Reducing the need for region-to-region communications can help keep costs within budget.

Availability Zones in AWS

Each region contains two or more AZs (availability zones), as illustrated in Figure 4-9 (this diagram will be developed further as you read through this section, adding more layers of information). AZs address HA requirements by providing multiple isolated zones within each region. For the most part, an event that causes a partial or even a full outage in one AZ should not affect services or resources in a different AZ, even if both AZs are in the same region. While some communication between regions crosses the public Internet, communication between AZs in the same region never leaves the AWS infrastructure.

Figure 4-9 Redundancy with multiple availability zones

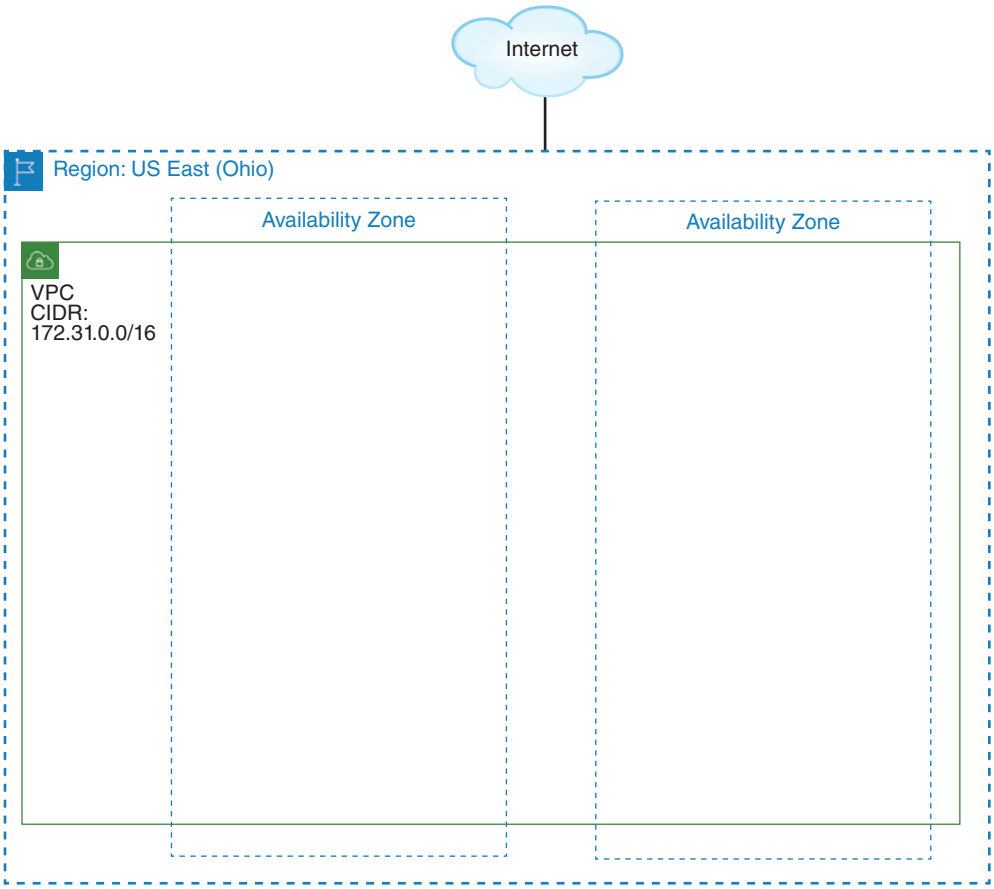


Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

VPCs in AWS

When you create an AWS account, a default VPC is created for you in each region that is available to you. Each default VPC is automatically assigned a /16 IPv4 CIDR range, though you can assign smaller spaces to your custom VPCs (as small as /28). Every VPC spans all AZs in its region (Figure 4-10). Also, DNS and DHCP options are initially set at the VPC level. Many—but not all—AWS resources must be or may be associated with a VPC.

Figure 4-10 An AWS VPC spans all AZs in a region



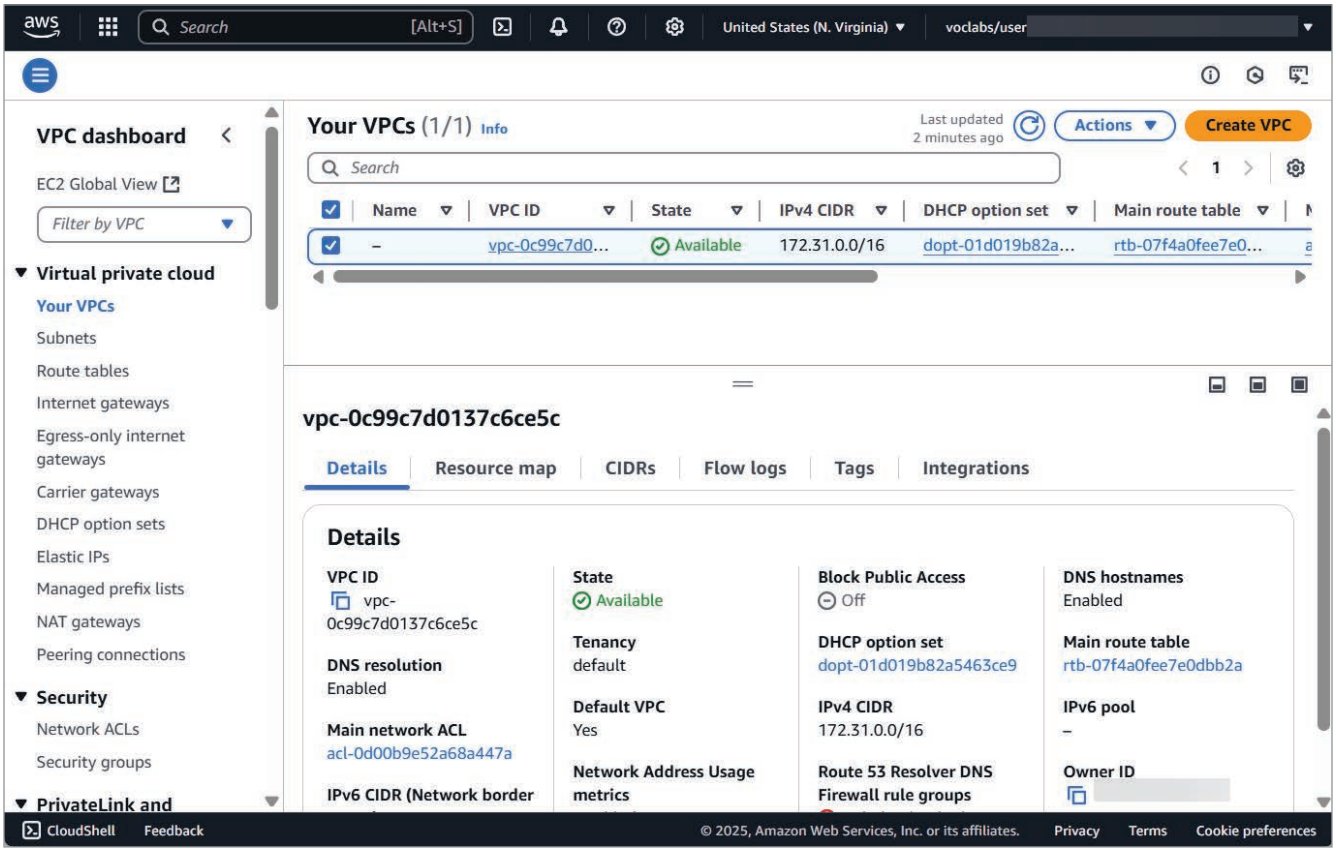
Note 4-7 | The default IPv4 CIDR block range for a default VPC is /16. However, you can instead use any CIDR block size from /16 (that is, 65,536 IP addresses) to /28 (that is, 16 IP addresses) when setting this configuration manually. The same restriction on CIDR block size applies to a VPC's subnets. However, a subnet's CIDR block must be contained within its VPC's CIDR block.

Additionally, some CIDR block ranges are not permitted to be used because AWS services use these ranges. For example, the CIDR block 172.17.0.0/16 is restricted.

In Figure 4-11, notice the default VPC's /16 CIDR range and the options associated with it, such as DHCP, DNS, a route table, and the network ACL. (You learn more about route tables and network ACLs in later chapters.) All these settings will, by default, carry over to any VM instances positioned within the VPC. You can change many of these settings for the default VPC (Figure 4-12), create nondefault VPCs with different settings, or replace the default VPC with a new one.

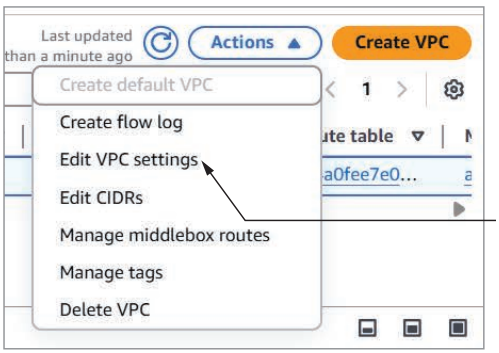
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-11 VPC configurations affect subnets and VM instances



Source: Amazon Web Services, Inc.

Figure 4-12 Change a VPC's CIDR range, DHCP options, or DNS settings



Source: Amazon Web Services, Inc.

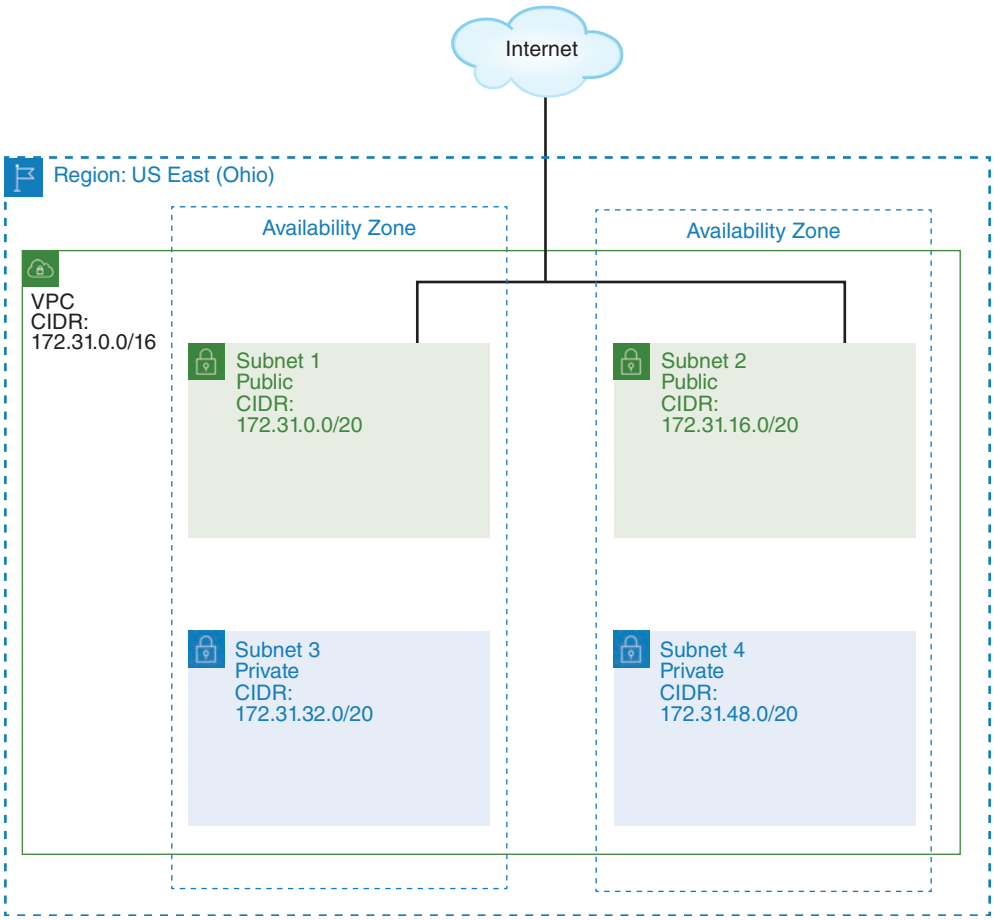
Note 4-8 | You can also enable IPv6 within a VPC, but you cannot disable IPv4. The configurations for IPv4 and IPv6 operate independently of each other.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Subnets in AWS

By default, each default VPC in each region contains one /20 subnet in each AZ for that region. That's 4,096 IPv4 addresses per subnet. Five of those addresses are reserved for use by AWS, leaving 4,091 available host addresses in each default subnet. In Figure 4-13, the default subnets are the green-shaded subnets labeled "Public." Default subnets have access to the Internet, which is why they are called **public subnets**. You'll find out in a moment how this is accomplished. **Private subnets**, such as the blue-shaded ones at the bottom of Figure 4-13, do not have direct access to the Internet. You can create multiple subnets within each AZ, and you can choose whether these subnets are public or private. Subnets cannot span AZs.

Figure 4-13 Public subnets in AWS have access to the Internet, whereas private subnets do not



Note 4-9 | CIDR block ranges for VPCs and subnets conform to the RFC 1918 specifications for private IP addressing. While you *can* use publicly routable IP addresses for your CIDR block ranges, it's not recommended and will not, by itself, result in Internet access for instances within those VPCs or subnets. Internet access is only provided through an Internet gateway, which is discussed next.

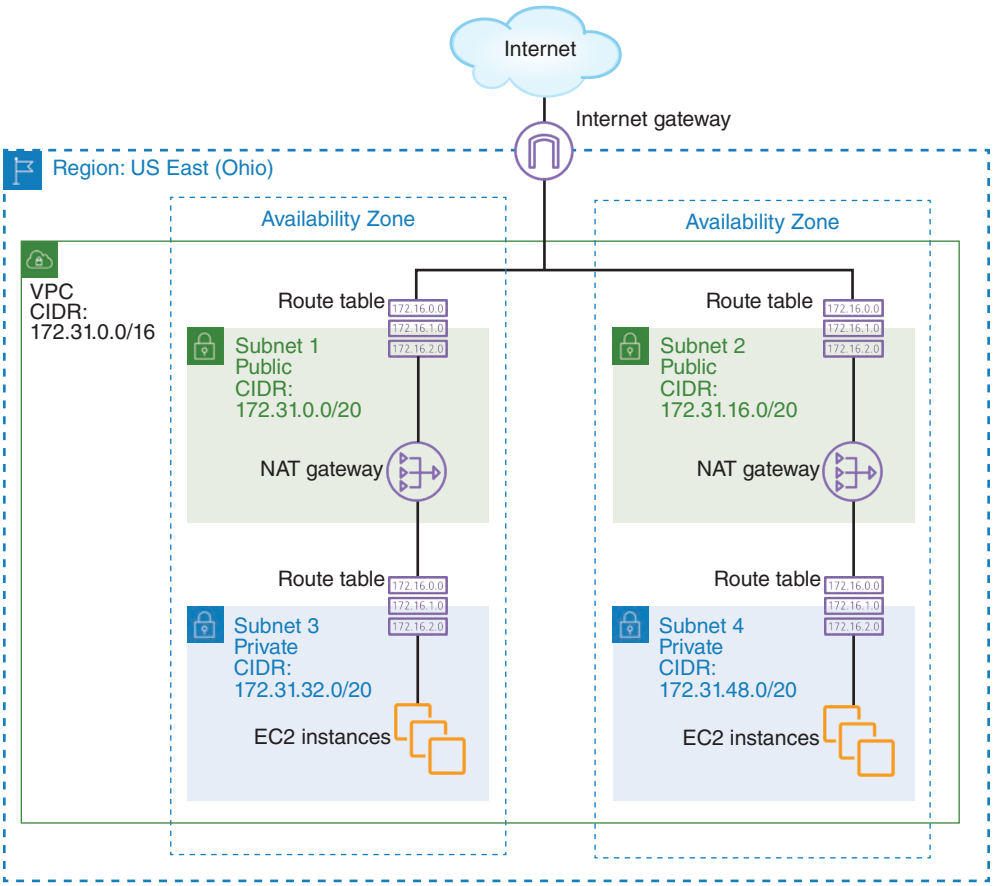
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Gateways and Route Tables in AWS

An Internet gateway is required to allow any of your VPC-hosted AWS resources to communicate with the Internet. In addition to this gateway, some common virtual network devices you might need in your VPCs include those identified in Figure 4-14 and described next:

- An **IG** or **IGW (Internet gateway)** is analogous to the cable modem on a home network in that it provides a path to the Internet for resources hosted within the VPC. IGs support both IPv4 and IPv6 traffic. They also provide NAT (network address translation) services for instances that have a public IP address within your public subnets.
- A route table is analogous to a router on a physical network. This list of routes determines where traffic from and within the VPC is routed. Route tables provide one of several layers of segmentation and security in AWS.
- An **NGW (NAT gateway)** device situated in a public subnet gives resources in a connected private subnet egress-only access to the Internet. This arrangement allows protected resources to communicate via the Internet, such as when obtaining server updates, without being exposed to public Internet traffic.

Figure 4-14 An Internet gateway, NAT gateways, and route tables in AWS



Routing to the Internet Gateway

A VPC's route table controls what traffic is directed where. By default, each VPC has a main route table, which is the default route table for all subnets in the VPC. You can make changes to this route table, or you can create new route tables for one or more subnets in your VPC. Many subnets can use the same route table, and each subnet must be associated with exactly one route table.

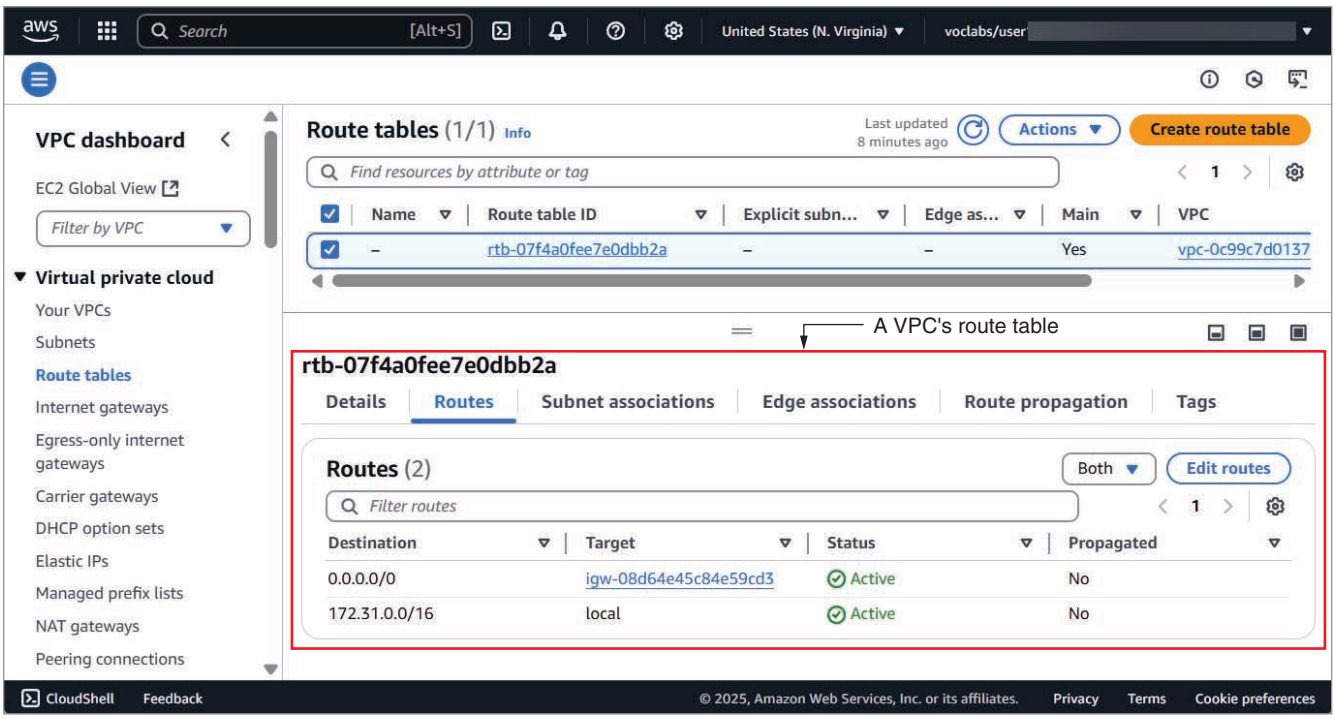
When a default VPC is created, an IG is also created and is connected to each public subnet within the VPC. This virtual device acts as a bridge, giving each public subnet and its instances access to the Internet.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Each time you create a new subnet within a VPC that has an attached IG, the subnet by default is configured as a public subnet with a route to the IG. Resources within a public subnet can communicate through the IG because a rule in the route table allows this communication. You can instead make a subnet private by associating with that subnet a route table that does not include a route to the IG.

Figure 4-15 shows the default route table for a default VPC. The table lists two routes: one that directs local traffic to the local subnet (enabling instances on the same subnet to talk to each other) and another one that directs all other traffic to the IG. The wildcard address 0.0.0.0/0 can be used to direct all unspecified traffic to the Internet.

Figure 4-15 Two routes send local traffic to the local subnet and all other traffic to the Internet gateway



Source: Amazon Web Services, Inc.

Consider the following details of these routes:

- The Destination field identifies the destination IP address or CIDR range the destination IP belongs to.
- The Target field identifies the IP address of the local device the traffic should be sent to for it to reach its destination. For example, the CIDR range 0.0.0.0/0 refers to all traffic not otherwise defined. The target for Internet-bound traffic (that is, traffic not already sent somewhere else) is the IG, which connects to the Internet.

This route table is designed to be used by a public subnet that allows all traffic between the subnet and the Internet. While you can delete the route allowing traffic to and from the IG (effectively making this subnet a private subnet), you can't delete or change the route allowing local traffic.

Applying Routes

Routes in an AWS route table are applied to traffic not by a ranked order like what you normally find on physical routers but by matching the traffic to the most specific route applicable to that traffic. For example, consider a route table with the routes listed in Table 4-5. The first route handles local traffic within the subnet. The second route provides a route to another VPC; this peering connection between the two VPCs allows resources in the VPCs to communicate via private IP addresses (as opposed to requiring public IP addresses). For performance reasons, you can even set a higher MTU (maximum transmission unit) for peered connections. VPC peering is somewhat analogous to adding a switch to connect physical networks. (You learn more about VPC peering in a later chapter.) The third route directs traffic to the Internet.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 4-5 Sample route table

Destination	Target
172.31.0.0/16 (This is the local subnet's CIDR range.)	Local (This traffic is directed to the local subnet.)
10.0.0.0/16 (This is the CIDR range of a peered VPC.)	pcx-1a2b1a2b (This traffic is directed to the peering connection, whose ID starts with "pcx".)
0.0.0.0/0 (This CIDR range covers all remaining traffic.)	igw-11aa22bb (This traffic is directed to the IG, whose ID starts with "igw".)

Consider communications coming from an instance on the local subnet and directed to an instance with the destination IP 10.0.0.25. This traffic most closely matches the route going to the peering connection pcx-1a2b1a2b and so would be sent to the peered VPC. However, traffic from the local subnet directed to the destination IP 8.8.8.8 doesn't match any of the more specific routes and would therefore be sent to the IG.

NAT Gateways

Instances running in a private subnet might need to communicate across the Internet while still being protected from Internet-initiated communication. This can be accomplished by creating an NGW in a public subnet and then routing Internet-bound traffic from the private subnet to the NGW instead of going directly to the IG.

Notice back in Figure 4-14, an NGW runs in each of the public subnets. A route table on each of the private subnets connects EC2 instances in those private subnets to one or the other NGW. This, in turn, allows the protected resources to communicate with the Internet. Using an NGW accomplishes multiple goals:

- Instances in the private subnets remain protected from Internet-initiated traffic.
- Instances in the private subnets can communicate outward to the Internet, such as when a server needs a software update.
- Fewer public IP addresses are required to allow this Internet-directed traffic.
- Traffic can be more tightly monitored and filtered.

NGWs have built-in redundancy. This means you do not have to create duplicate gateways for a single connection to ensure HA. Unlike IGs, NGWs cost money to use, both for the time the NGW is running and for the data processed over the connection. NGWs require an Elastic IP address, and the EC2 instances also incur data transfer charges.

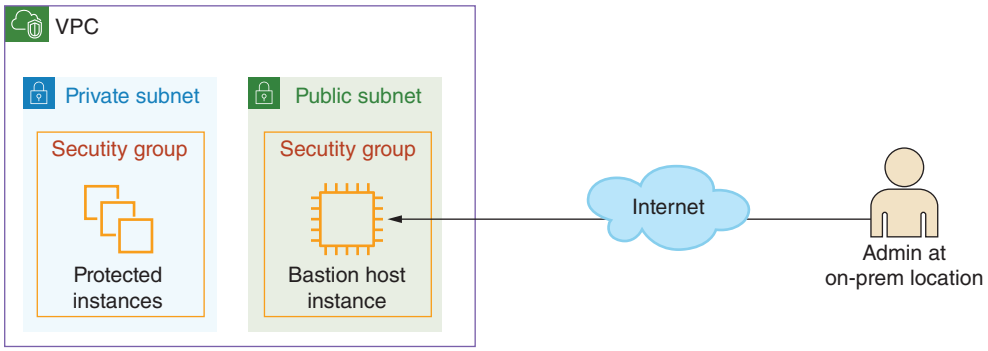
Note 4-10 | Alternatively, you can create a NAT instance, which is a VM instance hosting NAT services, to serve the same purpose as the NGW. Typically, however, the NAT gateway is the better option.

Remote Access to Cloud-Hosted Servers

There are times when you might need to support communication to these private subnets in the other direction—for example, you might need to access a server hosted in a private cloud subnet from your local computer via the Internet. This can be accomplished with a bastion host. A **bastion host** is a VM that resides in a public subnet, which you can access from your local computer over the Internet (Figure 4-16). Then the bastion host is given privileged access to the private subnet's resources. Through your access to the bastion host, you can then remote into a server that is hosted in the private subnet. This allows for stable protection around your private subnet while also supporting the administrative access you need from the Internet.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-16 Use a bastion host to remote into protected cloud VMs



Self-Check

7. Which AWS segment type is largest?
   - **a.** Availability zone
   - **b.** Region
   - **c.** Subnet
   - **d.** VPC

8. Availability zones provide _____.
   - **a.** increased throughput
   - **b.** decreased latency
   - **c.** decreased complexity
   - **d.** increased availability

9. Which AWS device's configuration options determine whether a subnet is public or private?
   - **a.** NGW
   - **b.** Route table
   - **c.** IG
   - **d.** Region

Check your answers at the end of this chapter.

Section 4-4: Networking in Azure

Certification

1.2 Explain concepts related to service availability.

1.3 Explain cloud networking concepts.

2.5 Given a set of requirements, provision the appropriate cloud resources.

Now that you've learned a bit about how networks are set up in AWS, you're ready to explore some of the variations you'll encounter in other cloud platforms. By identifying what tends to change between platforms, you'll develop a deeper understanding of the types of cloud networking services and the resources used to manage these network spaces.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As you've already read, Azure calls its virtual networks VNets instead of VPCs. You'll find there are more similarities than differences, however, between AWS's and Azure's networking services. As with AWS, you'll start with geographical regions and then work your way into smaller and smaller segmentation components.

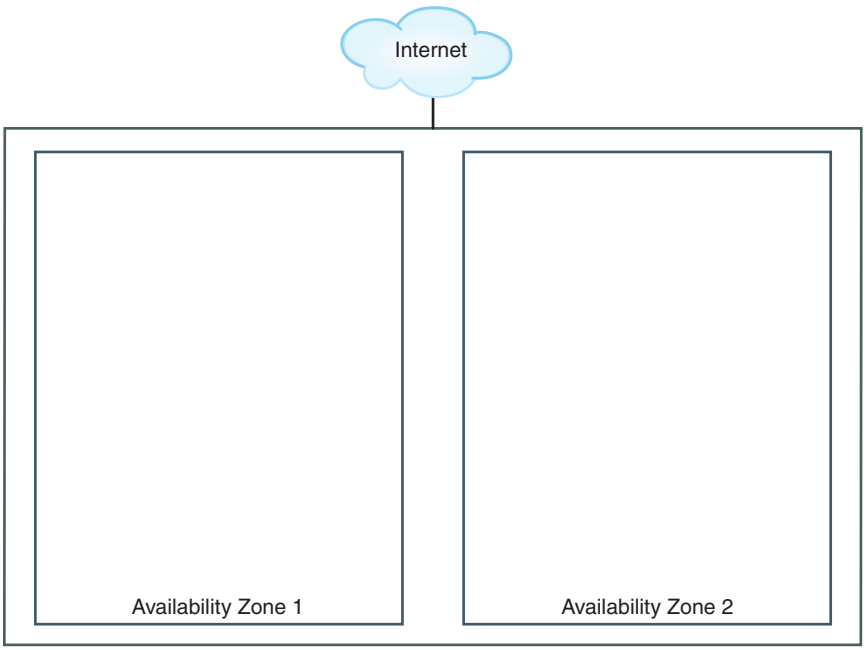
Regions and Availability Zones in Azure

At the time of this writing, Microsoft offers over 60 Azure regions worldwide with many more regions in the works. Each region offers different services, so it's important to check service availability in any region you're considering before you commit to it. Costs also vary by region, as do data residency compliance standards.

Regions are grouped into a larger category called geographies. A geography is a geographic area containing multiple Azure regions to provide data residency and compliance boundaries for discrete markets. For example, the United States geography contains 10 regions (plus additional government regions), such as West US in California and East US in Virginia. Other geographies include Canada, Australia, Europe, UK, India, and Asia Pacific.

Not all of Azure's regions currently offer AZs. The ones that do contain at least three AZs, as illustrated in Figure 4-17 (this diagram will be developed further as you read through this section, adding more layers of information). Each AZ consists of one or more completely separate physical data centers with independent power, cooling, and networking infrastructure. Microsoft guarantees at least 99.99% uptime when VM resources are backed up across at least two AZs. Although deploying VMs in multiple AZs can increase availability, there is also a charge for data transfer across AZs.

Figure 4-17 Azure availability zones reduce the risk of loss

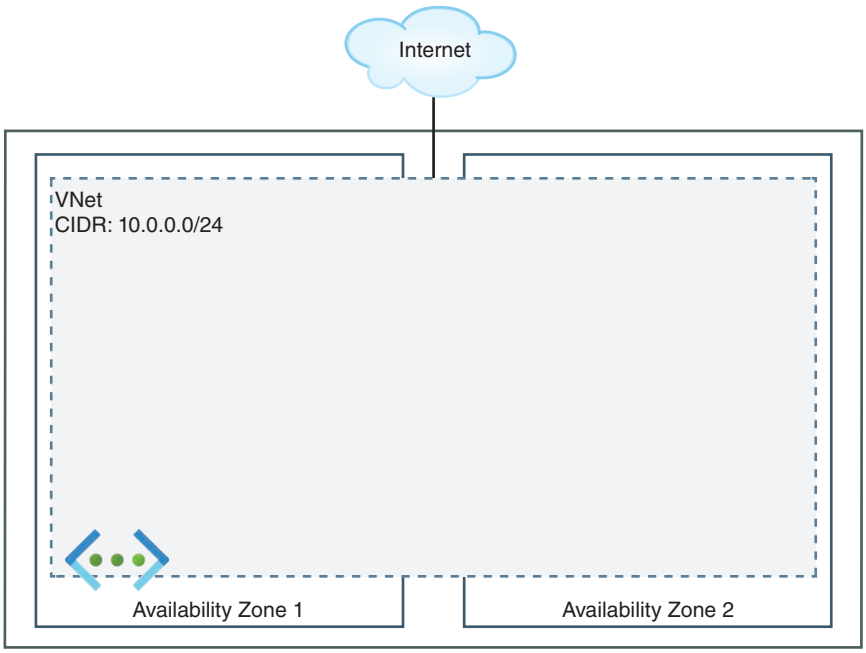


VNets and Subnets in Azure

Within each of your Azure subscriptions, you can create multiple VNets. Every Azure VM in your account must be associated with one VNet, even if it has multiple network interfaces. As with AWS's VPCs, each VNet can span multiple AZs within a region (Figure 4-18). This means a VNet can support resources residing in different AZs so long as those AZs are located within the same region.

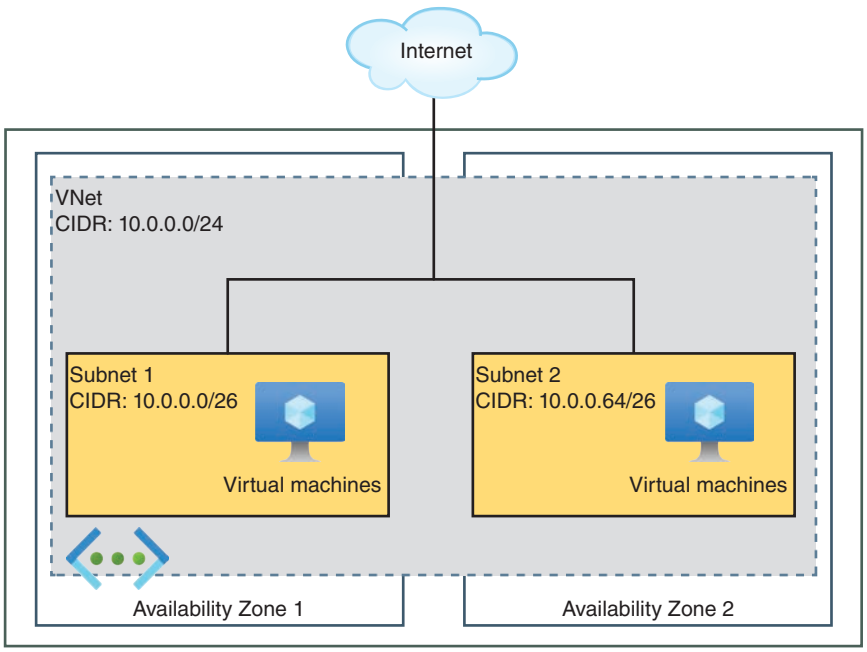
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-18 A VNet spans availability zones



Each VNet contains one or more subnets (Figure 4-19). Subnets must use an IP address space contained within the VNet's IP range, and each subnet must have a unique IP address range without overlapping any other subnets in the same VNet. By default, all subnets within a VNet can communicate with one another and with the Internet. You can change this setting if you need to restrict outside access to resources within a subnet. Similar to AWS, Azure reserves five IP addresses within each subnet.

Figure 4-19 Resources in Azure subnets can communicate by default



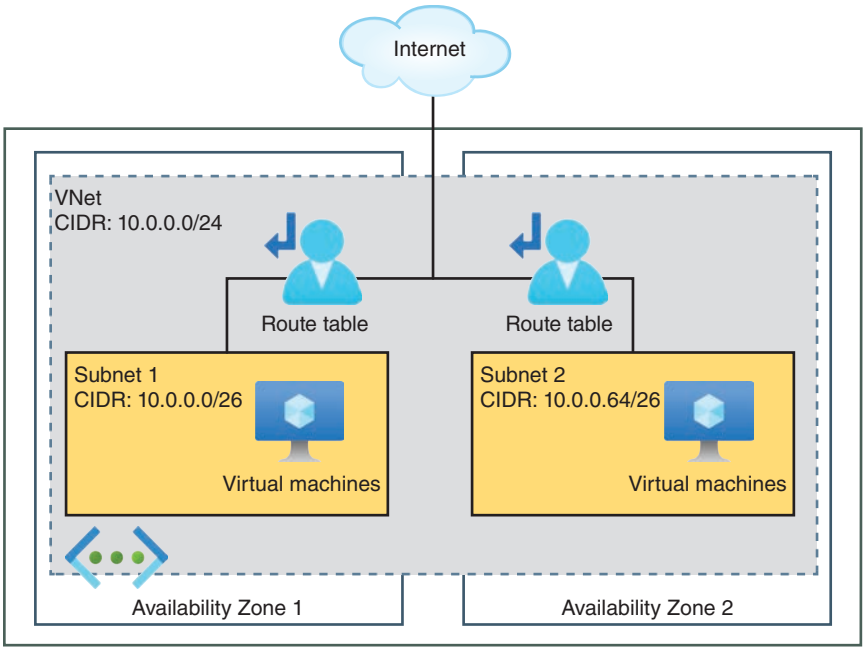
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 4-11 | The default IPv4 CIDR block range for a default VNet is /16. However, you can instead use any private CIDR block size from /8 (that is, 16,777,216 IP addresses) to /29 (that is, eight IP addresses) when setting this configuration manually, and you can add multiple private CIDR blocks to a VNet. The same restriction on CIDR block size applies to a VNet's subnets. However, a subnet's CIDR block must be contained within its VNet's CIDR block and not overlap any other subnet within the VNet.

Route Tables in Azure

When you first create a VNet, Azure automatically generates system routes that enable communication between all resources within the VNet, regardless of subnet, and allow for access to the Internet. These system routes can't be changed because you don't have access to them. However, you can create a route table with routes that will override the default system routes (Figure 4-20). Custom, static routes might be defined by the user, which is called UDR (user-defined routing), or routes can be learned from nearby gateways, such as a VPN gateway. Like AWS, Azure chooses the route to apply to traffic by how closely the destination IP address matches a route's address prefix.

Figure 4-20 An Azure subnet can be associated with a custom route table



Microsoft prefers that Windows VMs have access to the Internet in order to obtain updates. Therefore, Azure does not offer private subnets in the same way that AWS does. Although you can designate a subnet as private, outbound Internet connectivity is still available to resources located in that subnet. Azure also does not rely on IGs the way AWS does for managing outbound Internet connectivity. Instead, you can create a NAT gateway through the Azure Virtual Network NAT service. As with AWS, the NAT gateway instance is not a free resource.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

10. Which Azure segment type is largest?

a. Availability zone
b. Region
c. Geography
d. VNet

11. How does Azure choose which route to apply to traffic crossing a route table?

a. First match
b. Best match
c. Priority rating
d. Privacy setting

Check your answers at the end of this chapter.

Section 4-5: Networking in GCP

Certification

1.2 Explain concepts related to service availability.

1.3 Explain cloud networking concepts.

2.5 Given a set of requirements, provision the appropriate cloud resources.

GCP is set up very similarly to AWS with regions, VPCs, zones, and subnets. However, there are some significant differences. For example:

- A GCP VPC is global, meaning it spans all regions by default.
- A GCP VPC does not have an IP range assigned to it. CIDR ranges are defined only at the subnet level.

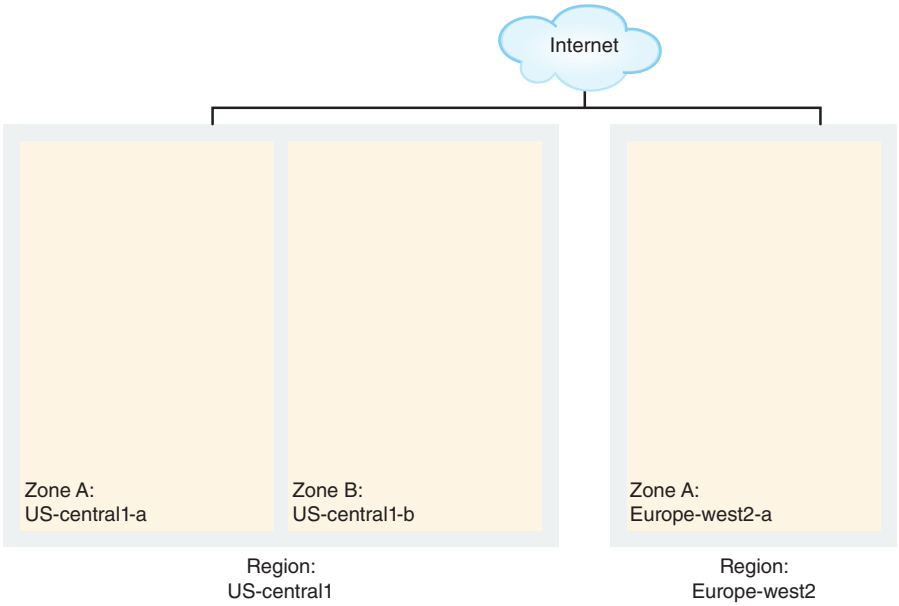
In this section, you explore the various components of networking in GCP.

Regions and Zones in GCP

At the time of this writing, GCP has over 40 regions worldwide, and most of these regions have at least three zones with independent power, cooling, networking, and control planes (which is the management layer of a software-defined infrastructure). Figure 4-21 illustrates how a zone can be associated with multiple regions (this diagram will be developed further as you read through this section, adding more layers of information). Because GCP VPCs span regions as well as zones, backing up resources and applications around the globe can be accomplished without leaving Google's private infrastructure. In other words, a VM in one region can communicate with a VM in another region without relying at all on public Internet infrastructure. As you might imagine, this provides a layer of security as well.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4-21 Most GCP regions have at least three zones

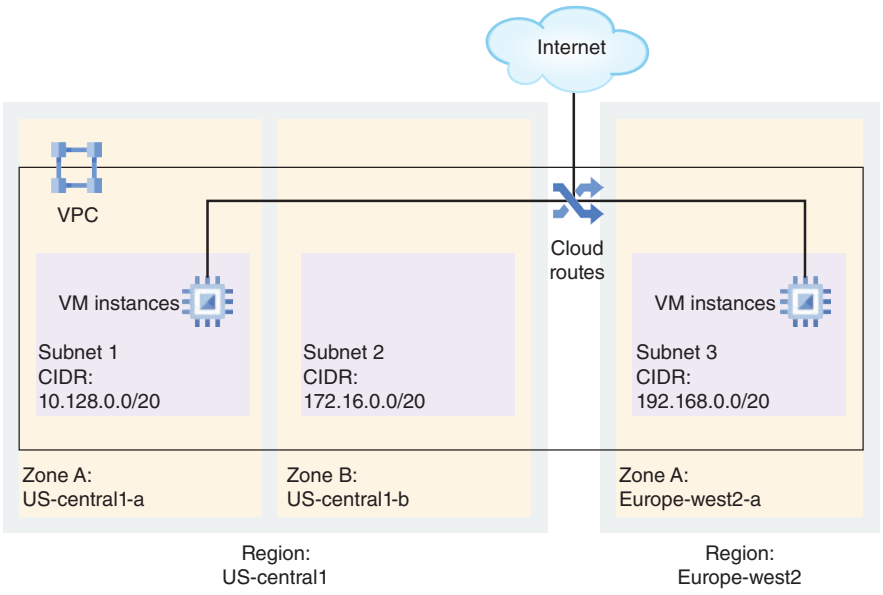


When choosing regions and zones to host your resources, select options that locate these resources geographically close to the users who will need to access them regularly. You might also choose a specific region or zone based on the services available there, the costs for those services, and the hardware available to support those services (such as CPU platforms, available machine types, and GPU availability).

VPCs and Subnets in GCP

As already mentioned, a GCP VPC spans multiple regions. When you create a VPC using auto mode, you automatically get a subnet in each region for that VPC. If a new region is added, auto mode adds a new subnet to your VPC for that region. Auto mode networks use preconfigured IP address ranges. You could instead use custom mode and choose where your subnets are located (Figure 4-22), or even whether to create any subnets at all. Somewhat similar to AWS, GCP reserves four IP addresses in each subnet, the first two and the last two of the primary IP address range.

Figure 4-22 A single VPC can span all of GCP's regions



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A significant advantage of global VPCs is that you can access distant regions from on-prem resources through a single VPC connection. In cloud platforms where virtual networks are limited to a single region, you would need a different connection from your on-prem data center to each region where you host resources, or you would need some other kind of transitory connection. (You learn more about network-to-network connections in a later chapter.) Much of the cost of networking in the cloud is related to connecting resources across regions. With GCP, you can avoid much of this expense due to the global access provided by a single VPC. Although there is still a fee for egress traffic (that is, traffic moving out of the VPC) across zones or regions, the rate is much lower than the cost for egress traffic to the Internet. In contrast, ingress traffic, or traffic entering a VPC, is not charged.

Note 4-12 | The default IPv4 CIDR block range for subnets in a VPC created using auto mode is /20. However, you can instead use any private CIDR block size from /8 (that is, 16,777,216 IP addresses) to /29 (that is, eight IP addresses) when setting this configuration manually. Keep in mind that a subnet's CIDR block cannot overlap any other subnet's CIDR block in the same VPC.

Routes in GCP

Routes in GCP are applied at the VPC level, not at the subnet level. Instances within the subnet maintain their own copy of applicable routes. When a route table is first created, two types of routes are automatically generated by the system. Additional, custom-defined routes can also be added. The different types of routes are as follows:

- System-generated
  - Subnet routes direct traffic between a VPC's subnets.
  - The default route sends all other traffic to the default Internet gateway.
- Custom
  - Dynamic routes are created automatically by a Cloud Router service.
  - Static routes are created manually by a cloud user.

While GCP does direct Internet-bound traffic to the default Internet gateway, this gateway is not a device you can access, reconfigure, delete, or create. You control a VPC's access to the Internet by managing the default route. You can prevent all Internet traffic between a VPC and the Internet by deleting the default route that connects to the Internet gateway. Routes in GCP are applied first by how closely a rule matches traffic and then by the priority rating of the route.

Self-Check

12. Which GCP resource type is assigned CIDR ranges?

a. Region
b. Zone
c. VPC
d. Subnet

13. Which GCP route type is created by a Cloud Router?

a. Static routes
b. Dynamic routes
c. Subnet routes
d. Default routes

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

Section 4-1: Networking Concepts in the Cloud

- SDN is a centralized approach to networking that removes most of the decision-making power from network devices and instead handles that responsibility at a software level.
- An overlay network is a virtual network that uses its own logical configurations to direct traffic on the underlying physical infrastructure, which allows customers to configure their own networking rules in their cloud environments without needing to configure underlying physical network devices.
- Just like with on-prem networks, you need to avoid any SPoF in your cloud networks, which often requires that you add redundant virtual devices, such as web servers or databases.

Section 4-2: IP Address Spaces

- Public IP addresses provide access to a resource from the Internet, and private IP addresses are only accessible from within the cloud-based network where that resource is located.
- In the cloud, private IP addresses are assigned chronologically from specific IP address ranges that are defined when you create the virtual network and its subnets.
- Subnetting helps manage IP address spaces in the cloud; within each subnet's IP address range are reserved addresses that cannot be used for hosts in the subnet.
- When a VM is created, a virtual network interface resource is also created with its own ID number and configuration settings; this vNIC provides a connection between the VM instance and its subnet in a virtual network.

Section 4-3: Networking in AWS

- A VPC (or VNet in Azure) is a software-defined portion of a larger, cloud-based network.
- Each AWS region is named for the geographic location of a group of AWS's physical data centers in that area and has an identity code.
- AZs address HA requirements by providing multiple isolated zones within each region.
- VPC settings—such as DHCP, DNS, route table, and the network ACL—carry over to any VM instances created within the VPC.
- Resources in public subnets have access to the Internet, whereas resources in private subnets do not have direct access to the Internet.
- A VPC's IG provides a path to the Internet for resources hosted within the VPC; a route table determines what traffic is directed where; and an NGW gives resources in a connected private subnet egress-only access to the Internet.

Section 4-4: Networking in Azure

- Azure regions are grouped into a larger category called geographies, where each geography contains multiple Azure regions to provide data residency and compliance boundaries for discrete markets.
- Each VNet can span multiple AZs within a region, and each VNet contains one or more subnets.
- Azure system routes enable communication between resources within a VNet, regardless of subnet, and allow for access to the Internet.

Section 4-5: Networking in GCP

- Because GCP VPCs span regions as well as zones, backing up resources and applications around the globe can be accomplished without leaving Google's private infrastructure.
- A significant advantage of global VPCs is that you can access distant regions from on-prem resources through a single VPC connection.
- While GCP does direct Internet-bound traffic to the default Internet gateway, this gateway is not a device you can access, reconfigure, delete, or create; instead you control a VPC's access to the Internet by managing the default route.

Key Terms

For definitions of key terms, check the Glossary.

bastion host
CIDR (classless interdomain routing) notation
IG or IGW (Internet gateway)
NFV (network functions virtualization)
NGW (NAT gateway)
overlay network
peering
private IP address
private subnet
public IP address
public subnet
RFC (Request for Comments) 1918
route table
scope exhaustion
SDN (software-defined networking)
subnetting
VNet (virtual network)
VPC (virtual private cloud)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

AZ availability zone
DNS Domain Name System or Domain Name Service
MTU maximum transmission unit
SDN software-defined networking
VPC virtual private cloud

Review Questions

1. What type of cloud network traffic incurs fees?
   a. Traffic between two VMs in the same subnet
   b. Ingress traffic to a storage volume
   c. Egress traffic from a database
   d. Traffic between two VMs in the same VPC
2. Which of the following are legitimate CIDR blocks for a subnet? **(Select two.)**
   a. 172.300.7.0/25
   b. 192.168.4.0/24
   c. 10.10.10.0/33
   d. 192.168.168.0/10
3. Which of the following concepts is ensured by redundant routers and switches?
   a. VPC
   b. HA
   c. SDN
   d. AZ
4. Which IP address belongs within the CIDR block 172.25.1.0/23?
   a. 172.25.2.10
   b. 192.25.1.100
   c. 172.25.0.254
   d. 192.24.1.0
5. You've created an EC2 instance in a private subnet to host a cluster of containers, and now you need to remote into it to make some configuration changes. Which technology will enable this connection?
   a. NAT gateway
   b. Port mapping
   c. CIDR block
   d. Bastion host
6. Which cloud stack layer corresponds to IaaS services?
   a. Physical layer
   b. Network layer
   c. Virtualization layer
   d. Image layer
7. Which redundancy in AWS would provide the greatest protection from a widespread outage?
   a. Duplicates across two subnets
   b. Duplicates across two VPCs
   c. Duplicates across two regions
   d. Duplicates across two AZs
8. How do you change applicable routes in an Azure VNet?
   a. Delete system routes.
   b. Edit system routes.
   c. Create a new VNet with different system routes.
   d. Override system routes with custom routes.
9. As you're setting up your VPCs for a new customer management application, you set all subnets across all VPCs to use the class B network 172.16.0.0/16. What is the most likely problem you'll encounter as your system grows and expands to eventually support a global market?
   a. Scope exhaustion issues
   b. Network overlap issues
   c. Service outage issues
   d. Regional service availability issues
10. Which cloud platform's VPC or VNet can extend beyond a single region?
    a. Azure
    b. VMware
    c. GCP
    d. AWS

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Scenario-Based Questions

Question 4-1

Ginny is building a series of subnets in her AWS cloud that will host VMs used to test applications her company is developing. The applications are tested by a team of users who are specialists in their respective industries. Ginny establishes the subnets, spins up a few experimental VMs, and starts testing her connections. She realizes there's a misconfiguration in one of the VMs and needs to remote into it to make some changes. What piece of information does Ginny need to collect from AWS first?

a. The VM's user password
b. The subnet's private IP address
c. The VM's public IP address
d. The AWS account's key pair

Question 4-2

Adan is mapping out address pools for the subnets he plans to create in his Azure account. He has determined the network will use a CIDR /24 address space, and he needs 16 small subnets within this space. How many bits should Adan borrow from the host ID to create these subnets?

a. 1
b. 2
c. 3
d. 4

Question 4-3

Tim is troubleshooting a problem with his Windows VM in AWS. He created a new VPC and a new subnet, enabled auto-assign public IPs, added an IG and a Windows VM, and attached the VM's interface to a subnet within the VPC. This is where he ran into problems. Tim has been trying for 30 minutes to open an RDP connection to the VM, and it's not working. He confirmed the VM is set to allow RDP traffic, and the VPC and subnet both also allow RDP traffic. He also confirmed the VM is running. What is the next step Tim must take to solve the problem?

a. He should install RDP on the VM.
b. He should configure a route to the IG.
c. He should assign the VM a public IP address.
d. He should create a new subnet in the VPC.

Hands-On Projects

Note 4-13 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Project 4-1: Configure a VPC and Subnets in AWS

Estimated time: 45 minutes
Objective 1.3: Explain cloud networking concepts.
Resources:

- AWS account
- Drawing app, such as Visio or app.diagrams.net
- Internet access

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Description: For security purposes, it's important to know the difference between public and private subnets in AWS and how to manage the traffic that can enter or exit each subnet. In this project, you will create a VPC and subnets and configure those subnets as either public or private. Complete the following steps:

1. In your AWS Management Console, go to the **VPC dashboard**. The VPC service is listed in the **Networking & Content Delivery** category.
2. In the left pane, click **Your VPCs** to list any existing VPCs. There is probably one listed already, which is your default VPC. In this project, you will create a new VPC. Click the **Create VPC** button.
3. AWS offers the option to create several networking resources at once. To display a diagram of these resources, under Resources to create, select **VPC and more**. A diagram gives a preview of the resources that would be created. Click **Show details** to display additional information about each resource. What resources are listed here?
4. In this project, you will create the needed resources manually. Under Resources to create, select **VPC only**. Under Name tag, give the VPC a name, such as MyVPC. What did you name your VPC?
5. Make sure **IPv4 CIDR manual input** is selected, and then specify an IPv4 CIDR block, such as 192.168.0.0/24. You don't need an IPv6 CIDR block, and make sure you use the **Default** tenancy. You also do not need to add any new tags in this project (you'll learn more about tags later). Click **Create VPC**.
6. Return to your list of VPCs to confirm your new, nondefault VPC is listed along with any other VPCs you already had in your account. What CIDR block is listed for your new VPC?
7. Click your new VPC's ID to go to the VPC's resource page.
8. AWS created a default route table for your new VPC. To find out what routes are included, in the new VPC's Details pane, note the main route table's ID—you'll need this information shortly. Click the main route table's ID to go to the list of route tables. You might have to click **Clear filters** for the route table to appear in the list. Click the correct route table's ID to go to its resource page.
9. Confirm the **Routes** tab is displayed. The route table should only include one route for local traffic to the VPC you just created and no other routes. What is the destination range for the local traffic route?
10. You do not yet have a subnet in this new VPC. To create a subnet, click **Subnets** in the left pane, and then click **Create subnet**. Select the VPC you created in Step 2. How do you know which VPC in this list is the new one you created earlier in this project?
11. Give the subnet a name, such as MyPrivateSubnet. You do not need to specify an AZ. Assign a CIDR block that is contained within and is smaller than the VPC's CIDR range, such as 192.168.0.0/26. Click **Create subnet**. Locate the new subnet in the list of subnets. What CIDR block did you use?
12. Your new subnet adopted the main route table from your VPC configurations. Click the new subnet's ID to go to its resource page. In the new subnet's Details pane, click the route table's ID, and check the routes to confirm this information matches what you found in Step 3.

Instances created within this private subnet will not be able to reach the Internet with the current configuration, as there is no route to the Internet. For VM instances within a public subnet to reach the Internet, you'll need to add an IG (Internet Gateway), as follows:

13. In the left pane, click **Internet gateways**. While you likely already have an IG listed here, for this project, you'll create a new IG. Click **Create internet gateway**. Give the IG a name, such as MyIG, and click **Create internet gateway**. What is the state of your new IG?
14. To attach this IG to your VPC, on the IG's page, click **Actions**, and then click **Attach to VPC**. Choose your VPC from the list and click **Attach internet gateway**. What is the state of your IG now?
15. Now you're ready to add a new subnet that will have access to the Internet through your new IG. Create a new subnet in the same VPC you created earlier in this project. This time, name it something like MyPublicSubnet. Use an adjacent CIDR block, such as 192.168.0.64/26. What CIDR block did you use?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Check the route table for your second subnet. This subnet also adopted the VPC's main route table and so only allows local traffic. To route Internet traffic through the IG, you could either change this route table (which would affect all resources using this route table, including your private subnet) or create a new route table. For this project, you'll create a new route table so the private subnet will stay private, as follows:

16. Go to the list of route tables and click **Create route table**. Give the route table a name (such as MyPublicRT), select the VPC you created earlier in this project, and then click **Create route table**. Go to the full list of route tables. You now have two route tables assigned to the same VPC ID. Which route table is the main route table for that VPC? How do you know?
17. Check the routes for your new route table. This new route table still only allows local traffic because you've not yet added a route to the IG. Click **Edit routes** and then click **Add route**. Traffic destined for the Internet can be described with the CIDR range **0.0.0.0/0**. Add this destination to your new route. For the target, choose **Internet Gateway** and then select your IG. Click **Save changes**. Confirm the new route was added to your public route table. Notice the item listed under Target for the new route. How can you tell this target for Internet traffic is an IG?
18. Now you need to add the new route table to the public subnet. Return to your list of subnets and select your public subnet. Click **Actions** and then click **Edit route table association**. Select your public route table to associate it with this subnet, confirm the routes are listed as expected, and click **Save**.
19. Unless you change this setting, nondefault subnets do not auto-assign a public IP address to resources in the subnet even if there's an IG attached to the subnet. To change this setting, on the **Subnets** page, select your public subnet, click **Actions**, and click **Edit subnet settings**. Select **Enable auto-assign public IPv4 address**, and then save your change.
20. To test your configuration of the public subnet, create a new VM instance. While configuring the VM, in the Network settings section, click **Edit**. Select your new VPC and the public subnet you created. Leave all other default network settings, and launch the instance (you do not need to download the key pair).
21. Return to the list of EC2 instances and examine the instance's configuration. Did your machine receive a public IPv4 address? If so, what is it? What is the VM's private IP address?
22. Use a drawing app, such as Visio or app.diagrams.net, to **draw the VPC and its resources** that you created in this project, showing logical connections and the relative, logical position of each resource. Submit this visual with your responses to this project's questions.
23. Delete all the resources you created in this project, including the VM instance, the VPC ***(do not delete the default VPC, only the VPC you created in this project!)***, both subnets, both route tables, and the IG. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
24. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 4-14 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 4-2: Configure a VNet and Subnets in Azure

Estimated time: 30 minutes
Objective 1.3: Explain cloud networking concepts.
Resources:

- Azure account and subscription
- Internet access

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Description: Each VM connects to a subnet through an interface. Given multiple interfaces, the VM can connect to multiple subnets. In this project, you'll create a VNet and two subnets, and then you'll practice managing VM interfaces. Complete the following steps:

1. In your Azure portal, create a new resource group. You don't need to use tags in this project (you learn more about tags in a later chapter). What did you name your new resource group?

> **Note 4-15** | Be sure to use this same resource group for all resources created in this project, and then delete the entire resource group at the end of the project.

2. Go to the **Virtual networks** dashboard. Click **Create virtual network**. Choose the resource group you just created, and give the network a name, such as MyVnet1. Click **Next** twice.
3. Assign an IPv4 address space, such as 192.168.0.0/24. What address space did you use?
4. In Azure, you create a new subnet at the same time you create a new VNet. By default, one subnet is already listed, named default. Edit the subnet and give it a new name, such as Subnet1. Change the IPv4 size to a smaller portion of the VNet's CIDR range, such as /26. How many addresses are included in the subnet's CIDR block? Click **Save**.
5. Click **Next** twice more to keep the remaining default settings for your new VNet, and then click **Create**.
6. Create a new VM in the VNet using the same resource group. On the Networking page, be sure to select the VNet you created and its subnet. Then create the VM.
7. After the VM is deployed, click the VM. Click **Networking** in the left pane of the VM's resource page, and then click **Network settings**. What is the VM's private IP address? What is the name of this network interface?
8. Above the network interface information, click **View topology**. Describe the information given in this diagram.
9. Now you're ready to add a second network interface to this VM. Power down the machine (but don't delete it). Make sure the VM is completely stopped.
10. Return to the VM's Network settings page and click **Attach network interface**. Click **Create and attach network interface**. Give the network interface a name, such as MyNewInt. What options do you have for selecting a VNet? Why do you think this is?
11. Leave the default settings in place and click **Create**.
12. When complete, the VM's Network settings page should show the information for the new interface. What IP address is assigned to this interface?
13. Click **View topology** and describe the information given in this diagram.

Both interfaces are included in the same subnet. Also, you were not able to add the new interface to a different VNet than the one the first interface was already attached to. Can you place the second interface in a different *subnet* from the first interface? To determine the answer to this question, complete the following steps:

14. Create a second subnet in the first VNet. Give the subnet a name and an address space within the first VNet's CIDR block, such as 192.168.0.64/28. What address space did you use?
15. After the subnet is created, return to the VM's Network settings page. Associate the second network interface with the new subnet. To do this, you will need to visit the resource page for the network interface's IP configuration, change the subnet selection, and apply the changes. What is the new IP address for this interface?
16. What changed on the VM's network topology? **Take a screenshot** of the topology; submit this visual with your responses to this project's questions.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

17. Delete all the resources you created in this project, including the VM instance, the VNet, both subnets, both interfaces, and the entire resource group. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
18. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 4-16 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 4-3: Configure a VPC in GCP

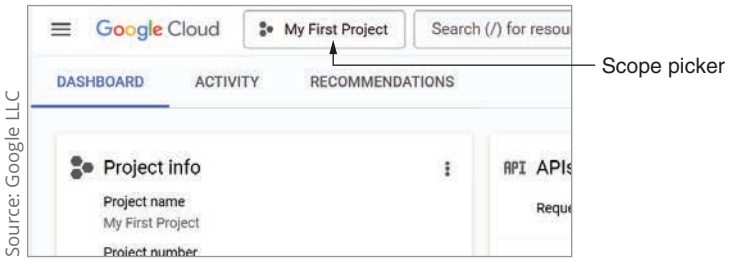
Estimated time: 30 minutes
Objective 1.3: Explain cloud networking concepts.
Resources:

- GCP account
- Drawing app, such as Visio or app.diagrams.net
- Internet access

Description: GCP VPCs are unusual. Subnets within a VPC do not need to use a CIDR block within the VPC's address space because the VPC doesn't have an address space assigned to it. In this project, you'll explore VPC configuration options in GCP, comparing auto subnet creation mode with custom subnet creation mode. Complete the following steps:

1. To make cleanup easier at the end of this project, first click the scope picker at the top of your GCP console, as shown in Figure 4-23, to create a new project. Click **NEW PROJECT** and give this project a name. Click **Create**. What did you name your project?

Figure 4-23 Creating a new project makes it easier to clean up everything later



2. Select the project you just created. Then go to the VPC networks dashboard. You might need to enable the Compute Engine API to proceed. When ready, click **CREATE VPC NETWORK**.
3. Give the network a name, such as myautovpc. Under Subnet creation mode, select **Automatic**. Leave all other default settings and click **CREATE**.
4. On your VPC networks page, locate your new VPC in the list. After it's fully deployed, how many subnets are included in your new VPC?
5. Click the VPC name, and then click the **Routes** tab. Select a region in the dropdown list, and click **VIEW**. If no routes are listed, try a different region. What is the next hop for most of the routes? What is the next hop for the Internet route?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6. Return to your list of VPC networks, and create a new VPC. This time, use the **Custom** subnet creation mode. Name the subnet and select a region. (Do not select a region in Asia, as it will cause an error later in the project.) Assign a private IP address range to the subnet, such as 192.168.0.0/24. Notice the subnet receives an IP address range but the VPC does not. What CIDR range did you choose?
7. Create a second subnet in a different region—make sure you're creating a second subnet, not just adding a secondary IPv4 range to the first subnet. (Do not select a region in Asia, as it will cause an error later in the project.) Give it the same CIDR range as the first subnet and try to create the VPC. What happens?
8. Try creating your custom VPC again, and this time give the second subnet a CIDR range that is not in the same class as the first subnet's CIDR range. For example, if you used 192.168.0.0/24, try 172.16.0.0/16 for the second subnet. What happens?
9. After the VPC is fully deployed, how many subnets are included in the newest VPC? Which two regions did you use?
10. Check the routes for this VPC. How many routes are included in each region? Where do they each go?
11. Use a drawing app, such as Visio or app.diagrams.net, to **draw the custom VPC and its resources** that you created in this project, showing logical connections and the relative, logical position of each resource. Submit this visual with your responses to this project's questions.
12. Delete all the resources you created in this project, including both VPCs and all subnets. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems? Check through your account to confirm that all related resources have been deleted.
13. Go to the list of projects and shut down the project you created in Step 1. Confirm the project is shut down and scheduled for deletion.
14. Make sure to properly sign out of your account.

Note 4-17 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Challenge Projects

Note 4-18 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 4-1: Deploy a Cloud Network

Estimated time: 45 minutes (+10 minutes for group work, if assigned)
Objective 2.5: Given a set of requirements, provision the appropriate cloud resources.
Group work: This project includes enhancements when assigned as a group project.
Resources:

- Public cloud account (such as AWS, Azure, or GCP)
- Drawing app, such as Visio or app.diagrams.net
- Internet access

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Description: You've now learned how to deploy compute and networking resources in the cloud. In this Challenge project, you will choose a cloud platform and create resources for a cloud deployment. Complete the following steps:

1. If you need to create a project or resource group to manage project resources when you're finished, do so now.
2. Using the tools and services available to you in this platform, create resources needed for the following cloud deployment:
   a. One VPC or VNet
   b. Two subnets in the VPC or VNet using the CIDR blocks 172.16.0.0/24 and 172.16.1.0/24.
   c. A VM instance in each subnet with access to the Internet (including a public IP address)

Consider the order of steps you should take to create these resources and think about which resources require which configurations to meet the deployment requirements (for example, what CIDR block should be assigned to your VPC or VNet, if any?).

3. Use a drawing app, such as Visio or app.diagrams.net, to **draw the VPC or VNet and its resources** that you created in this project, showing logical connections and relative position of each resource.
4. When your deployment is complete and fully running, **take screenshots** to document the configurations required in Step 2. Submit this visual with your responses to this project's questions.
5. **For group projects:** Each group member should share the public IP address of one of their VMs with another group member. When you receive a public IP address from one of your group members, create a route in the route table for one of your subnets or for your VPC, depending on the platform you're using. The route should direct traffic from that network space to your group member's public IP address. Consider the following:
   - If the destination field requires a CIDR block rather than a single IP address, use the slash notation /32 to indicate the CIDR block represents a single host.
   - You might need to identify an Internet gateway or some other device as the target for traffic traveling to that destination.
6. Delete all the resources you created in this project. Check through your account to confirm that all related resources have been deleted.

Note 4-19 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Solutions to Self-Check Questions

Section 4-1: Networking Concepts in the Cloud

1. Which theoretical plane moves traffic from switch port to switch port?

   Answer: b. Data plane

   Explanation: The **data plane** is where transmissions on the network actually traverse physical devices. The control plane, also called the management plane, is the layer where devices are configured to determine rules for directing traffic. SDN (software-defined networking) separates the control plane where decision making occurs from the data plane where bits are transferred from device to device.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. At which cloud layer are user access accounts managed?

   Answer: d. Network layer

   Explanation: The **network layer** is where server resources, access control, many costs, cybersecurity, and other services are managed. The virtualization layer abstracts virtual hardware from physical hardware through virtualization software. The image layer provides instances or platform-native services that support features at the application layer. The physical layer contains physical infrastructure, including cables, servers, networking devices, cooling systems, and power systems.

3. Who configures physical devices hosting the cloud?

   Answer: c. The CSP

   Explanation: The underlying physical infrastructure that supports the cloud is not accessible to the cloud customer and can only be configured by the **CSP (cloud service provider)**. The IaaS (Infrastructure as a Service) user can create and manage virtual networks at the network layer. The PaaS (Platform as a Service) user can manipulate various OS (operating system) environments and can set some infrastructure preferences at the image layer. The SaaS (Software as a Service) user accesses cloud-hosted applications at the application layer.

Section 4-2: IP Address Spaces

4. What kind of address is 8.8.8.8?

   Answer: d. Public IP address

   Explanation: Public IP addresses provide access to a resource on the Internet and are not found within a private IP address range. The IETF's RFC (Request for Comments) 1918 identifies three blocks of private IP addresses reserved for use on private networks. IPv6 (Internet Protocol version 6) is a newer version of IP that offers significantly more IP addresses, increased security, and greater efficiency for network traffic

5. If a VM is assigned the address 172.16.18.10/24, what is its network ID?

   Answer: b. 172.16.18

   Explanation: The /24 indicates the first three octets identify the network ID, which are **172.16.18**. A CIDR of /16 could have a network ID of 172.16 or 192.168. A CIDR of /24 could have a network ID of 192.168.18, but that value does not match the IP address given in the example.

6. When you delete a vNIC in the cloud, what else is also deleted?

   Answer: d. Nothing—the vNIC is a separate entity

   Explanation: vNICs (virtual network interface connections) are **separate entities**, or resources, in cloud networking that can be created and deleted independently of these actions performed on VMs. You can also move vNICs between VMs or subnets. The VM instance's storage drive is not affected by changes to its vNIC.

Section 4-3: Networking in AWS

7. Which AWS segment type is largest?

   Answer: b. Region

   Explanation: Each AWS **region** contains multiple AZs (availability zones) and multiple physical data centers. Subnets are contained within AZs and VPCs (virtual private clouds), which are contained within regions.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8. Availability zones provide _____.

Answer: d. increased availability

Explanation: Availability zones address HA (**high availability**) requirements by providing multiple isolated zones within each region. Spreading resources across AZs can increase latency and complexity while potentially decreasing throughput, depending on configurations.

9. Which AWS device's configuration options determine whether a subnet is public or private?

Answer: b. Route table

Explanation: Resources within a public subnet can communicate through the IG (Internet gateway) because a rule in the **route table** allows this communication. An NGW (NAT Gateway) gives resources positioned in a private subnet one-way access to the Internet. A region is a collection of data centers within a geographic area.

Section 4-4: Networking in Azure

10. Which Azure segment type is largest?

Answer: c. Geography

Explanation: Azure regions are grouped into a larger category called **geographies**. A geography is a geographic area containing multiple Azure regions to provide data residency and compliance boundaries for discrete markets. Each VNet can span multiple AZs within a single region.

11. How does Azure choose which route to apply to traffic crossing a route table?

Answer: b. Best match

Explanation: Like AWS, Azure chooses the route to apply to traffic by **how closely the destination IP address matches** a route's address prefix. It does not use first match, priority rating, or a privacy setting to select a route.

Section 4-5: Networking in GCP

12. Which GCP resource type is assigned CIDR ranges?

Answer: d. Subnet

Explanation: In GCP, CIDR ranges are defined only at the **subnet** level. A GCP VPC does not have an IP range assigned to it. Regions and zones are not used to manage IP address spaces in GCP.

13. Which GCP route type is created by a Cloud Router?

Answer: b. Dynamic routes

Explanation: Dynamic routes are created automatically by a Cloud Router service. Static routes are created manually by a cloud user. Subnet routes direct traffic between a VPC's subnets. The default route sends all other traffic to the default Internet gateway.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Chapter 5

Cloud Connectivity

Chapter 5 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Compare technologies used to connect hybrid and multi-cloud networks.
2. Explain how common network services can be extended across hybrid and multi-clouds.
3. Describe technologies that extend the cloud to its points of use.
4. Explain troubleshooting strategies to address common cloud networking issues.

Chapter 5 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 5-1: Hybrid Cloud and Multi-Cloud Networking

- Connecting Networks
- Connecting Cloud Networks

Section 5-2: Extending Network Services

- DHCP Services
- DNS Services
- Routing
- Load Balancing

Section 5-3: Extending the Cloud

- The Case for Edge Computing
- CDN (Content Delivery Networks)
- IoT Cloud Networking

Section 5-4: Troubleshooting Cloud Connectivity

- Common Connectivity Problems
- HTTP Status Codes
- Misconfigured Routes
- NTP (Network Time Protocol)
- VLAN Issues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 5 Scenario "Doodles!" squawks Liz as she rounds the corner into the IT Department's lobby area. Liz is the school's HR director and a close friend of Kendra's. A couple of years ago, Liz started calling Kendra "Doodles" because she always drew little pictures on the back of her notes during staff meetings. Kendra returned the favor with a nickname for Liz...

"Sunshine! You're all bright and happy today, as usual. What's in the box?" Kendra asks.

"Doughnuts for you and your gang," replies Liz with her typical warm tone. "They're running a fundraiser upstairs. I can't eat these rings of sugary goodness because of my blood sugar, but I can live vicariously through you! Here, Malik, pass them around." Liz hands the box of doughnuts to a drooling Malik.

"That's so *sweet* of you!" Kendra smirks as she scoops up a warm doughnut.

"Ha-ha, aren't you the clever one," Liz says. "Lunch today? Our usual time?" Liz and Kendra always go to lunch together on Fridays.

"Absolutely! See you then," Kendra mumbles through a mouthful of doughnut.

As Liz walks out, Kendra swallows then lets out a long sigh.

"What's wrong?" you ask as you wipe a few crumbs off your desk and take a sip of coffee to put the finishing touch on your own doughnut. "Don't you want to go to lunch with Liz?"

"Oh, it's not that," replies Kendra. "Seeing Liz just reminded me of something. HR uses an old, custom program to manage contacts with alumni and other school supporters. It's so old and so convoluted that we haven't even been able to migrate it to an updated Windows server. There's no way we'll be able to put that thing in the cloud."

"Can we move the data to a SaaS solution?" you ask. "Surely there's something out there that would work."

"Maybe," says Kendra. "But this software is pretty customized. The school had someone build it from scratch back in the early '90s. Migrating the data to a different system will be a nightmare."

Malik heard the last part of Kendra's statement as he walked back into the office with an empty doughnut box. "What happens if we can't move everything to the cloud? Is this an all-or-nothing deal?"

"No," says Kendra. "We can certainly run a hybrid cloud. That means we'll have some of our systems in the cloud, and some things will just have to stay on-prem."

"So how does that work?" you ask. "Do we have to find a dividing line, or can those systems communicate with each other somehow?"

"That's a good question," replies Kendra. "There are ways to allow communication between our cloud and our on-prem network. Some options are better than others, and there are limitations on how seamlessly we can blend resources across that gap. Let's research this some more." She scribbles some notes on the back of yesterday's staff meeting agenda. She hands the sheet to you, doodle side up, where you can read the following questions:

- How does the on-prem network connect with the cloud?
- What network services can be extended across a hybrid cloud?
- What are some ways to improve performance by reducing latency?
- What problems are likely to crop up in managing these connections?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 5-1: Hybrid Cloud and Multi-Cloud Networking

Certification

1.3 Explain cloud networking concepts.

1.7 Compare and contrast virtualization concepts.

1.10 Compare and contrast methods for optimizing workloads using cloud resources.

2.5 Given a set of requirements, provision the appropriate cloud resources.

Recall that a hybrid cloud is an organization's network that exists both on-prem and in the cloud. Similarly, a multi-cloud is an organization's network that spans two or more cloud platforms, such as AWS and GCP or Azure and Salesforce. It seems this architecture would introduce significant complexity in getting the various network services and resources to talk to and support each other. So why would an organization create this metaphorical monster?

The reality is that hybrid and multi-clouds are becoming more reasonable and popular due to their flexibility and potential cost benefits. A hybrid cloud is a typical transition phase as companies migrate to the cloud, and many companies ultimately decide to continue indefinitely with a hybrid infrastructure. This often makes sense financially because, while many cloud services such as compute can be relatively inexpensive, other cloud services, such as data transfer, can be quite costly. Where it might make good sense to run web servers in your cloud, it might be more cost-effective to keep media storage on-prem for cheaper access by the company or its customers. In other cases, the reverse might be true. Another example is when application development might benefit from cloud adaptability, but application hosting might work better in your own data center's production environment, and cloud storage could be used for inexpensive archival storage that is rarely accessed.

Consider the following potential use cases for a hybrid cloud:

- **Application development**—Develop and test applications in the cloud, and then move code and related files to the on-prem production environment for stable releases of the application.
- **Market testing**—Determine the demand for a new application before investing on-prem infrastructure into hosting that application.
- **Cloud bursting**—Temporarily extend a workload that has exceeded its on-prem capacity until the short-term demand decreases to a normal range.
- **High availability and disaster recovery**—Inexpensively establish redundancy that can share workloads with on-prem resources or that can take over in an emergency.
- **Compliance**—Retain regulated data storage on-prem to comply with relevant restrictions while offloading less sensitive data, applications, servers, or other resources to the cloud.
- **Archives**—Store in the cloud archived data that is not likely to be accessed often, thereby greatly reducing the cost of data storage.

Reasons for deploying a multi-cloud infrastructure are similar in nature to those of a hybrid cloud along with additional benefits, including:

- **Features**—Choose services from different platforms that best fit the organization's needs, and maintain portability where possible to benefit from emerging cloud technologies.
- **Cost optimization**—Negotiate for the best prices from each vendor, and then avoid vendor lock-in to continue to take advantage of potential cost savings offered by various CSPs.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Edge locations**—Leverage regional availability and proximity of various CSPs to maximize performance and minimize latency.
- **Disaster recovery**—Spread redundancy across multiple CSPs to minimize the effects of one CSP's outage by having another CSP host part of the network.

Hybrid and multi-clouds do present significant challenges, however. For example, not all services are portable across CSPs. For that matter, not all skill sets are portable across cloud platforms, either. Additional cross-network connections will need to be created and maintained, and additional security measures are needed to protect data in and between the various cloud platforms. While the debate continues as to whether it's better to focus an organization's resources on deeply integrating a single cloud platform (such as AWS) or spread its interests across multiple platforms, you as an IT professional can offer your company more options if you understand how to blend network services across hybrid and multi-clouds.

Connecting Networks

When extending your network to the cloud, the connection between your on-prem data center and your cloud becomes a crucial factor in the overall success of your deployed systems. Multiple infrastructure options offer varying degrees of cost savings, privacy, performance assurance, and complexity. Three major categories of options to make the connection from on-prem to the cloud include the following:

- Use the public Internet as you've been doing in this course's projects.
- Rely on VPN connections. A **VPN (virtual private network)** is a virtual connection over the Internet between two geographically distant locations that provides access to network resources between a client and a remote network, between two remote networks, or between two remote hosts.
- Subscribe to some variety of direct connection service.

Relying on the open, public Internet with no added layers of security or encryption is certainly the least secure of these options. It can also be the least predictable in terms of performance. Many Internet connections run over broadband, where overcommitment of network bandwidth by the ISP (Internet service provider) can result in slower throughput when many customers access the Internet at the same time. The direct connection option can be secured more thoroughly, yet the characteristics and service features differ depending on the cloud provider. This also tends to be the most expensive option. Somewhere in the middle between these options is the use of a VPN, which might run over public Internet infrastructure or over a private WAN (wide area network) connection. In this section, you examine the issues to consider when comparing VPNs to direct connection options from some of the major CSPs.

VPN (Virtual Private Network)

VPNs are an inexpensive and relatively secure mode of connection to your cloud resources. Regardless of what other options you use, many organizations rely on VPNs to give remote workers, traveling salespeople, and contract partners access to their cloud resources. AWS, for example, calls this service Client VPN. In contrast, some organizations use VPNs as the primary connection between the on-prem data center and their cloud-based VPCs. AWS calls this service Site-to-Site VPN.

A VPN creates an encrypted tunnel for traffic to safely traverse the public Internet, which means that latency can be an issue depending on Internet "weather conditions." Availability of the connection might not be covered by an SLA (service level agreement), or high availability might be guaranteed only if you establish and maintain multiple VPN tunnels for the same connection. Alternatively, a VPN tunnel can be established across a private connection, although this is more expensive.

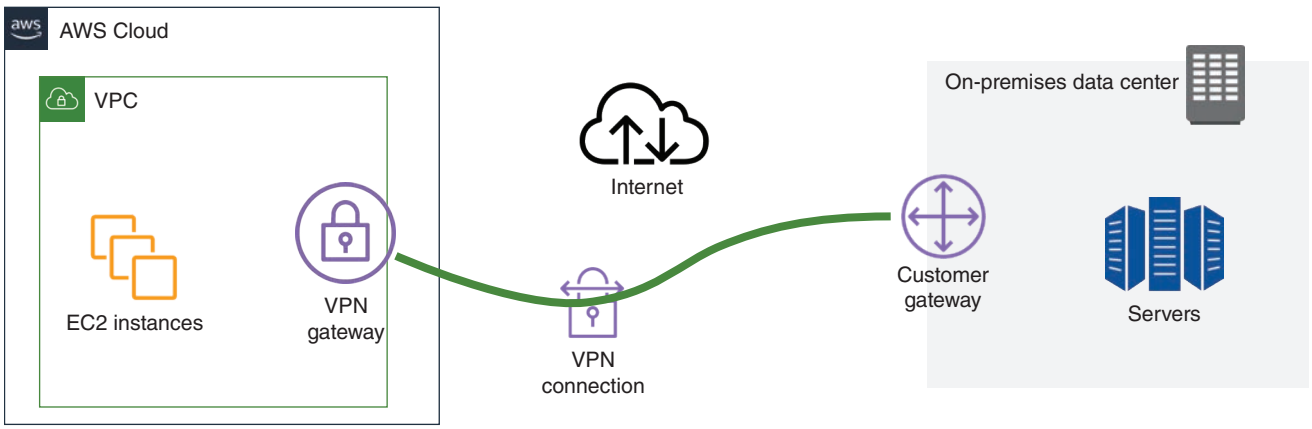
To support cloud-based VPN connections from individual users, such as remote workers, you need a VPN endpoint resource configured in your VPC. You then send a client configuration file to your users, who need a VPN client application installed on their computers. The VPN service scales automatically as user demand increases or decreases. Your organization is typically charged by connection-hours.

To create a VPN connection from a VPC to another VPC or to an on-prem data center, you need a VPN gateway (also called VPN endpoint) in each network. For example, if you're connecting your data center in

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Memphis to your VPC on AWS, you need a CGW (customer gateway) appliance in your Memphis network and a VGW (virtual gateway) service running on your VPC (Figure 5-1). The CGW might be a hardware VPN gateway device or a virtual VPN gateway application hosted on a server. The VGW is hosted in a specific VPC and provides access to resources running within that VPC.

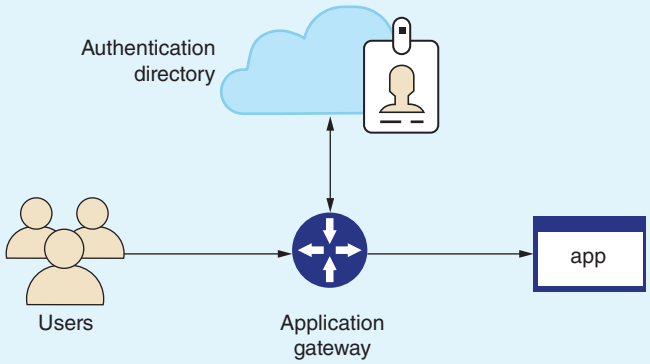
Figure 5-1 A VPN tunnel carries encrypted traffic over the Internet between two VPN gateways



To increase availability of the connection, you can create multiple VPN tunnels hosted on each VPN endpoint. In some cases, you'll also need a separate tunnel (or group of tunnels) for each VPC in your cloud.

Note 5-1 | An **application gateway**, or app gateway for short, is a service that manages user authentication to give users remote access to applications without the requirement of a VPN (Figure 5-2). The application gateway can source the same credentials for authentication as are used for the organization's other network services and applications, such as email. Many cloud-hosted applications often already rely on shared authentication directories to simplify sign-in credentialing for users. However, many companies face challenges when they try to incorporate these same authentication processes for their own applications hosted on-prem. Instead, they require users to remote into the network using a VPN to access internal applications. An application gateway running on the company's own servers can give users a unified, convenient, and secure access experience to company-hosted apps and cloud-hosted apps.

Figure 5-2 The application gateway handles authentication to provide secure app access without a VPN



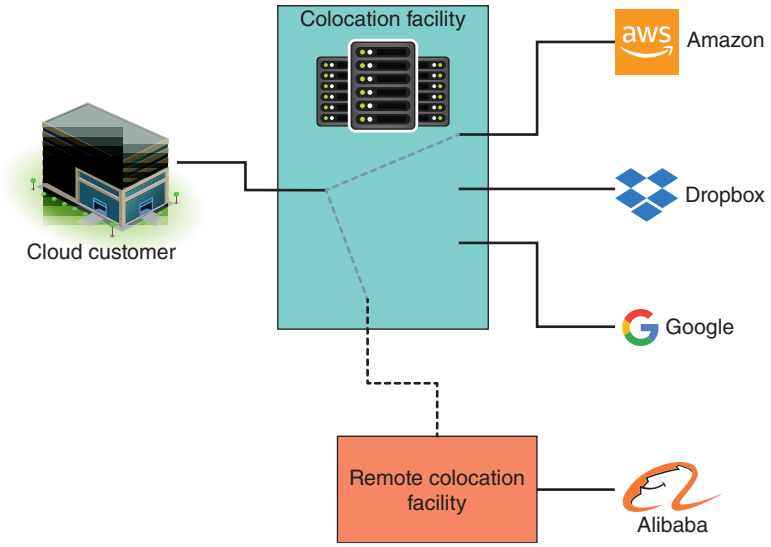
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Direct Connection

In the context of cloud networks, VPNs are often used to connect a single device to cloud resources or to connect a cloud network to an on-prem network over the public Internet. In contrast, a direct connection uses dedicated, private infrastructure to connect an on-prem network with a cloud provider's physical infrastructure. Creating a direct connection from your data center to your public cloud can be accomplished by meeting the CSP in a common location called a colo (colocation). The colo is a data center facility dedicated to interconnecting service providers (both ISPs and CSPs) with their business partners and customers. For example, Equinix currently offers more than 250 colocation facilities worldwide and offers direct connect options for customers to connect with AWS, Azure, GCP, Oracle, Zoom, and many other service providers.

By connecting to the colo, a cloud customer has the option to request direct connections within that facility to any number of service providers to support a hybrid or multi-cloud deployment (Figure 5-3). Each of these connections offers low latency and assured bandwidth that don't rely on public Internet conditions or physical infrastructure. Although direct connect options are more expensive than relying on the Internet, customers maintain much greater control of the connections and related costs while maximizing availability and performance. AWS calls its direct connection service Direct Connect. In Azure, the service is ExpressRoute. And in GCP, it's called Cloud Interconnect. These direct connections are managed by VLAN (virtual local area network) technology, which you learn more about later in this chapter.

Figure 5-3 A colocation facility for creating direct connections between service providers and their customers



Network Throughput Capacity on Direct Connections

When establishing a connection to the cloud using tools such as VPNs or a dedicated service such as AWS's Direct Connect, one of the factors to consider is the throughput supported by that connection. For example, AWS's Direct Connect is available at various tiers of throughput capacity, from 50 Mbps up to 400 Gbps. Once the connection is created, its bandwidth can't be changed—a new connection must be created with the new bandwidth amount to replace the old connection.

Importantly, configured bandwidth is used to determine much of the cost associated with these services as well as the performance and user experience of the connection and anything that relies on it. Additionally, other virtual devices in the path of the connection, such as a transit gateway or load balancer, can also contribute to the overall cost. The amount of outgoing data affects cost, but interestingly, incoming data often is not charged on direct connections. Careful consideration for the location of these devices in the context of your VPCs, subnets, workloads, and other services can optimize costs and throughput.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Segment Direct Connections

For security purposes, traffic over a direct connection between an on-prem data center and the cloud often must be segmented. VLAN configurations are used to manage these controls. You can configure VLANs on this CSP connection to manage access to various subnets within your VPC. For example, one VLAN might host traffic for an Internet-facing subnet in your VPC, while a different VLAN would host traffic for a private subnet in your VPC, thereby maintaining traffic segmentation over the interconnect. However, once you get traffic into the cloud, other technologies are used to blend or segment traffic and IP address spaces.

In the on-prem data center, VLANs play a significant role in segmenting the network for both security and efficiency purposes. While you have likely learned about VLANs in a networking class, let's briefly review how VLANs work.

A VLAN groups ports on a switch so that some of the local traffic on the switch is forced to go through a router to reach ports in a different VLAN, even if those ports are on the same switch. This allows a network admin to limit which network clients can communicate with other network clients without having to install a different switch for each group. VLAN information is added to each packet in a VLAN tag (Figure 5-4), which is inserted in the data link header of each message (recall that the data link layer is OSI layer 2). The VLAN tag identifies which VLAN the message belongs to, and that message can only be delivered to other switch ports assigned to the same VLAN.

Figure 5-4 The VLAN tag is inserted after the Source address field in an Ethernet frame

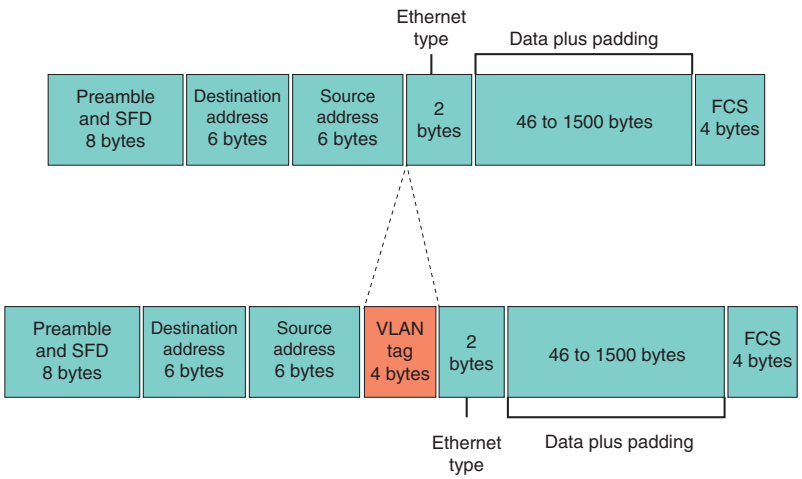
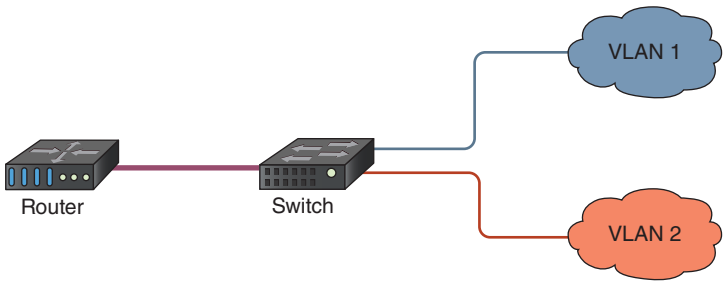


Figure 5-5 illustrates what happens when ports on a switch are partitioned into two VLANs. Traffic within each VLAN still goes through the switch as normal to reach other devices on the same VLAN. Traffic to hosts on other networks still goes through the router. However, traffic between hosts on VLAN 1 and VLAN 2 must now also go through the router, which is called inter-VLAN routing. This simple VLAN configuration, where one router connects to a switch that supports multiple VLANs, is sometimes called an ROAS ("router-on-a-stick").

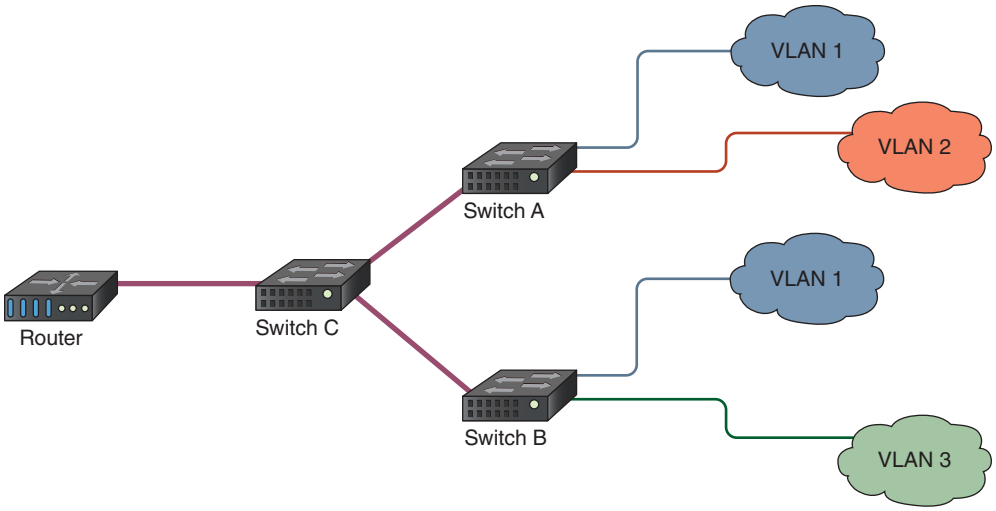
Figure 5-5 A switch with its ports partitioned into two groups, each belonging to a different VLAN



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As just described, a switch can support more than one VLAN. Similarly, a VLAN can include ports on more than one switch. Suppose you add a couple more switches to the LAN (local area network), as in Figure 5-6. In this example network, Switch B's ports can be configured with the same or different VLANs as the ports on Switch A. Traffic from one device on VLAN 1 connected to Switch A can travel to another device on VLAN 1 connected to Switch B and be treated as local traffic (that is, using private IP addresses). However, devices on separate VLANs—even if they're connected to the same switch—can't talk to each other without going through the router. Therefore, transmissions from a device on VLAN 1 connected to Switch B must go through the router to reach a device on VLAN 3, even though both devices are plugged into the same switch.

Figure 5-6 Three switches on a LAN with multiple VLANs



Exam Tip | The CompTIA Cloud+ exam likely will not expect you to configure or manage VLANs on a network. However, you might be given a scenario that requires you to identify symptoms related to a misconfigured VLAN. Having a basic understanding of how VLANs work and how to deploy them on-prem will help you to recognize these symptoms.

Applying Concepts 5-1: Testing Cloud Latency

Estimated time: 10 minutes
Objective 1.10: Compare and contrast methods for optimizing workloads using cloud resources.
Resource:

- Internet access

Description: When selecting a region to host your cloud resources and to connect your on-prem data center to, you'll need to check latency metrics to determine which location might give you the best performance. It might not be the closest region because latency depends on so many other factors, such as the quality of infrastructure spanning the routes to these regions. A latency testing tool can give you helpful insight into what kind of performance you can expect from different cloud locations for yourself and for your customers.

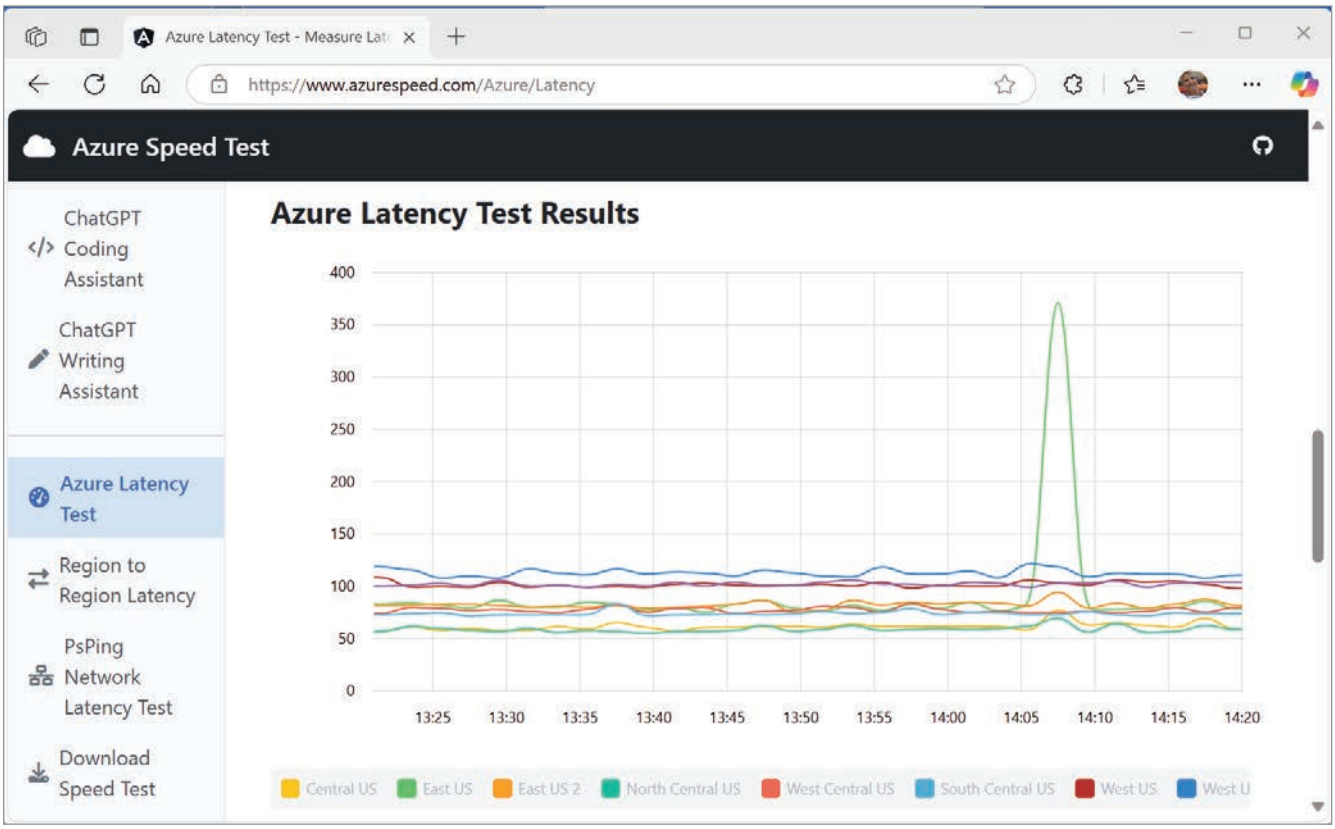
Complete the following steps:

1. Select a public cloud platform to research, such as AWS, Azure, or GCP. Search online for a latency tool that can compare latencies to various regions for that platform, such as Azure Speed Test (azurespeed.com) to assess latency to Azure. Notice that these sites typically are not developed or managed by the CSP. Consider searching for site reviews before selecting a site, to ensure your safety. What site did you choose?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. Most of these latency tools have you select a number of regions to perform a latency test on your various options. For example, Figure 5-7 shows latency scores graphed and compared for all Azure US regions. Select several or all the regions in your area and run the latency test for at least a few minutes. Which of these regions consistently yields the best (lowest) latency measurements?

Figure 5-7 Latency scores to Azure US regions from a location in the SouthEast



Source: blrchen/AzureSpeed.com

3. Select a few regions around the world from you. For example, you might select a region across one ocean from you, a region on the other side of the world from you, and a region in a different hemisphere (across the equator) but not across an ocean from you. Which of these regions yields the worst (highest) latency measurements?
4. Explore any other testing tools available on the same site. What other information are you able to gather about different regions for your selected cloud platform?
5. Explore available information on this site about how the site works. For example, the Azure Speed Test site provides a link to its GitHub repository. Site users can report feedback or bugs by creating a GitHub issue, and developers are invited to contribute by creating a code branch on GitHub and submitting a pull request.

Connecting Cloud Networks

The tools you've learned about so far—VPNs and direct connections—are designed to connect on-prem networks with cloud networks. Although it's sometimes possible to use these tools to connect cloud networks to each other, this is not common. Instead, cheaper and more efficient tools native to the cloud environment can connect cloud networks to each other, even across regions and around the world. Some of these tools can also be used in conjunction with VPNs and direct connections to expand the functionality of these resources.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Cloud Peering

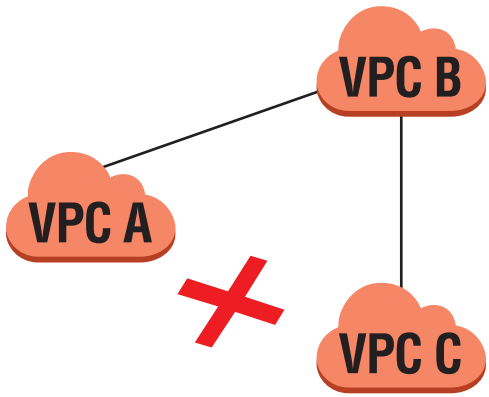
In the context of cloud computing, cloud peering refers to connecting multiple VPCs to each other in such a way that ensures traffic between the networks can communicate via private IP addresses (as opposed to requiring public IP addresses). This is similar in concept to network peering used between ISPs, where peering organizations are equal parties with equal needs for the network peering connection. Traditional ISP network peering creates a "settlement-free" relationship where neither party pays the other (unlike a direct connection, which is provided by a CSP and paid for by a customer).

Similarly, cloud peering allows for fast and secure connections between an organization's VPCs, or between VPCs owned by different organizations. This configuration provides lower latency, simpler configuration, and higher security than what could be achieved by connecting VPCs over the Internet or across a VPN connection. Typically, peering features are built into the major IaaS platforms—such as AWS, Azure, and GCP—and involve creating a simple connection that bridges the gap between each virtual network.

Cloud peering gives instances within each VPC access to instances in the peered VPC. In effect, you're blending the routes available to instances in each VPC. Because routes rely on distinct CIDR ranges, the CIDR blocks of peered VPCs cannot overlap. This is an important restriction of peering that must be considered when first assigning CIDR blocks to VPCs. Most cloud architects take this limitation into account from the very beginning of their design process.

Another common concern with VPC peering is the question of whether communications can cross multiple peering connections, which is called **transitory peering**. In the interest of security, a cloud peering relationship is typically not transitory, although there are exceptions. Consider Figure 5-8. Imagine that VPC A is peered to VPC B, and VPC B is also peered to VPC C. Because this peering is not transitory, VPC A is not automatically peered with VPC C. You would need a direct peering relationship between VPC A and VPC C for resources in those two VPCs to communicate directly. In some cloud platforms, however, you do have the option to enable transitory connectivity.

Figure 5-8 VPC A's peer connection with VPC B does not give VPC A access to VPC C's instances

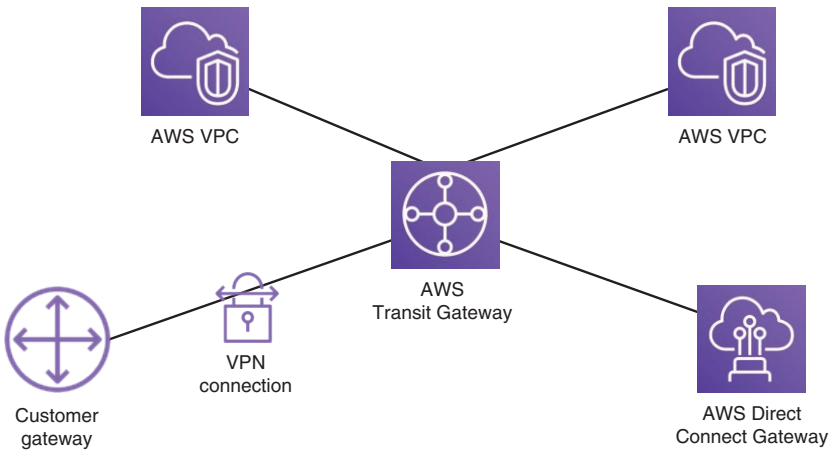


Transit Gateways

Peering arrangements are not limited to connecting two VPCs with each other. However, consider the complexity peering would require if you wanted to connect a few dozen VPCs across multiple cloud accounts along with multiple on-prem data centers and perhaps with remote users. To do all this, it's more effective and efficient to use an infrastructure designed to support this complexity.

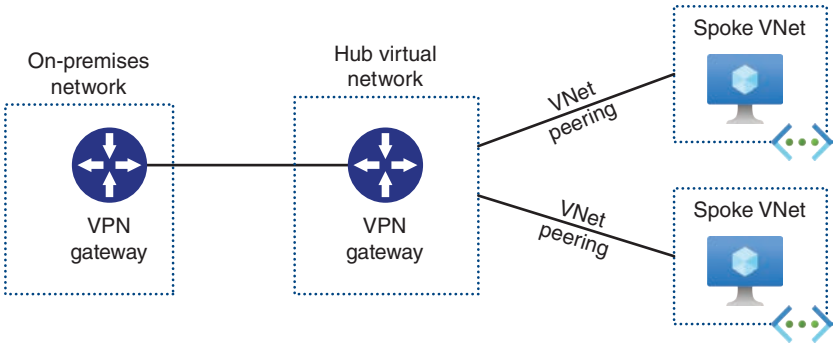
In many cloud platforms, you can deploy a central device (such as AWS's Transit Gateway) or a central network (such as Azure's hub VNet) to connect many VPCs to each other. This architecture is called a **hub-and-spoke topology**, or just hub-spoke (Figure 5-9 and Figure 5-10).

Figure 5-9 Hub-and-spoke VPC peering architecture using AWS Transit Gateway



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-10 Hub-spoke VPC peering architecture using an Azure hub virtual network



Note 5-2 | As with VPC peering, VPCs connected via a transit gateway typically cannot have overlapping CIDR blocks.

The hub in each case can be used to connect large numbers of VPCs in the cloud to each other as well as connect on-prem networks or individual users via VPNs or direct connects to resources contained within not just one, but many VPCs. The hub at the center of the topology contains a **transit gateway**, which acts as a cloud router directing traffic across all connected networks, whether those locations are on-prem or in the cloud. The cloud customer determines which of the connected networks can communicate with each other. This capability allows for security-driven segmentation.

Transit gateways cannot connect directly to all cloud resources because not all cloud resources are hosted within a VPC or VNet. For example, AWS's S3 (Simple Storage Service) allows customers to create storage buckets to store data; however, these buckets are not associated with a VPC-defined network space. Instead, they reside outside of the customer's VPCs with very little network context or configuration that the customer can access or manage.

In other words, an S3 bucket typically has no private IP address and no need to know about DNS (Domain Name System) servers or default gateways. All of this network configuration is handled behind the scenes by the cloud provider and is not accessible to the cloud customer even to manage network connections or direct traffic. Instead, a different, cloud-native type of configuration resource gives cloud customers options for private access.

For example, a VM instance inside a VPC can access an S3 storage bucket if its traffic crosses an Internet gateway and public Internet infrastructure to the bucket's public IP address. Alternatively, the instance's VPC can be given private access to S3 via a special interface endpoint that keeps this traffic off the open Internet. This interface endpoint is like a secret passage between the VPC and the S3 bucket. Traffic across this passage is not exposed to the open Internet, but it still doesn't rely on networking concepts such as VPNs or even cloud peering.

The same restriction applies to a transit gateway and any devices using the transit gateway to access cloud-hosted resources. There must be an interface endpoint in a VPC that is connected to the transit gateway, and this interface endpoint, in turn, provides private access to the storage bucket.

For this course, it's not so important that you thoroughly understand these more complex network configurations in the cloud. However, you should be aware that access to some cloud resources does not rely on the traditional concepts of networking that apply in VPCs.

Cloud WAN

Even a transit gateway has capacity limitations, especially when supporting a global network of connected locations. For this reason, many public cloud platforms are now offering a newer service type called something like Cloud WAN or Virtual WAN. These services provide a centralized platform for managing a large

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

number of connections in widely dispersed geographical areas, such as across continents and regions. All of the typical cloud connection types are supported, including VPNs and private connections. The cloud WAN or virtual WAN also relies on a hub-and-spoke topology, similar to a transit gateway. However, the functionality and features are designed to support large numbers of connections and connection types more effectively throughout available global regions.

Self-Check

1. Which of the following technologies represents a physical device in the customer's on-prem network?
   - **a.** VGW
   - **b.** CGW
   - **c.** Cloud WAN
   - **d.** VLAN
2. Which cloud connection type is the most expensive for the cloud customer?
   - **a.** Public Internet
   - **b.** Direct connection
   - **c.** VLAN
   - **d.** VPN
3. Which of the following characteristics is increased by a transit gateway?
   - **a.** Complexity
   - **b.** Latency
   - **c.** Connectivity
   - **d.** Throughput
4. Which of the following factors directly affects the cost of a direct connection?
   - **a.** Bandwidth
   - **b.** Routing
   - **c.** Segmentation
   - **d.** Incoming data

Check your answers at the end of this chapter.

Section 5-2: Extending Network Services

Certification

1.3 Explain cloud networking concepts.

1.7 Compare and contrast virtualization concepts.

2.5 Given a set of requirements, provision the appropriate cloud resources.

6.2 Given a scenario, troubleshoot network issues.

When you extend a network's infrastructure to the cloud, several network and security-related services are involved in bridging the gap between on-prem and cloud-based resources. Making the connection between the two locations, which you just learned about, is only the first step. For example, you've already learned how IP address spaces cannot overlap across a VPN, interconnect, or peered connection. Therefore, it's important to manage your address spaces efficiently. This section begins with a discussion of DHCP (Dynamic Host Configuration Protocol) services across hybrid and multi-clouds. The section also covers DNS services, routing, and load balancing across platforms.

DHCP Services

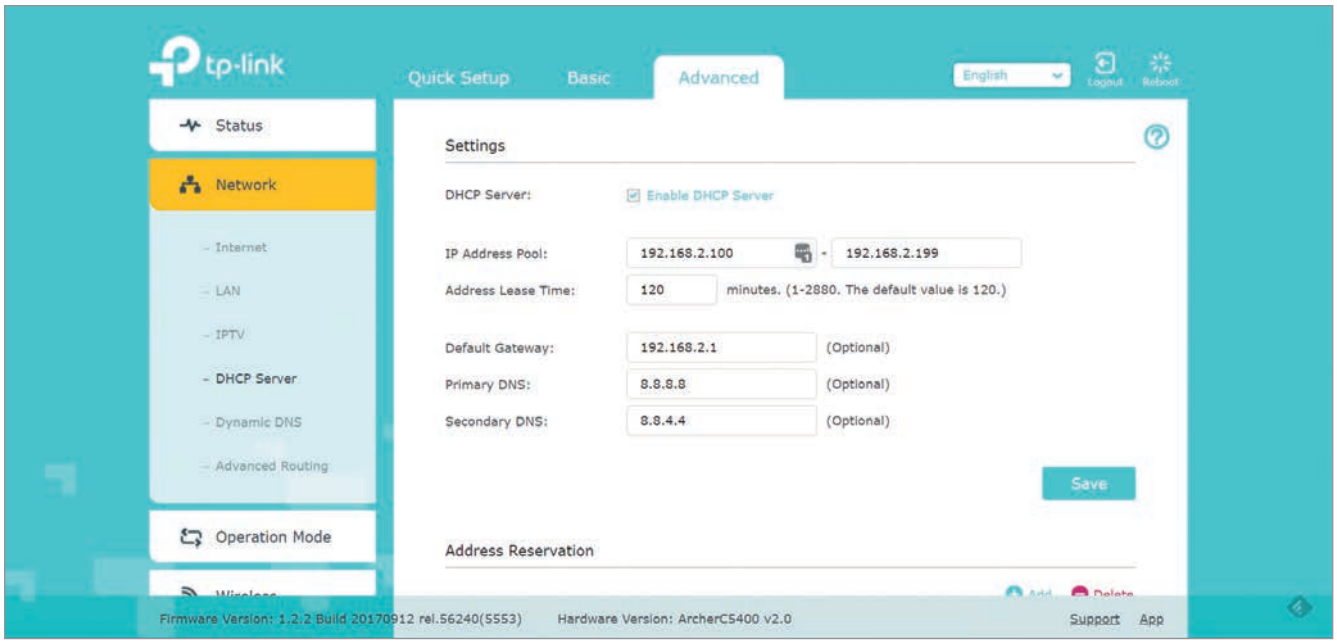
Consider where a computer on a traditional network gets its IP address. A network administrator manually assigns a static IP address, while a dynamic IP address is assigned by the DHCP server each time a device connects to the network. To enable DHCP, a network admin configures specific IP address pools from which

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the DHCP server can assign addresses to clients (Figure 5-11). Along with the IP address comes additional information called scope options, which include the following:

- A time limit called a lease time, after which the DHCP client must request an IP address renewal from the server
- The IP address of the network's router interface, which is called the default gateway
- Primary and secondary DNS server addresses, which the DHCP client can use to resolve IP addresses

Figure 5-11 DHCP options on a small network router



Source: TP-Link Technologies Co., Ltd.

In addition to allocating one or more address pools for DHCP to use, the network admin might also configure reserved or excluded addresses. For example, a network printer might have a static IP address. If that IP address is included in the range of addresses in the DHCP pool, then the IP address should be listed as an exclusion so it won't be used for another client. Alternatively, the network admin might configure an IP reservation based on the printer's MAC address so the DHCP server reserves this address only for the printer. The server then assigns the printer this reserved address any time the printer connects to the network.

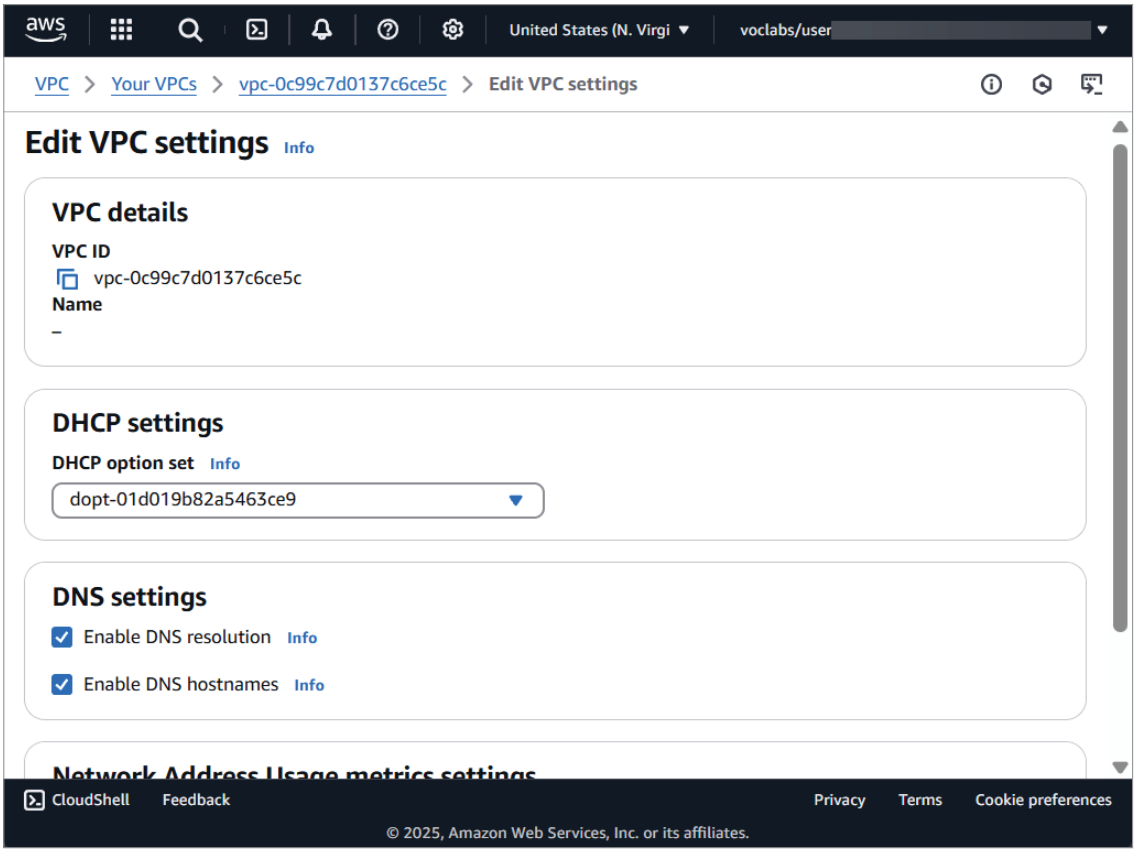
DHCP in the Cloud

DHCP in the cloud accomplishes many of the same tasks but not necessarily from a server instance. For example, consider how instances in AWS receive dynamic IP addresses. When you create a VPC, you define an address space from which addresses are allocated to new VM interfaces in the VPC. This address allocation process happens in the background. Other than setting the CIDR block, you don't need to do anything else to ensure that your instances receive a private IP address along with all the other information they need to communicate with other resources in the VPC. By design, each interface keeps its private IP address indefinitely until the interface is deleted (although this is not necessarily true in other cloud platforms).

If desired, however, you can change some of the DHCP options that a VM instance receives when its interface is created. This might be important, for example, if you need to use your own DNS servers instead of relying on the cloud provider's default DNS servers. In AWS, changing DHCP options is accomplished by creating a new DHCP options set from the VPC dashboard and then editing the options set associated with an existing VPC (Figure 5-12). If instances already exist using a previous options set, you must restart the instances for the new settings to take effect. Figure 5-13 shows the options you can configure in the DHCP options set.

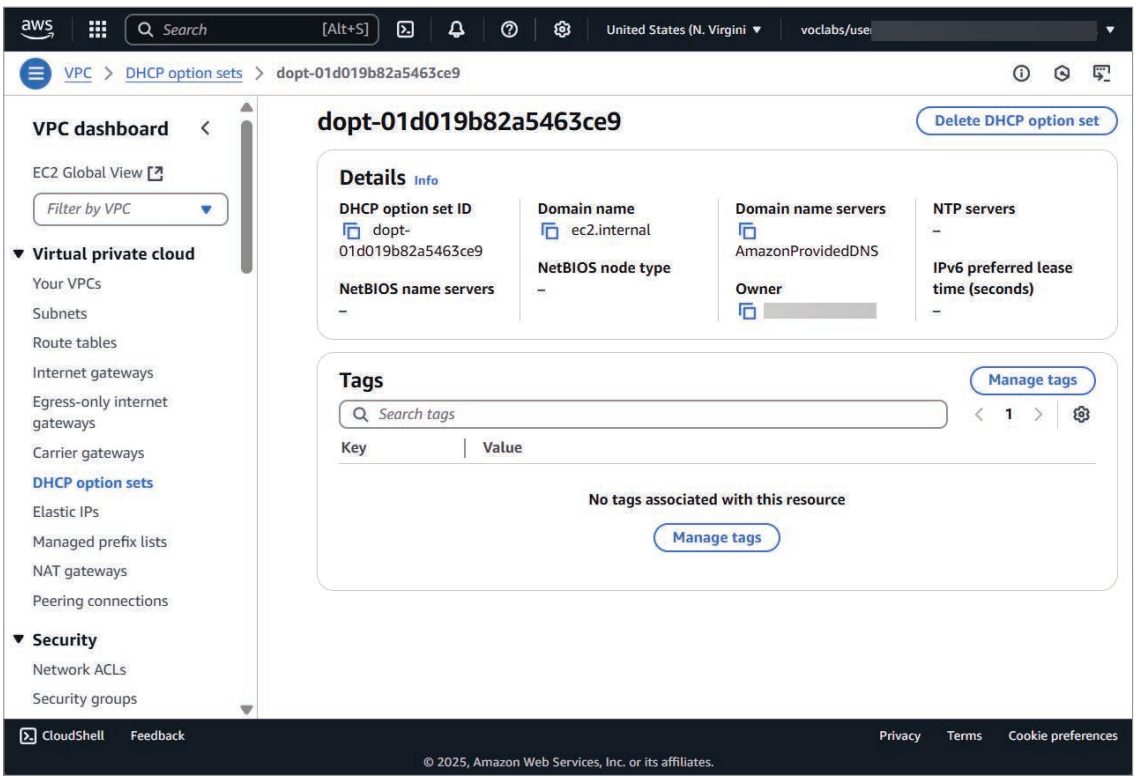
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-12 A DHCP options set associated with a VPC



Source: Amazon Web Services, Inc.

Figure 5-13 Set one or more of these options

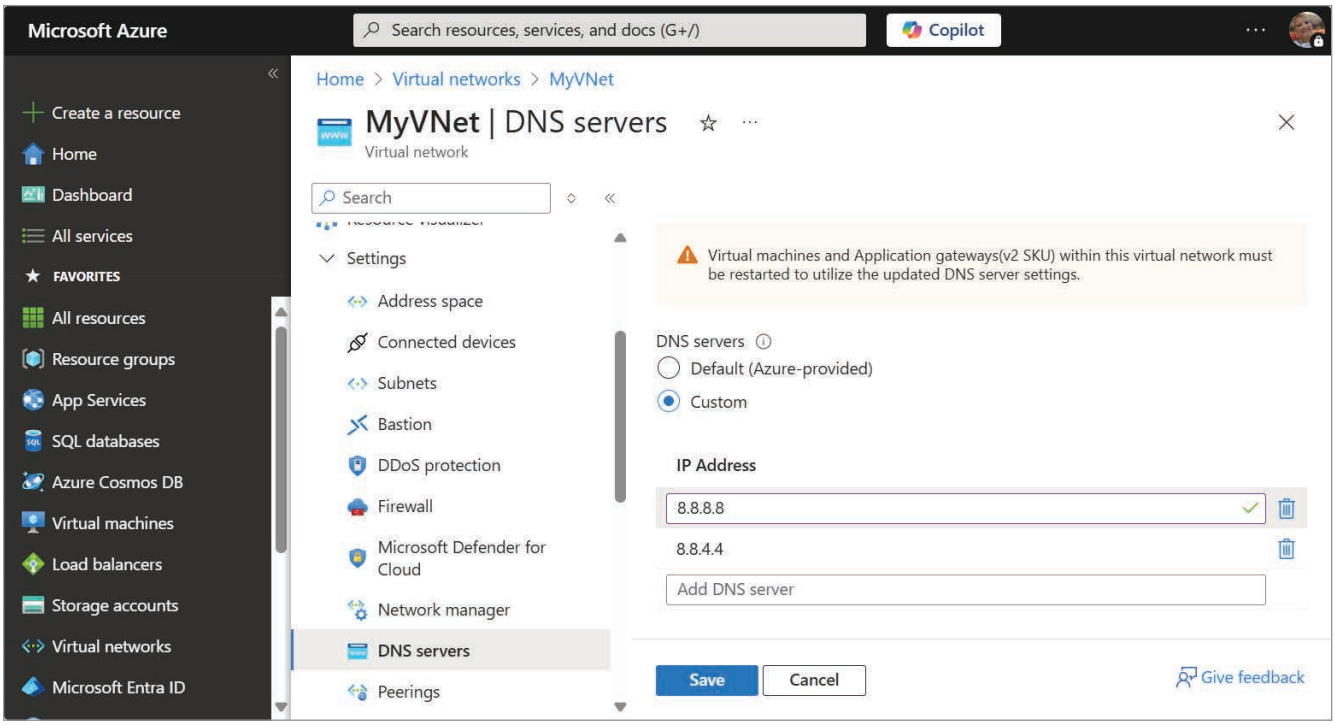


Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Similar to AWS, you can choose the address space for a VNet in Azure. You can also change the DNS servers the VNet's instances use. To change DNS servers, go to the VNet's page (Figure 5-14), click DNS servers in the left pane, and change the setting to Custom. Then enter the DNS server addresses you want to use.

Figure 5-14 Choose custom DNS server addresses for instances in an Azure VNet



Source: Microsoft

In GCP, you can set IP address ranges when creating a VPC and its subnets. You can also reserve static IP addresses as needed. Although it's possible to set a VPC's DNS server addresses with a DNS server policy, at the time of this writing, these steps cannot be performed in the console, only through shell commands or scripting.

DHCP Across a Hybrid Cloud

Now that you've figured out how to adjust your cloud's DHCP options, the next question is, how can you configure DHCP across a hybrid cloud so that one system manages DHCP services both on-prem and in the cloud? The solution is IPAM (IP address management), which is a service that provides a way to plan, deploy, and monitor a network's IP address space. IPAM tools can automatically detect and manage IP address ranges, assignments, reservations, and exclusions; integrate this information with data from DNS records; and provide constant monitoring for growth, security, and troubleshooting purposes.

AWS offers an IPAM service that can manage IP pools. A public scope handles public IP addresses separately from private IP addresses in a private scope. The free tier works for a single account and a single region, and the advanced tier works across regions and accounts. However, at the time of this writing, neither of these options provides IPAM services into the on-prem environment.

For now, IPAM functionality is not built into public cloud platforms like AWS and Azure for hybrid deployments other than being able to reserve address spaces for on-prem use, which prevents overlapping of address spaces. Instead, third-party, software-defined IPAM solutions, such as Infoblox and FusionLayer, are designed to integrate with popular cloud platforms and on-prem network resources. These tools can provide a unified interface for managing address spaces across hybrid and multi-clouds.

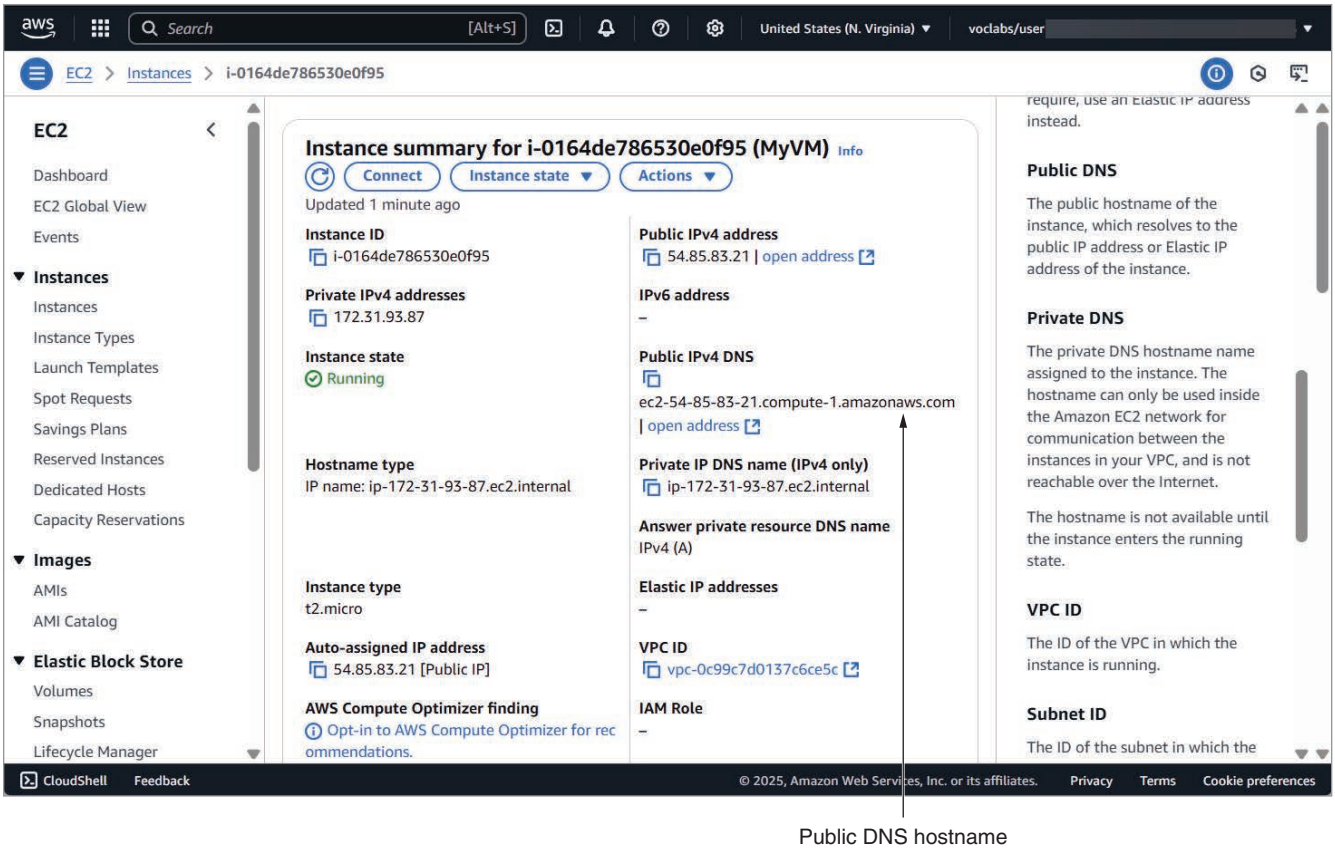
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DNS Services

You read earlier that DHCP scope options can be used to define a virtual network's DNS servers. Why is this an important capability? And how does this feature apply to a hybrid or multi-cloud environment? A quick review of what DNS does will help answer these questions.

You might have noticed that VM instances are typically assigned a host name in the cloud. For example, Figure 5-15 shows an example of an instance in AWS with its auto-generated public and private host names along with their corresponding IP addresses.

Figure 5-15 Each host name corresponds to an IP address



Source: Amazon Web Services, Inc.

Host names, such as the ones in Figure 5-15, and domain names, such as google.com, were created because character-based names are easier for humans to remember than numeric IP addresses. While domain names are convenient for humans, a computer must convert the domain name to an IP address before it can find the referenced machine. This is where **DNS (Domain Name System** or **Domain Name Service)** comes in. Through a process called name resolution, DNS associates computer names with IP addresses.

Each organization that provides host services (such as websites or email) on the public Internet is responsible for providing and maintaining its own authoritative DNS servers. An authoritative server is the authority on computer names and their IP addresses for computers in their domains. An organization also uses DNS servers for network clients, which need to access public DNS data for resolution of domain names across the Internet.

When a computer requests name resolution of a particular domain, the computer first reaches out to its designated DNS server, whether that's a private DNS server managed locally by the organization, or a public DNS service offered by an ISP or organization like Google. For example, Google's public DNS server addresses are 8.8.8.8 and 8.8.4.4. If that DNS server already has the requested information cached, it will return that information to the DNS client. Otherwise, the DNS server will query other DNS servers that function at higher

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

layers of the name server hierarchy until it locates the authoritative name server for the requested domain name. The authoritative name server can then report the associated IP address for the domain. The entire process might take a few seconds, at most.

DNS servers collect lists of computer names and their associated IP addresses in namespace databases. These databases are stored on thousands of servers around the world rather than being centralized on a single server or group of servers. In other words, DNS doesn't follow a centralized database model but a distributed database model across a hierarchy of servers. Because data is distributed over thousands of servers, DNS will not fail catastrophically if one or a handful of servers experiences errors.

DNS in the Cloud

Major public cloud platforms offer DNS services for resources hosted in their cloud. For example, by default, all AWS VPCs are configured to use Amazon DNS servers. AWS also offers its more sophisticated Route 53 DNS service (DNS's default port is 53, thus the name). Organizations can use the highly available and scalable Route 53 as its authoritative DNS service for domains that it owns.

Alternatively, you can set the VPC's instances to use your own DNS servers or public DNS servers by creating a new DHCP options set, as described earlier. Many organizations run their own DNS servers to have more control over their network's traffic and to provide name resolution for internal network traffic.

DNS Across a Hybrid Cloud

AWS's Route 53 service can support a hybrid cloud as well. An organization's on-prem resources can access cloud-based DNS services through a VPN or Direct Connect connection. This enables fully integrated DNS across the entire hybrid cloud without requiring the use of a third-party service provider. Azure offers similar functionality by running Active Directory on Azure-hosted VMs. In GCP, you can use Cloud DNS forwarding capability for hybrid cloud DNS management.

> **Note 5-3** | If you'd like more information about DNS and how DNS servers work, use your Cengage Unlimited subscription to go to *CompTIA Network+ Guide to Networks*, 10th edition, Chapter 3, and read the section titled "Domain Names and DNS."
>
> If you don't have a Cengage Unlimited subscription, you can find more information at cengage.com/unlimited.

Routing

You've already learned how route tables are used to manage a virtual network's access to the Internet. When you create connections between on-prem and cloud networks, routing plays a key role in determining how traffic flows between resources across the connection. However, establishing and maintaining these routes works a little differently in the cloud than it does for a traditional, on-prem network.

You might have studied routing protocols in a networking class. These routing protocols are categorized into two major groups: interior and exterior gateway protocols. Interior gateway protocols propagate routes within an organization's domain, while exterior gateway protocols communicate routes between organizations. CSPs and other large organizations rely on one or more interior gateway protocols, such as RIPv2 (Routing Information Protocol, version 2), OSPF (Open Shortest Path First), and EIGRP (Enhanced Interior Gateway Routing Protocol). However, you typically won't manage these protocols directly for configuring public cloud services. The only routing protocol you'll work with in the cloud is **BGP (Border Gateway Protocol)**, which is the only modern exterior gateway protocol.

VPN Routing

Routing for VPNs can be configured using static routes set by the cloud admin (which do not automatically adapt to route outages) or using dynamic routing, where routes are configured automatically by the cloud platform. When configuring traditional VPNs, you often have many routing protocol options to choose from.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

However, in the cloud, you are most often required to use BGP. BGP dynamically manages routes between organizations, such as across the Internet. A router using BGP calculates the best path between two locations and accumulates this information in its route table. If congestion or failures affect the network, a router can detect the problems and reroute messages through a different path. To use BGP, it must be supported by the device (physical or virtual) on each end of the connection.

When configuring a VPN, you also have the option to enable route propagation between connected networks. For example, if your VPC is connected by VPN to an on-prem network, route propagation will allow the VPC to learn all the routes that the on-prem network already knows. This can be helpful in some circumstances and problematic in others.

Direct Connect Routing

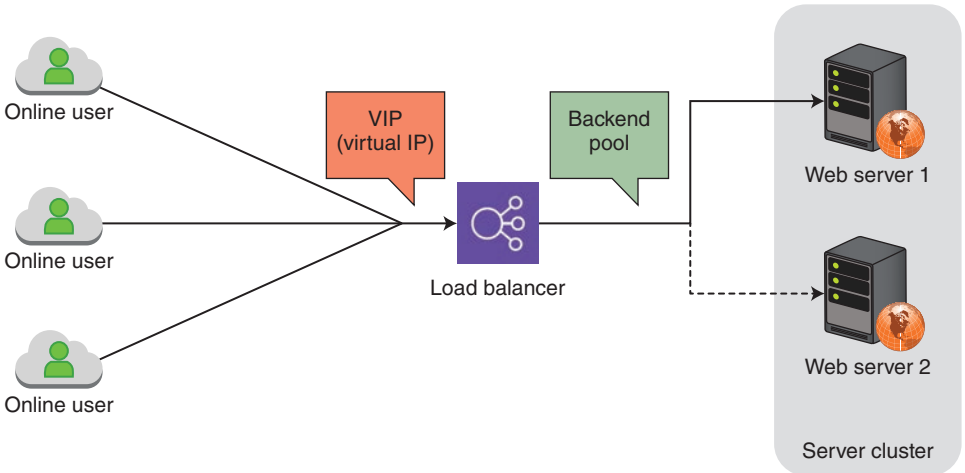
Direct connection services require BGP routing on each virtual interface that is part of the connection. A virtual interface is basically a subinterface on the connection that allows different connections to access various cloud services. This feature functions similarly to VRF (virtual routing and forwarding), which is a technology that allows a single router to maintain multiple route tables instead of only one. A single direct connection can have many virtual interfaces, and each one has its own VLAN. Recall that VLANs are used to segment a network into smaller subnets. The direct connection between an on-prem data center and the cloud is itself a small network, and each subinterface represents a smaller portion of that network. Segmenting a direct connection in this way allows a single connection to be used for many purposes, which provides significant cost savings for the cloud customer.

Load Balancing

Routed cloud traffic doesn't always go directly to a single destination. It's common for organizations to establish a cluster of multiple servers that take turns performing essentially the same task. This way, the cluster can support much larger volumes of traffic, and a failure on one server won't disrupt the service itself. For example, consider a typical web server setup for a popular website. A single web server can't possibly handle all the website's traffic. The company must run multiple web servers to spread the demand of traffic across several machines. However, DNS still might point to only a single IP address. How does this work? The secret ingredient is a load balancer.

Load balancing is the distribution of traffic over multiple components to optimize performance and fault tolerance. Basically, the load balancer stands between incoming Internet traffic and a cluster of web servers. It distributes incoming traffic to all the servers in the cluster according to rules defined by the administrator. Examine Figure 5-16, which illustrates Internet traffic on the left (called the frontend) and the organization's web server cluster on the right (called the backend).

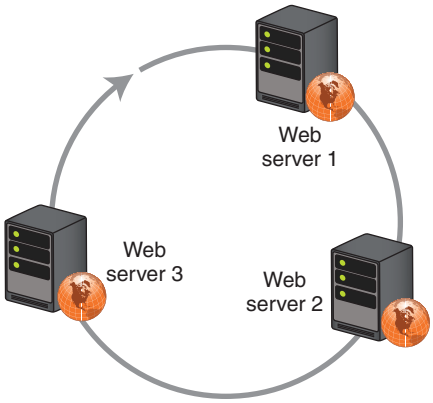
Figure 5-16 A load balancer distributes traffic to multiple destinations



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The public IP of the website, called a VIP (virtual IP), represents the entire cluster and directs traffic to the load balancer, not directly to any of the web servers. The load balancer uses NAT (network address translation) to forward traffic to one web server at a time. Each web server is part of the cluster identified by a backend pool, which is a list of private IP addresses belonging to the servers.

Figure 5-17 With a round-robin algorithm, each server takes a turn in a set order



Load-Balancing Algorithms

The load balancer chooses a server according to one or more load-sharing algorithms. For example, the load balancer might use a round-robin rule (Figure 5-17), where each server in the cluster gets a turn in a particular order. However, this method does not take into consideration how much work a server is already doing. In this scenario, web server 1 might receive a large, time-consuming request. Then web servers 2 and 3 each get a quick and easy request and are finished with their tasks long before web server 1 has completed its task. In the meantime, the next request comes in. Using a round-robin rule, the load balancer would send the newest request to web server 1 even if that server is still busy with the first request.

Other algorithms can be used to balance workloads more efficiently across a cluster. For example, the load balancer might instead send the next request to the least busy server. Collecting information about the servers' activity levels and health might be achieved by installing on each server a small piece of monitoring software called an agent, or the load balancer might periodically probe the server for health information. A server that is experiencing problems will be removed from the load balancer's queue until that server is up and running again. More complex algorithms might be handled though a PAC (proxy automatic configuration or proxy auto config) file so traffic is balanced according to various characteristics of the traffic itself.

Sticky Sessions

Load balancers can also maintain session information so a client communicating with a server can continue with the same server until the session is complete. For example, consider an online shopper who has added a few items to the website's shopping cart. The live data of which items were placed in the cart resides on the initial server. If the session is suddenly transferred to a different server by a load balancer, that server wouldn't have access to this data because the order has not been completed—this information doesn't yet exist in any official records.

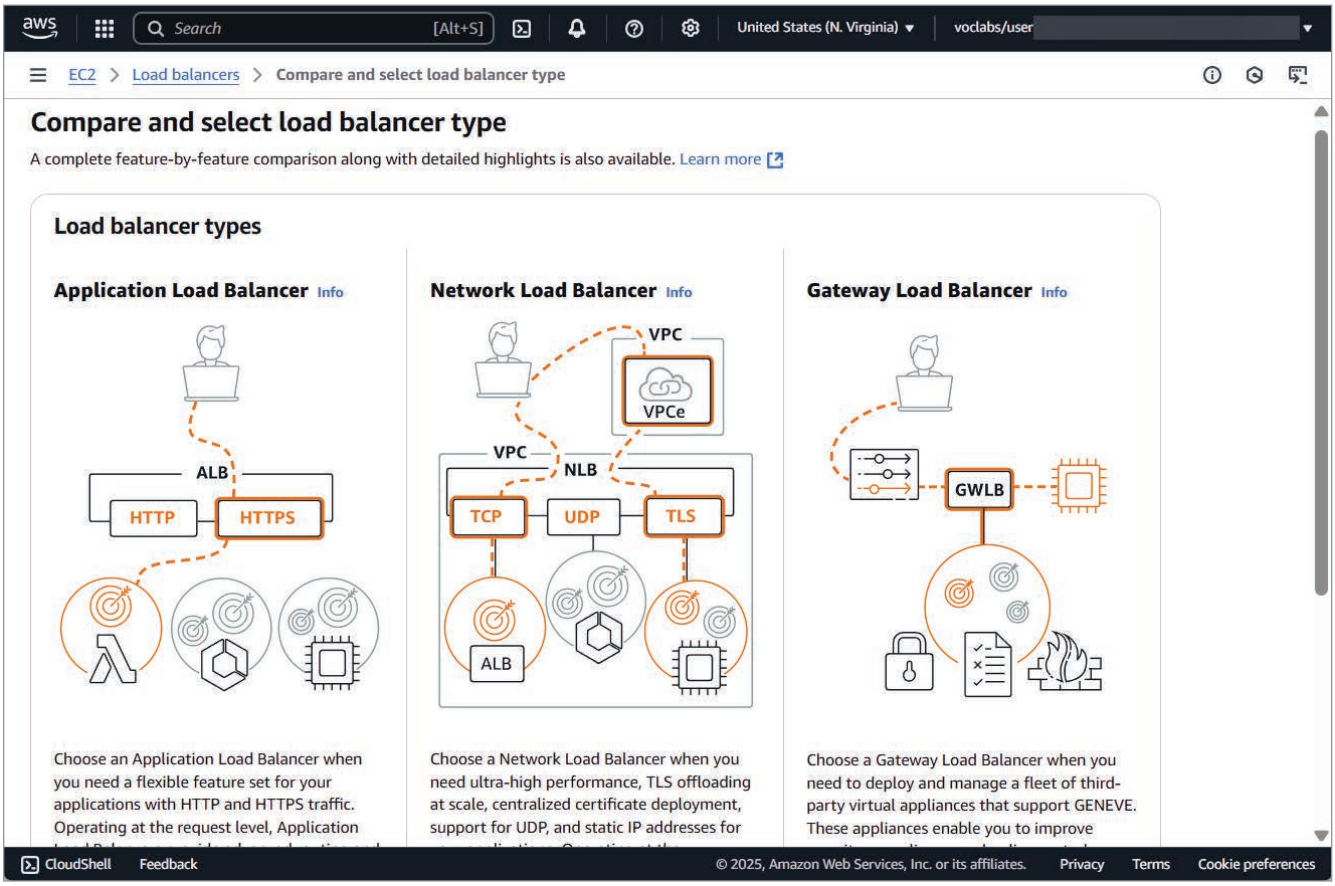
It's the responsibility of the load balancer to ensure the shopper's session continues to be handled by a single server so temporary data (such as items in a shopping cart) isn't lost. The load balancer might use PAT (port address translation) to accomplish this goal. When the session is first created, it's assigned to a port on a single server. Further communications between the web client and the web server are tracked by this port number, ensuring that future messages are directed to the one server while the session is active. This is called a sticky session.

Load Balancing in the Cloud

The major public clouds like AWS, Azure, and GCP offer various load balancer services. These load balancers can handle traffic to and from the Internet or traffic internal to the organization's cloud. Specific load balancer services are typically differentiated by the OSI layer in which they function. For example, AWS, Azure, and GCP all offer application load balancers that sort traffic based on layer 7 information, such as HTTP (Hypertext Transfer Protocol) requests. In contrast, their network load balancers evaluate header information at layer 4 (transport layer), including TCP (Transmission Control Protocol), UDP (User Datagram Protocol), or TLS (Transport Layer Security) traffic. Of these two options, the network load balancer is the more robust service, capable of handling much higher volumes of traffic at much higher speeds (Figure 5-18). In AWS, load balancer options are accessed through the EC2 service. In Azure and GCP, you'll find load balancing under the Networking category.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-18 Load balancer services are classified by the OSI layer on which they function



Source: Amazon Web Services, Inc.

Note 5-4 | TCP is a connection-oriented protocol, meaning it makes a connection with the host on the other end, checks whether the data is received, and resends it if it is not. UDP is a connection-less protocol or best-effort protocol. It does not guarantee delivery and does not check to confirm if the data was received.

To configure load balancers, you need to understand the different strengths and weaknesses of each load balancer type and when to choose one type versus another. An **NLB (network load balancer)** uses network information contained in the transport layer of packets to distribute traffic among systems in a target group (Figure 5-19). For example, the NLB might direct traffic based on port numbers or protocols, such as TCP, UDP, or TLS. An NLB is simpler in its functionality in comparison to application load balancing and, as a result, handles higher volumes of traffic better. When the workload needs to support high volumes of traffic with high scalability and low latency, an NLB is usually the better choice.

An **ALB (application load balancer)** examines application-layer data to determine where to send incoming traffic (Figure 5-20). For example, your website might contain a product directory, a shopping cart, and the ability for customers to check out and make purchases. The ALB can distinguish between these various types of traffic, forwarding each request to a cluster of servers or a serverless service designed to handle each request type. However, because the ALB must search deeper into each packet to route the request, ALBs tend to be slower than NLBs and cannot handle the same volume of traffic. For example, in

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-19 An NLB distributes traffic based on transport-layer information

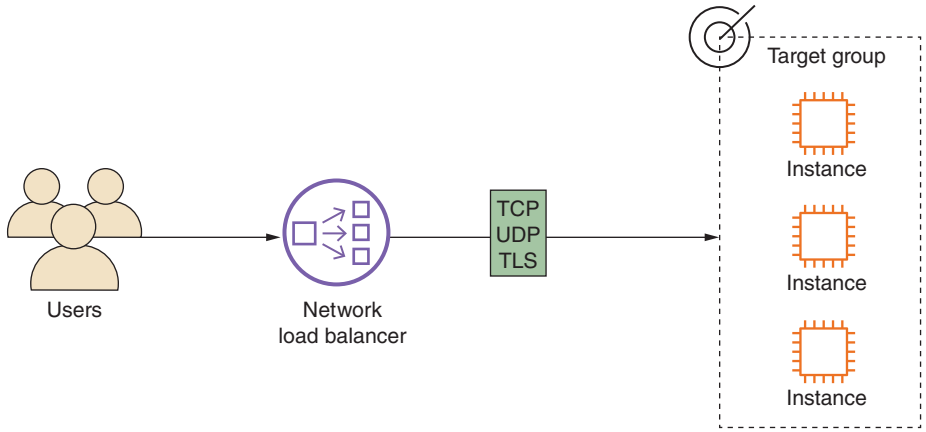
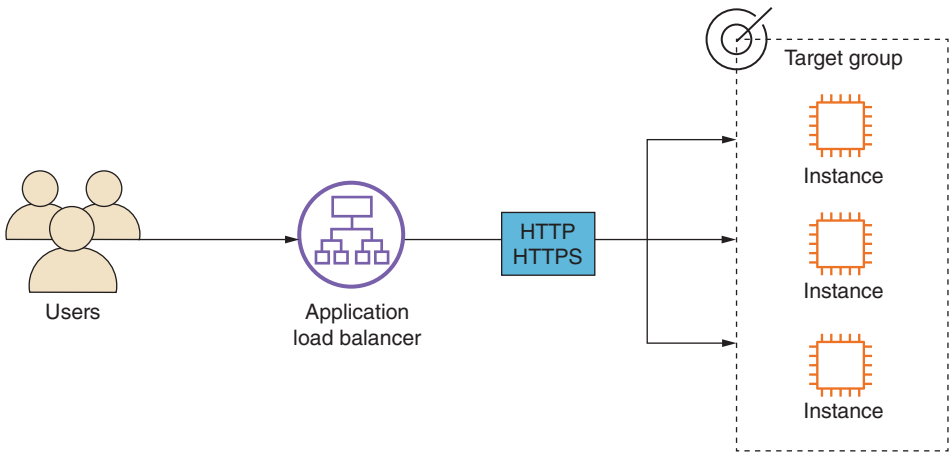


Figure 5-20 An ALB distributes traffic based on application-layer information



AWS, an ALB supports traffic that uses HTTP or HTTPS (HTTP Secure). An ALB is a good choice when you need to manage traffic based on application-layer data, such as when supporting microservices, containers, or web applications.

Each cloud platform's load-balancing services use different approaches to evaluating rules and are limited in different ways. When working with a specific load-balancing cloud service, it's important that you learn that particular service's limitations, features, and functionality.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 5-5 | In AWS, both an NLB and an ALB can route traffic to specific IP addresses or EC2 instances. An NLB can also route traffic to an ALB, which provides the opportunity for more complex and layered load-balancing configurations. An ALB can route traffic to a Lambda function, which is a serverless service in AWS, but an NLB cannot.

You don't need to know this level of detail for this course. However, when working with different types of load balancers in a cloud platform, make sure to explore the various configuration options available for each load-balancing service.

Note 5-6 | Azure has recently released its Application Gateway service, which offers application-level rule-based routing similar to AWS's ALB service. This is not the same thing as the application gateway that provides an alternative to VPNs for authenticated app access, as discussed earlier in this chapter, despite using the same term.

Load Balancing Across a Hybrid Cloud

Now that you understand the basics of how load balancing works, the next relevant question is, how do load balancers work across a hybrid or multi-cloud? As it turns out, CSPs have already thought about this. Most cloud-based load-balancing services can be extended to cover resources hosted on-prem and in other clouds, which is helpful for disaster recovery and business continuity planning. In fact, GSLB (global server load balancing), also called multi-site load balancing, can be configured to direct traffic to servers based on their proximity to users. This minimizes latency as users are accessing cloud-based resources distributed in various geographic areas.

For example, suppose you place an order for a book on Amazon. Amazon has warehouses distributed throughout the world. If you place an order for a book that is stored in multiple warehouses, the copy sent to you likely will be shipped from the warehouse closest to you, thereby minimizing transit time. This is similar in concept to what happens when a user wants to sign into their AWS cloud console. AWS data centers are located all around the world. Global load balancing, made possible by a partnership between AWS's load-balancing services and their Global Accelerator service, allows the user's traffic to be directed to servers hosted geographically near the user's own location. The user's traffic is directed to other locations only if needed to complete their requests.

Self-Check

5. Which of the following devices is least likely to need a static IP address?
   - **a.** Web server
   - **b.** Web client
   - **c.** DNS server
   - **d.** Default gateway
6. Which routing protocol is used across the Internet?
   - **a.** OSPF
   - **b.** BGP
   - **c.** RIPv2
   - **d.** EIGRP
7. Which device helps ensure HA for a web server?
   - **a.** DNS server
   - **b.** DHCP server
   - **c.** Router
   - **d.** Load balancer
8. Which type of device can load balance the highest volumes of traffic?
   - **a.** ALB
   - **b.** NLB
   - **c.** App gateway
   - **d.** Transit gateway

☐ Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 5-3: Extending the Cloud

Certification

1.2 Explain concepts related to service availability.

1.3 Explain cloud networking concepts.

1.10 Compare and contrast methods for optimizing workloads using cloud resources.

1.11 Identify evolving technologies in the cloud.

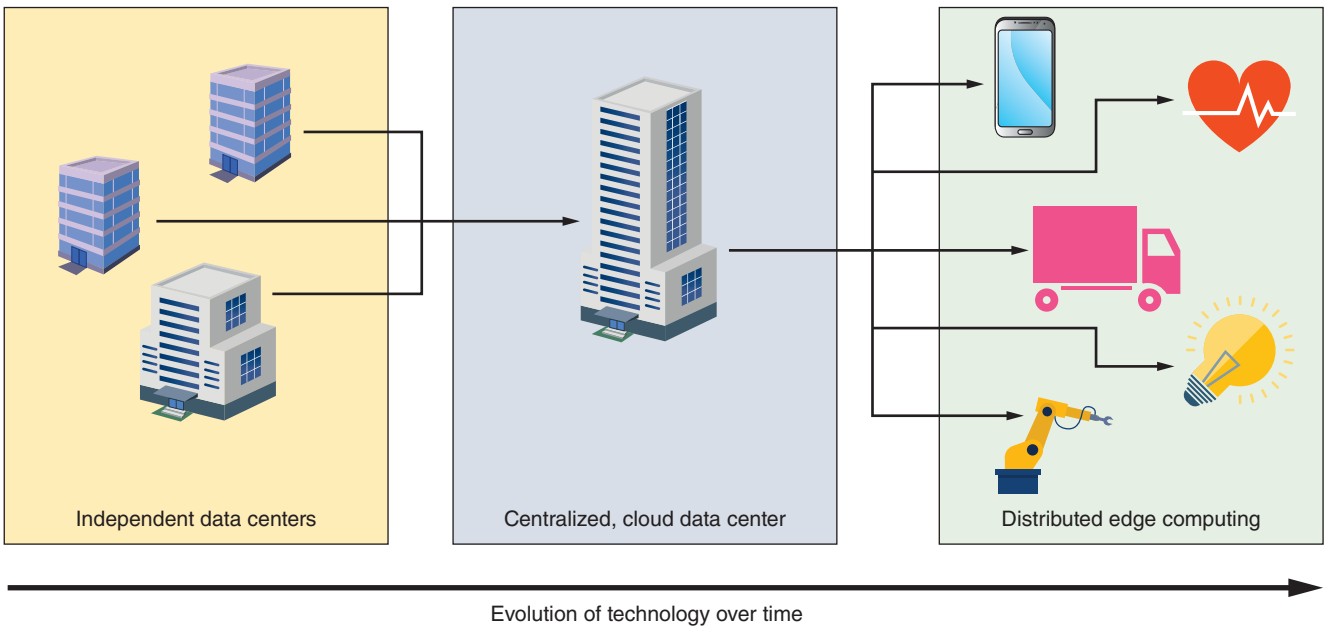
Edge computing is an evolution of cloud computing that distributes data processing closer to where data is used. For example, consider the latency involved with sending data to a central processing point, which could be located on a different continent, and then waiting again for the response to return to the point of use. If you're interacting with a retail website, perhaps this slight delay is not problematic. However, if you're driving a smart car that has just detected a pedestrian in its path, this kind of delay is unacceptable, even potentially fatal. Edge computing processes data at or near its point of use to nearly eliminate this delay.

Since the inception of cloud computing, the primary focus of cloud technologies has been to ubiquitously centralize compute resources away from smaller data centers to a few, massive data centers. One of the primary benefits of cloud computing is the economy of scale that offers lower prices on shared data center resources. Edge computing appears to be a shift in the other direction—bringing compute back to the point of use. If edge computing is, by nature, distributed instead of centralized, how is it still cloud computing? To make more sense of this seeming incongruity, let's dig a little deeper into the design of edge computing.

The Case for Edge Computing

To be clear, edge computing is not a shift back to smaller data centers. Instead, edge computing places compute power at the point of use. Consider that today's computing demands extend far beyond any data center. Data processing is needed in mobile devices, self-driving automobiles, industrial sensors, smart home devices, and implanted medical devices (Figure 5-21).

Figure 5-21 Edge computing shifts data processing to the points of use



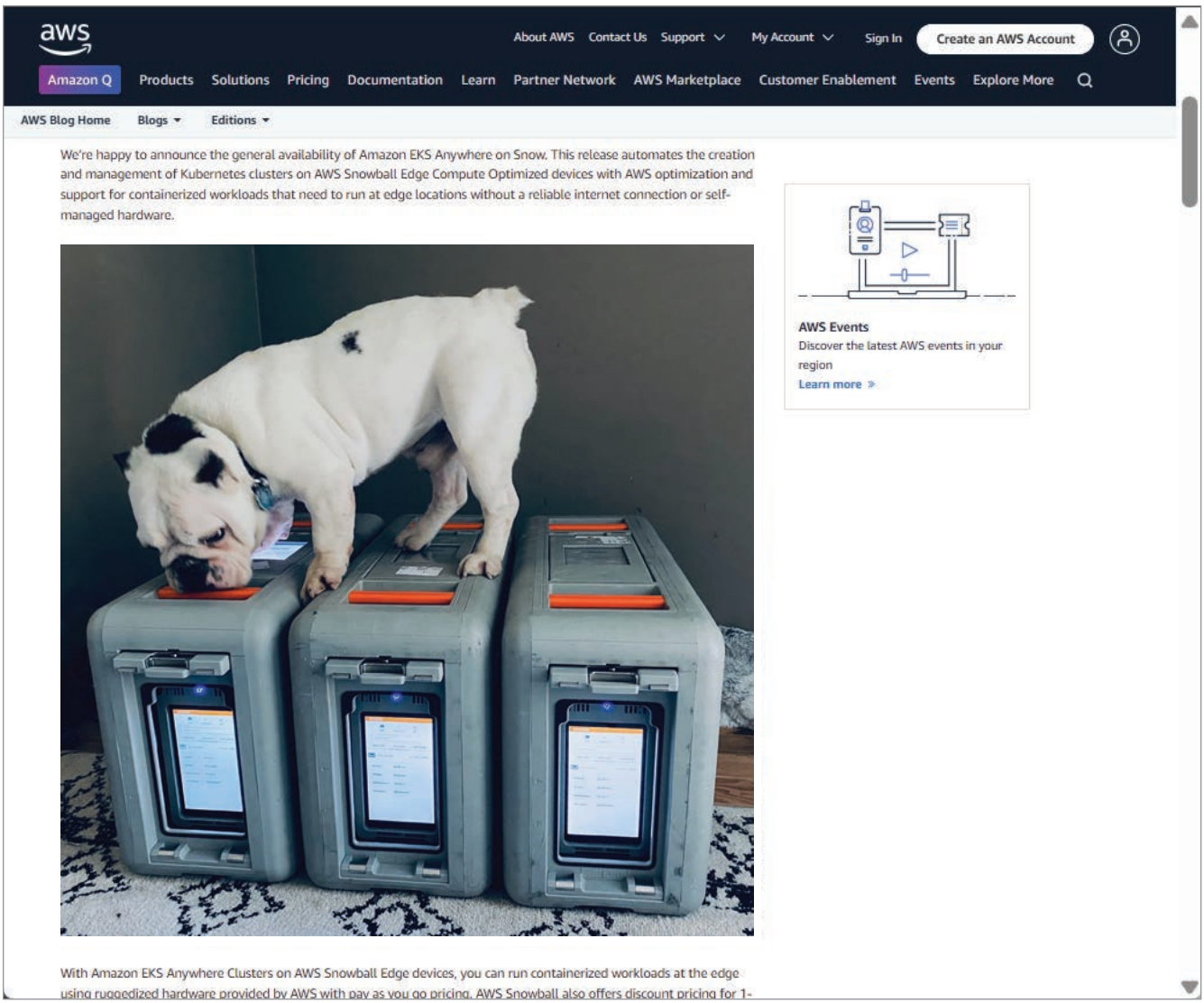
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In addition to reducing latency, edge computing increases resource availability. By processing data at its point of use, network bandwidth is freed for other traffic and less demand is placed on centralized resources. Data can still be copied to central locations for backup or additional processing.

Many CSPs offer a variety of edge computing services. For example, consider the following services from AWS to help you visualize the many ways cloud services can support edge technologies:

- Snowball Edge (Figure 5-22) is a self-contained data processing and storage device in a rugged case—about the size of a large suitcase—that can be set up in adverse environments. For example, Snowball Edge is used for military operations, on-site videography processing, and data transfer from remote data centers.
- The Local Zones service offers limited use of data centers owned by other organizations, such as ISPs, to AWS customers. This service improves customers' ability to meet data residency requirements and to further reduce latency.
- Networking services such as CloudFront and Global Accelerator keep AWS managed traffic off the public Internet as long as possible, enabling customers to take advantage of optimized AWS infrastructure for more of their traffic's journey. This strategy improves security, increases performance, and reduces latency.
- IoT Core is used to manage customer IoT (Internet of Things) devices, such as industrial systems or autonomous vehicles.

Figure 5-22 Ruggedized Snowball Edge devices



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CDN (Content Delivery Networks)

A popular form of edge computing is CDN. A **CDN (content delivery network)** is a distributed storage structure that allows customers to store files in locations closer to where their users are located. For example, suppose you run a video repository website. If your users in India try to stream a video that is stored on a server located in Virginia, those users will experience significant latency. A CDN such as AWS CloudFront allows you to cache a copy of your video at an edge location that is physically located in India. The proximity to your users will decrease latency and increase the quality of the user experience.

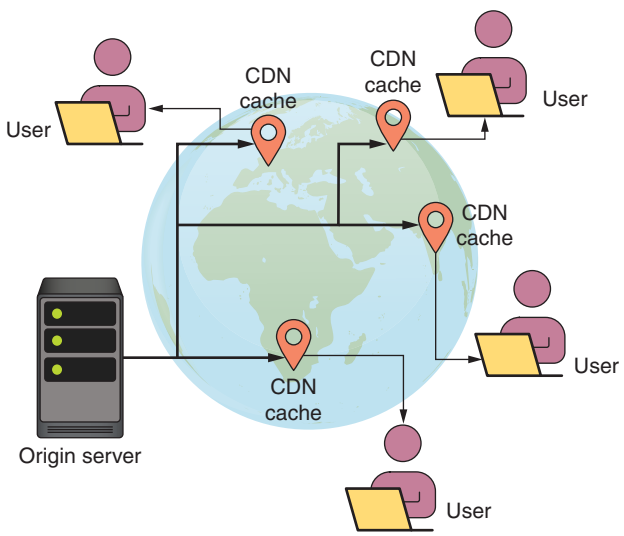
The CDN service routes user traffic to the closest edge location with the best performance at the time of the request. It also manages the TTL (time to live) of files cached at those edge locations according to parameters you define. These parameters are usually determined by factors such as the following:

- Resistance of the data to becoming stale (that is, becoming outdated over time)
- Sensitivity of the data to customization for the user (such as caching data used by many users but not caching data specific to only one user)

When cached files expire (that is, they have exceeded their TTL) or have not been used within a given period, they are discarded at the edge location. If the files are requested again from a user near that edge location, the CDN will again refer to the data's origin point to retrieve updated files.

As a general explanation, you set up a CDN distribution by starting with a central storage location for your files (Figure 5-23). You then configure a CDN access point, which might be able to perform authentication on file requests. Requests for the data are sent to this access point's URL (uniform resource locator) rather than accessing the files directly at their storage location. The CDN service then stores copies of requested files at distributed locations for a brief period, such as 24 hours, after each request, depending on the configurations you set.

Figure 5-23 Basic CDN infrastructure



Exam Tip | CDN is a type of overlay network, which you learned about in the previous chapter. VPN, VLAN, and cloud peering are also types of overlay networks.

IoT Cloud Networking

A perhaps unexpected area of cloud growth is in supporting IoT devices and services. Earlier in the course, you read about IoT and a couple of ways cloud technology is involved. Most IoT devices communicate over the Internet with some sort of cloud-based service that allows users to monitor and control their devices. Additionally, many IoT services rely on the cloud for data storage, such as video files collected from security cameras or historical measurements and logs. How does the cloud interact with these devices, and what kinds of technologies are involved?

To begin this discussion, let's elaborate on what types of devices can be included in an IoT deployment's infrastructure. You might already be familiar with common home IoT devices, such as security cameras, smart thermostats, smart door locks, leak sensors, smoke detectors, and smart light bulbs. In more industrial environments, you can find a variety of sensor types (such as temperature and humidity, as pictured in Figure 5-24) and actuators (Figure 5-25), which are components used to control other devices (such as motors to move things or relays to control electrical current). Also consider equipment

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-24 IoT weather sensors



iStock.com/Scharfsinn86

Figure 5-25 IoT in manufacturing



iStock.com/Coffeekaii

fleets managed via IoT, such as service vehicles, shipping vehicles (such as trucks or trains), or even e-bikes or e-scooters (referred to as micromobility vehicles) often available to rent in larger cities (Figure 5-26).

Typically, an **IoT gateway** serves as the intermediary to connect these various IoT devices to the cloud, similar to a network router (Figure 5-27). Many cloud services are needed to manage communications with these systems, and many others are used in the process of developing new IoT applications.

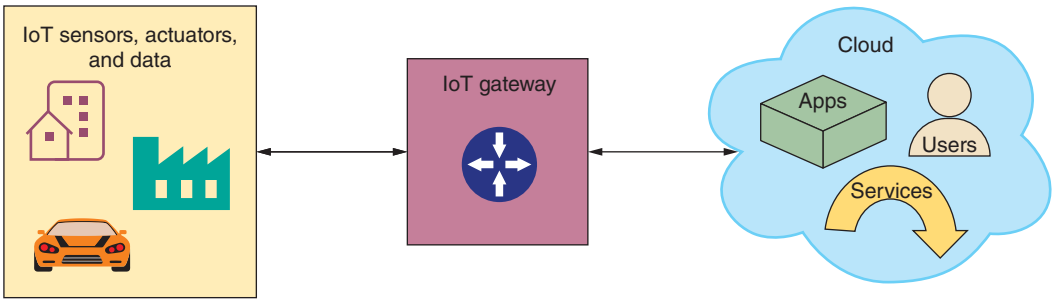
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-26 Fleet of e-bikes



iStock.com/Ivan Pantic

Figure 5-27 An IoT gateway connects IoT devices with the cloud



The following list describes a few of the AWS services in both categories to give you an idea of what kinds of IoT services most of the major cloud platforms offer:

- **AWS IoT Device Management** tracks, monitors, and manages connected device fleets.
- **AWS IoT Core** facilitates communication between IoT devices and cloud-hosted applications.
- **AWS IoT Device Defender** audits IoT device configurations to enforce security best practices.
- **AWS IoT Device Tester** is used to automate IoT testing.
- **AWS IoT Greengrass** brings IoT interactivity to edge devices to reduce latency.
- **AWS IoT Analytics** automates the analysis of data collected from IoT devices.

These services rely on a variety of protocols to support an IoT deployment. Many of these protocols are already familiar to you from other forms of networking, such as HTTPS, TLS, Wi-Fi, Bluetooth, NFC (near-field communication), and cellular standards (such as 5G). Some protocols used by IoT that might not be familiar include the following:

- **MQTT (Message Queuing Telemetry Transport)**—This is a lightweight, open-source messaging protocol, which means it is used to send messages between IoT devices and IoT applications. MQTT is a good choice for IoT devices with limited processing power and networks with low bandwidth, low reliability, or high latency.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **CoAP (Constrained Application Protocol)**—CoAP is also a good choice for networks with limited resources because it limits the size of network packages. CoAP can also be used for document transfer. One strength of CoAP is that it works well for the wide variety of devices often found in IIoT (industrial IoT) environments. However, one of its weaknesses is a lack of built-in security features, instead relying on some other protocol to layer on security.
- **AMQP (Advanced Message Queuing Protocol)**—This open-source messaging protocol is a good fit for connecting a high variety of systems, devices, and applications where cross-compatibility is needed because it relies on a middleware layer to act as an intermediary. AMQP is also known for providing better security than some of the other options.
- **XMPP (Extensible Messaging and Presence Protocol)**—XMPP, also an open-source protocol, was originally developed for instant messaging systems. It has been adapted for IoT systems that require near-real-time communication.
- **DDS (Data Distribution Service)**—DDS is a peer-to-peer communication protocol intended to allow devices to communicate with each other within the same layer of an IoT deployment's infrastructure. For example, IoT monitoring systems can use DDS to communicate with IoT storage systems. DDS also works well when scaled to large deployments, such as for IIoT applications.
- **LoRaWAN (long-range WAN)**—This is a low-power wireless WAN standard that provides long-distance communication in a range of up to 10 miles. Low power not only increases battery life but also decreases data throughput.
- **RTMP (Real-Time Messaging Protocol)** and **RTSP (Real-Time Streaming Protocol)**—These communication protocols support media streaming over the Internet, such as when video and audio are transmitted from a security camera or webcam to an IoT app or website (such as YouTube). RTSP is a better fit for streaming on a local network, such as for video surveillance. But RTMP provides low latency and simple setup to support live streaming over the Internet.

AI in IT | AI (artificial intelligence) enables much of the automation provided by IoT systems. AI analyzes data collected by IoT devices, which supports intelligent decision-making, automated responses, predictive capabilities, and other enhanced features. For example, AI can monitor manufacturing machines and predict when failures are likely to occur. These AI systems can ensure needed parts are ordered and in place before they're needed, and schedule preventive maintenance before failures happen.

When AI and IoT are deeply integrated, the combined technologies are sometimes referred to as AIoT (artificial intelligence of things). This strategy moves AI processing closer to the point of use, which often relies on edge computing infrastructure.

Self-Check

9. Which feature of CDN files helps ensure customers get current copies of cached data?
   a. Proximity
   b. Storage location
   c. TTL
   d. URL
10. Which IoT messaging protocol is most known for providing security?
   a. AMQP
   b. MQTT
   c. CoAP
   d. LoRaWAN

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 5-4: Troubleshooting Cloud Connectivity

Certification

1.3 Explain cloud networking concepts.

2.5 Given a set of requirements, provision the appropriate cloud resources.

6.1 Given a scenario, troubleshoot deployment issues.

6.2 Given a scenario, troubleshoot network issues.

Troubleshooting in the cloud often revolves around addressing lower performance or handling failed instances. In regard to networking, common problems include connectivity issues or misconfigured routes. When troubleshooting cloud networks, it's important to understand how typical network services, such as NAT, routing, and even network time keeping, function differently in the cloud than on-prem. In this section, you learn about common connectivity concerns, status codes, misconfigurations, and other common issues related to cloud networking.

Common Connectivity Problems

When troubleshooting cloud connectivity, whether a connection isn't happening at all or traffic isn't traversing the connection as expected, you'll encounter several common issues. For example, incorrectly managing IP address spaces can often cause connectivity difficulties. As you read earlier in this chapter, peered VPCs cannot use overlapping IP address ranges. Other common causes of peering problems are incorrect or missing route tables or firewall rules. In this section, you further explore cloud connectivity concerns you might encounter.

Unreachable Instance

When you can't connect to a VM instance, it could be a problem with the VM's OS or connectivity settings internal or external to the VM. Some cloud platforms allow you to access task logs for the instance. For example, to access logs for an EC2 instance in AWS, select the instance, click Actions, Monitor and troubleshoot, and Get system log. Use the logged data for further troubleshooting.

When connectivity is slow, consider that you might have insufficient bandwidth at your location for the quantity of communication you're attempting to support. Latency can also be caused by Internet "weather conditions" or by poorly designed routing or security configurations.

Each cloud platform provides additional tools for troubleshooting running instances and their connections. It helps first to understand some of the typical underlying causes of network connectivity issues. Consider the following possibilities:

- **Incorrect IP address or incorrect subnet**—Mistyping a static IP address into an instance interface can result in no connectivity to that interface. Similarly, mistyping a virtual network's or subnet's CIDR range can prevent communications. If you're trying to reach an instance across the Internet, make sure you're targeting the instance's public IP address. If it doesn't have one, you'll likely need to give it one.
- **DNS errors**—You might instead attempt to use the instance's public DNS host name, if available. DNS host names should be enabled in the account, network, or VPC where the instance resides. Also check any relevant DNS records to ensure there are no typos or other errors.
- **Misconfigured firewall rule**—Check to make sure the connection protocol's traffic is allowed through any firewalls, both at the on-prem computer's location and at the cloud instance's location.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For example, if using Remote Desktop to connect to a Windows VM, make sure RDP's port 3389 is enabled on all intervening firewalls. For Linux instances, make sure SSH's port 22 is enabled.

- **Incorrect routing or incorrect gateway**—A missing, misconfigured, or misplaced route in a route table is a common cause of broken connections in the cloud. For example, suppose your AWS VPC has an IG (Internet gateway), but instances in the VPC still can't reach the Internet. Check to make sure the static route to the VPC's IG is properly configured and prioritized in the route table. If the VPC is instead using a NAT gateway, again, check to make sure the route is pointing to the correct gateway.
- **Tunneling issues**—Unused VPN tunnels might time out, causing tunneling issues. You can configure a host or some VPN endpoints to send a message every so often to keep the tunnel active.
- **Misconfigured proxy**—A proxy server acts as an intermediary between external and internal networks, screening all ingress and egress traffic. This helps increase security of the internal network (called network hardening) by masking information about internal resources from outside Internet systems and by providing some degree of content filtering. Proxies are commonly used to protect web servers from external threats. Due to the proxy's role on the network, a misconfigured proxy can prevent or delay network communications.
- **QoS issues**—QoS (Quality of Service) is a group of techniques for adjusting the priority a network assigns to various types of transmissions. QoS configurations don't work across the open Internet. However, QoS can be configured within a cloud network or across a direct connection with your cloud. This is often done for IoT traffic or sometimes for UC (unified communications) traffic, such as VoIP (voice over IP). With IoT, QoS settings are used with the MQTT protocol. The correct QoS level must be set for each message type; confusing, duplicate communications can cause misconfigurations of the devices being controlled.
- **Partial cloud service outage**—Many of the regions and AZs (availability zones) in public cloud platforms function on top of multiple data centers that benefit from isolated power, cooling, and communications infrastructure. Sometimes, one data center is affected by a localized outage, but other data centers in the same AZ or region are unaffected and continue to function. This could affect your access to your data or instances running on dedicated hosts, but might not affect serverless services or managed services. Cloud platforms also sometimes experience partial outages that aren't necessarily related to the underlying physical data center. For example, software bugs or cyberattacks might affect VM instances but not storage services. If your troubleshooting reveals no identifiable misconfigurations on your part, you should also consider checking for any cloud outages, especially a partial outage that might not be so obvious from the symptoms your system is experiencing.

Note 5-7 | You can ping a cloud server, gateway interface, or similar resource just as you would your on-prem interfaces. When using ping in the cloud, carefully consider which two interfaces are supposed to be able to communicate, and then ping from one to the other. If it doesn't work, consider first that the problem could be as simple as enabling ICMP (Internet Control Message Protocol) on an intervening firewall somewhere along the route. Also consider that some cloud platforms place limits on which devices you can ping. For example, in Azure, you can't ping the VNet's default router.

NAT Issues

Earlier in this course, you learned that you can use either a NAT gateway or a NAT instance in AWS to establish communications between the Internet and resources in a private subnet. While AWS recommends using a NAT gateway because the NAT instance feature is being deprecated, you can still manually configure a NAT instance in EC2 using the latest Amazon Linux 2 AMI. By default, an EC2 instance performs source/destination checks to confirm that any traffic it receives lists that instance as either the source or the destination before it accepts that traffic. A NAT instance, however, will need to send and receive traffic that does not list itself as the source or destination. For this reason, you'll need to disable the SrcDestCheck attribute on the NAT instance. Figure 5-28 shows how to access this option from the listed instance in EC2, and Figure 5-29 shows the source/destination option disabled.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-28 Use the Actions menu to change the source/destination check option in AWS

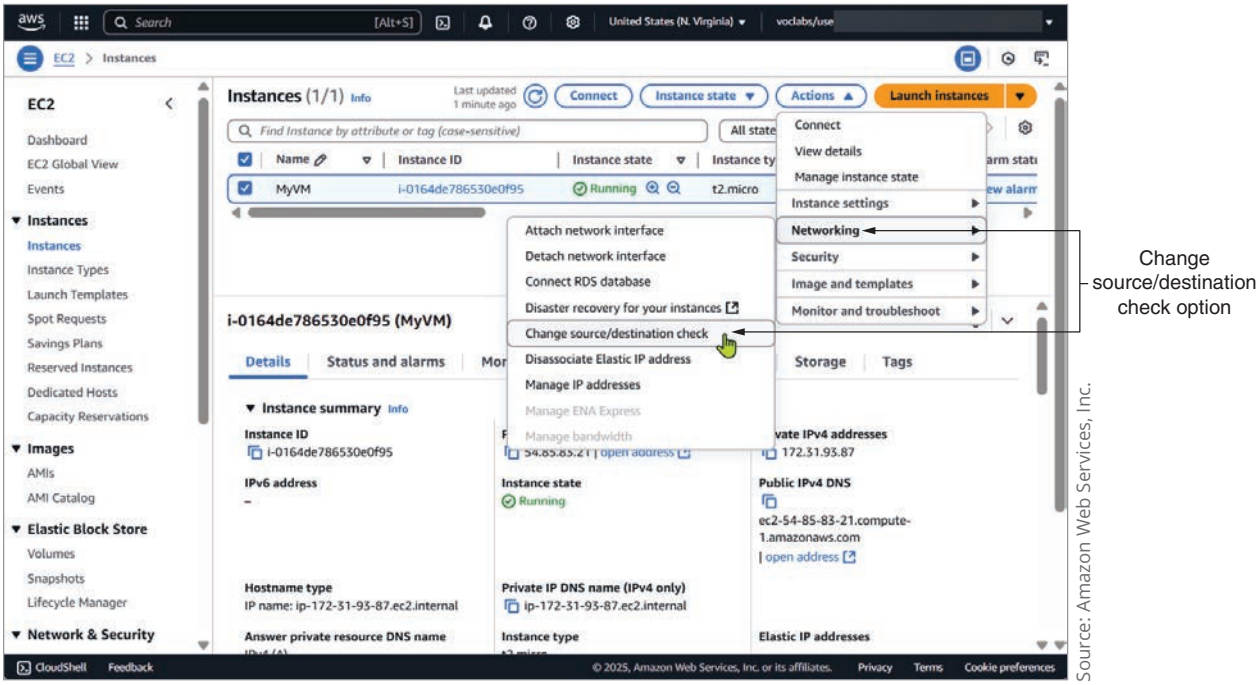
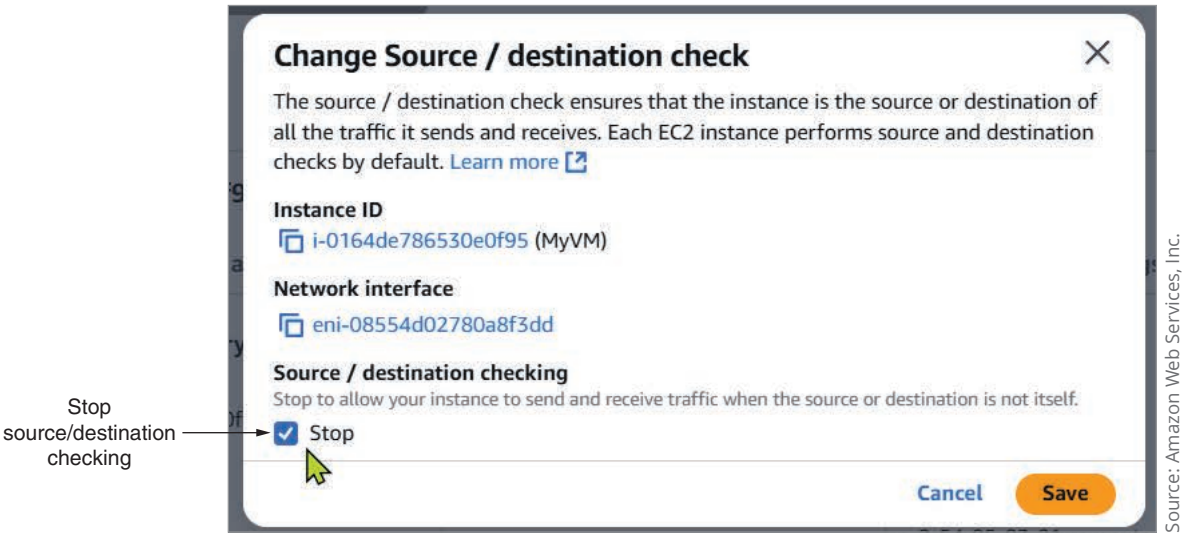


Figure 5-29 Disable source/destination checks



Load-Balancing Issues

Most load-balancing issues result from incorrectly configured load balancers or failed devices somewhere along the load-balanced path. Consider the following possibilities:

- **Load-balancing algorithm**—Typically, load balancers use a round-robin approach so traffic is distributed evenly across targets. Another popular method is weighted round robin where traffic is distributed according to varying weights (such as sending 80 percent of traffic to one target and 20 percent of traffic to another target). Or traffic might be sent to the target currently handling the fewest connections, which is called least connection. Using an inappropriate method for the network can result in one or more targets being overloaded with traffic while other targets are underutilized. In this case, consider changing the load-balancing algorithm or adjusting weights to better distribute traffic across targets.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Persistence**—To maintain the integrity of a conversation between a load-balanced server and a client, the persistence option might be enabled on the load balancer, which you read earlier is called a sticky session. This ensures that communication from a single client IP address continues to be sent to the same server. If client–server communications require consistency throughout the conversation, enabling persistence might be necessary. However, this might also result in a target server being overwhelmed with too much traffic during a session. If this becomes a problem, you can use HTTP cookies to support conversations between a client and different servers in the load-balanced cluster.
- **Headers**—Load balancers require that message headers include necessary information for directing traffic, and these headers must be edited in the process of being redirected so that source and destination IP addresses are properly handled. Common header issues include headers that are too long, that are missing key information, and that include too much, irrelevant, or invalid information. When there's a problem with a header, the load balancer will generate an error code. Refer to the load balancer's vendor documentation to interpret the error code and resolve the issue.
- **Health check protocol**—Load balancers rely on health checks to ensure target systems are functioning within expected parameters. Confirm the required port is open to allow this protocol's traffic to pass any security groups or ACLs (access control lists) protecting target systems—this information will vary by platform and vendor, so check load balancer documentation. In many cases, the health check port will be the same as the port used for other traffic to target servers, but not always.
- **Unhealthy backend server**—A failed health check on a backend server may or may not indicate a problem with the server itself. Confirm that traffic is supported through the entire route from the load balancer to the server, such as checking security groups or NACLs (network access control lists), which you learn more about in a later chapter. Also confirm that health messages from the load balancer are being monitored by the servers and that these servers are responding appropriately. If it's clear that the servers themselves are experiencing issues, you'll need to troubleshoot these VM instances or replace them.
- **Unreachable frontend**—Confirm the load balancer has access to the Internet. For example, in AWS, a load balancer must be attached to a subnet. If it's attached to a private subnet, Internet traffic can't reach the load balancer. Ensure the IP address assignment is correct and that no security devices are blocking legitimate traffic to this device. Also check relevant DNS records. If the frontend is being overloaded with traffic, you might need to configure a pool of public IP addresses, add a backup load balancer appliance, or scale up the load balancer's resources.
- **Security of data in transit**—You can increase the security of data contained in load-balanced traffic by employing encryption protocols, such as HTTPS and TLS. You can reduce processing demands on the backend servers by terminating encrypted connections at the load balancer, which requires the load balancer to perform encryption and decryption of traffic before reaching the backend servers.

HTTP Status Codes

Many cloud services that rely on HTTP, such as gateways and APIs, can return 3-digit **HTTP status codes** to clients. This provides some insight into the state of an HTTP request and any related problems, whether those problems are due to misconfigurations, service outages, or user error. These codes, originally defined by the IETF (Internet Engineering Task Force), are grouped by value into five classes:

- **100-level code**—Indicates the request is being processed and is typically used for informational purposes only.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **200-level code**—Indicates the request was successfully received and processed. For example:
  - **201 Created**—This message says a new resource was created.
  - **204 No Content**—This message says the server successfully processed a request but did not have a response.
- **300-level code**—Indicates further action is needed from the client for the server to complete the request. For example:
  - **307 Temporary Redirect**—This message says that a requested resource has been temporarily moved.
- **400-level code**—Indicates a problem on the client or user side, such as a failed authorization request from providing invalid credentials or a request that is too large. For example:
  - **403 Forbidden**—This message says the server understood the request but has denied fulfilling it, such as when provided credentials are incorrect.
  - **404 Not Found**—This message says the requested information is not available or not located.
  - **429 Too Many Requests**—This message says the client has exceeded the request limit.
- **500-level code**—Indicates a problem on the server or service side, such as a misconfigured API or a service outage.
  - **502 Bad Gateway**—This message says the server processing the original request received an error from a server it attempted to communicate with to fulfill the request.
  - **503 Service Unavailable**—This message says there is a service or server outage, which is preventing a response.
  - **504 Gateway Timeout**—This message says the server processing the original request could not get a timely response from another server it needs to communicate with.

Sometimes the error message will include additional information that can be helpful in troubleshooting the problem. Always examine the error message carefully for insights on where to start searching for the source of the problem. When working with a specific cloud platform, be sure to check that provider's documentation for more detailed information about HTTP status codes and error responses.

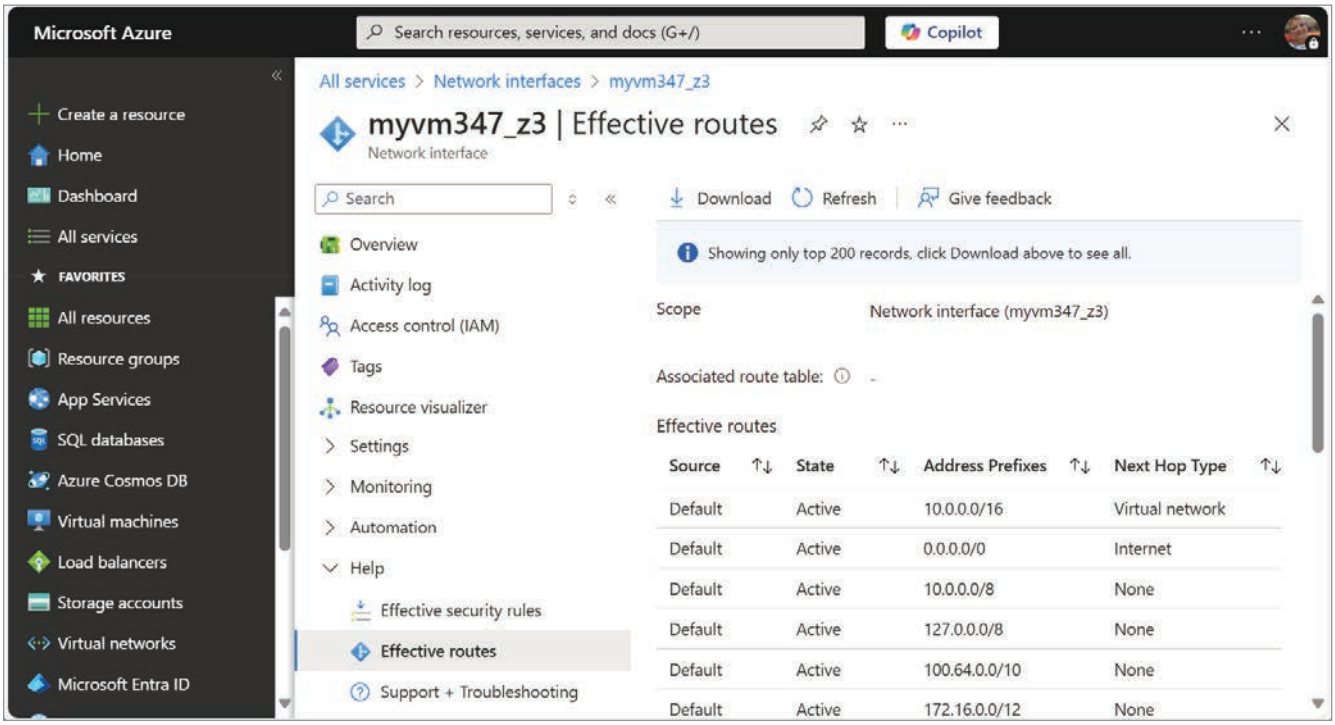
Exam Tip | In preparation for the CompTIA Cloud+ certification exam, consider reading through the IETF's list of HTTP status codes (Section 15 of the linked document) at https://www.rfc-editor.org/rfc/rfc9110.html#name-status-codes.

Misconfigured Routes

Configuring correct routing is essential both for your connection to the cloud and for your connections within the cloud. For example, VM instances must have access to routes to find gateway devices. The gateways themselves (such as AWS's Transit Gateway) must be configured with the needed route tables for sending traffic between various network spaces and locations. Azure offers a helpful feature for its Virtual machines service called Effective routes (Figure 5-30). You can use the feature to identify any routes affecting each network interface on a VM instance. For other tools and services, such as gateways, make sure to check all related route tables when troubleshooting a routing issue.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-30 The effective routes feature for a virtual network interface in Azure



Source: Microsoft

NTP (Network Time Protocol)

NTP (Network Time Protocol) is used to provide time synchronization across a network. The genius of NTP is how it can almost completely account for the variable delays across a network, even on the open Internet. It does this through a hierarchy of time servers where stratum-1 servers communicate directly with a primary time source, such as GPS (global positioning system) or Galileo (Europe's version of GPS). These servers track UTC (Coordinated Universal Time) and provide this information to lower-strata servers. Each hop between NTP servers increases the **stratum** number by 1 up to 16. NTP servers at any stratum can then convert the provided UTC into its local time zone. Not every network has its own time server, but those that do can maintain accuracy for its NTP clients to within a millisecond of each other and are closely synced to the UTC. NTP communicates over port 123.

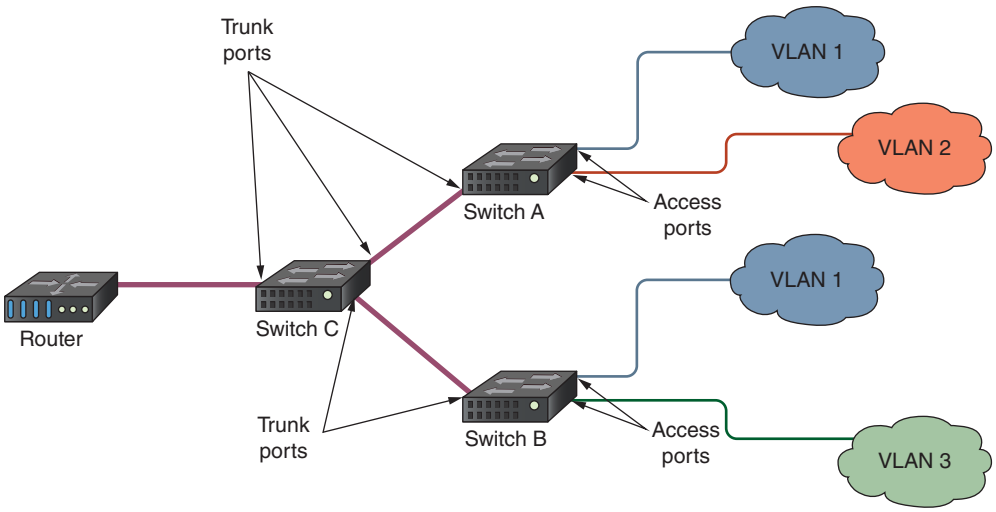
Many cloud platforms provide NTP services native to the platform. For example, the Amazon Time Sync Service is available to all EC2 instances at no charge using the IPv4 address 169.254.169.123 or the IPv6 address fd00:ec2::123. One benefit of this type of service is that you don't need to provide Internet access for your EC2 instances simply to support NTP synchronization. Many network admins set their network clocks according to UTC rather than local time, as local time zones experience some fluctuations due to users being located in different time zones or from shifts required by daylight savings time. Using UTC provides consistent timing for all users and in all logs. In AWS Windows images, the Amazon Time Sync Service is configured by default. For many Linux images, you'll need to edit the chrony configuration file.

VLAN Issues

One potential area for VLAN issues is the configuration of trunk ports and access ports (Figure 5-31). An **access port** on the switch is used to connect VLAN client devices, such as servers or workstations, which are intended to communicate only on a single VLAN. A **trunk port** on the switch is used to connect the switch

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-31 Access ports and trunk ports for VLANs



to its router or to other switches, which are connections that must be able to handle traffic for multiple VLANs. The connection between trunk ports is called a trunk line, and this connection uses VLAN tags to properly direct each network packet according to its VLAN.

Misconfigured VLAN tags typically indicate a misconfiguration on the port itself, which might require a correction of the VLAN assignment or the port type (that is, access port versus trunk port). When working with trunk ports, make sure each port on both sides of each connection is configured with all the VLANs it needs to support. When working with access ports, make sure the device connected to that port is configured with the appropriate VLAN's IP addressing information (such as subnet mask and default gateway).

Self-Check

11. Which protocol requires port 22 be open on a firewall?

a. RDP
b. SSH
c. ICMP
d. ARP

12. Which device is used to camouflage internal network resources from the outside world?

a. Load balancer
b. Proxy server
c. DNS server
d. Default gateway

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

Section 5-1: Hybrid Cloud and Multi-Cloud Networking

- A hybrid cloud is a typical transition phase as companies migrate to the cloud, and many companies ultimately decide to continue indefinitely with a hybrid infrastructure.
- Three major categories of options to make the connection from on-prem to the cloud include the public Internet, VPN connections, and a direct connection service.
- Cheaper and more efficient tools native to the cloud environment can connect cloud networks to each other, such as cloud peering, transit gateways, and cloud WAN.

Section 5-2: Extending Network Services

- IP address spaces are managed by DHCP services, both on-prem and in the cloud; IPAM tools can be used to manage IP spaces across hybrid and multi-clouds.
- DNS servers collect lists of computer names and their associated IP addresses in namespace databases, and some CSPs can allow an organization's on-prem resources to access cloud-based DNS services through a VPN or direct connection.
- Routing can be configured using static routes set by the cloud admin (which do not automatically adapt to route outages) or using dynamic routing, where routes are configured automatically by the cloud platform.
- A load balancer stands between incoming traffic sources and a cluster of servers, and distributes that traffic to all the servers in the cluster according to rules defined by the administrator.

Section 5-3: Extending the Cloud

- Edge computing is an evolution of cloud computing that distributes data processing closer to where data is used.
- A CDN service routes user traffic to the closest edge location with the best performance at the time of the request.
- Most IoT devices communicate over the Internet with some sort of cloud-based service that allows users to monitor and control their devices, and many IoT services rely on the cloud for data storage.

Section 5-4: Troubleshooting Cloud Connectivity

- When troubleshooting cloud connectivity issues, whether a connection isn't happening at all or traffic isn't traversing the connection as expected, common issues include unreachable instances, NAT issues, and load-balancing issues.
- Many cloud services that rely on HTTP can return 3-digit HTTP status codes to clients, which provides some insight into the state of an HTTP request and any related problems.
- Configuring correct routing is essential both for a connection to the cloud and for connections within the cloud.
- NTP relies on a hierarchy of time servers to provide time synchronization across networks and the Internet.
- Misconfigured VLAN tags typically indicate a misconfiguration on the port itself, which might require a correction of the VLAN assignment or the port type.

Key Terms

For definitions of key terms, check the Glossary.

access port
ALB (application load balancer)
application gateway
BGP (Border Gateway Protocol)
CDN (content delivery network)
DNS (Domain Name System or Domain Name Service)
edge computing
HTTP status code
hub-and-spoke topology
IoT gateway
load balancing
NLB (network load balancer)
stratum
transit gateway
transitory peering
trunk port
VPN (virtual private network)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

BGP	Border Gateway Protocol
ICMP	Internet Control Message Protocol
LAN	local area network
RTMP	Real-Time Messaging Protocol
TCP	Transmission Control Protocol
VLAN	virtual local area network

Review Questions

1. What cloud technology can increase the speed of compute processes in smart cars?
   a. NLB
   b. Edge
   c. Peering
   d. CDN

2. Which device hosted in a VPC provides a cloud-based application with a VPN connection to an on-prem database?
   a. Client VPN
   b. VGW
   c. Site-to-site VPN
   d. CGW

3. What type of virtual network can be peered with many other virtual networks to allow communication between all connected networks?
   a. Private
   b. Spoke
   c. Public
   d. Hub

4. Why are hosts in a VLAN restricted to layer 2 communications?
   a. VLANs must use MAC addresses.
   b. VLANs rely on UDP.
   c. VLAN tags are added at layer 2.
   d. VLANs define broadcast domains.

5. Suppose you have a large cluster of web servers in one availability zone and a much smaller cluster in a different availability zone. You want to make sure your larger cluster takes a larger portion of the website traffic. Which load-balancing method is the best fit?
   a. Least connection
   b. Round robin
   c. Persistent
   d. Weighted round robin

6. Which protocol can secure load-balanced traffic?
   a. TLS
   b. HTTP
   c. DNS
   d. VIP

7. What would you need to change in your VPC configurations to ensure VM instances use DNS servers hosted by your organization?
   a. DNS records
   b. Scope options
   c. Default gateway
   d. CIDR block

8. At what level is an IP address space typically defined in the public cloud? **(Select two.)**
   a. VPC
   b. Region
   c. Subnet
   d. Gateway

9. What misconfiguration will result in failed health checks from a load balancer's server cluster?
   a. Misconfigured firewall rule
   b. Incorrect load-balancing algorithm
   c. Misconfigured sticky session
   d. Misconfigured headers

10. Which load-balancing method is the best fit for a server cluster where workloads can be highly unpredictable in the time required to complete each series of tasks?
    a. Least connection
    b. Round robin
    c. Persistent
    d. Weighted round robin

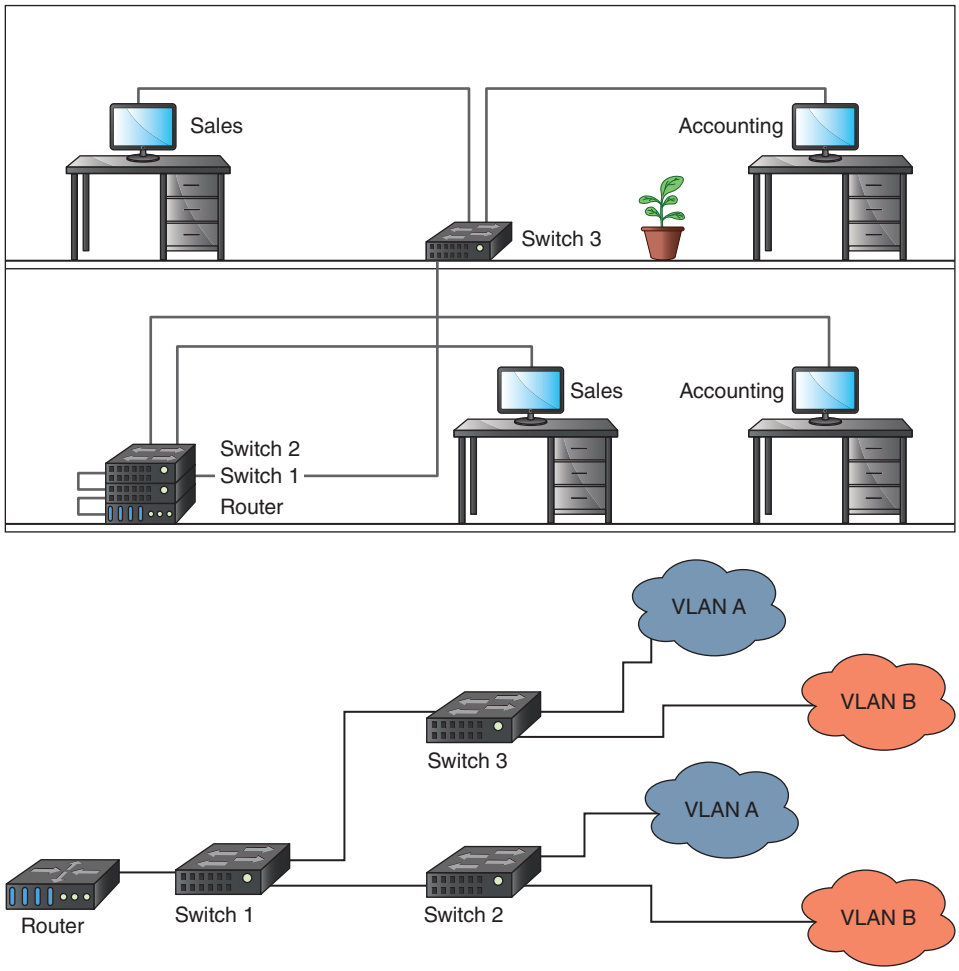
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Scenario-Based Questions

Question 5-1

Marc has configured two VLANs on his office network: VLAN A and VLAN B. He has three switches: Switches 1, 2, and 3. As illustrated in Figure 5-32, Switch 1 on the first floor of Marc's office building is connected to his router. It's also connected to Switch 2 on the first floor and to Switch 3 on the second floor. VLAN A handles traffic for the Accounting Department, and VLAN B handles traffic for Sales. Both VLANs have network clients on each floor.

Figure 5-32 Marc's office network



Question 5-2

Marc is having trouble getting clients on different VLANs talking to each other, even when they're on the same floor. For example, the sales computer on the first floor can't connect to the accounting computer sitting right next to it. It's the same problem on the second floor. Which device is most likely the source of the problem?

a. Switch 1
b. Switch 2
c. Switch 3
d. The router

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Question 5-3

Sami is trying to ping between instances in her newly created VNet, but it's not working. She double-checked the IP addresses she's using, and confirmed the instances are running. She decides to check the network's firewall settings to ensure that the ping protocol is allowed. Which protocol should she check for?

a. BGP
b. SSH
c. DHCP
d. ICMP

Question 5-4

Amari is troubleshooting a problem with a website, and he keeps receiving a cryptic error message, "Error 400 – Request failed." On which device should he start his troubleshooting efforts?

a. Load balancer
b. Web client
c. Web server
d. Transit gateway

Question 5-5

Luca has just added a new workstation to a VLAN on her network. During testing, she finds the workstation is not able to communicate with the rest of the network or the Internet. Which of the following is most likely to be the misconfiguration?

a. The workstation is plugged into a switch port configured as a trunk port.
b. The workstation is configured with the wrong default gateway information.
c. The workstation is configured with the wrong VLAN tag information.
d. The workstation is plugged into a switch port configured as an access port.

Hands-On Projects

Note 5-8 | None of the Hands-On Projects in this chapter can be completed without incurring costs for resources used. If students are using special access to cloud platforms with free educational credits (such as with AWS Academy), this should not be a problem. The resource costs are low and will not use up much of their free educational credits. However, if students are using a standard account (not an educational account such as AWS Academy) with limited or expired free trial credits, you might consider having students complete both Challenge Projects in each cloud platform rather than completing these Hands-On Projects. Note that an extra Challenge Project is provided in this chapter as an alternative to the Hands-On Projects if students are not able to use free credits for these VPN projects.

Note 5-9 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Project 5-1: Explore VPN Options in AWS

Estimated time: 30 minutes
Objective 1.3: Explain cloud networking concepts.
Resources:

- AWS account
- Drawing app, such as Visio or app.diagrams.net
- Internet access

Description: A VPN connection in the cloud, with all its protocols and configuration options, seems like an abstract concept until you start building your own VPNs. Once you explore and build the various components yourself, the whole idea starts to make more sense. In this project, you'll configure a VPN connection to a VPC. The VPN will connect a virtual gateway device in your VPC to a customer gateway device in your on-prem network. If you have an actual VPN gateway device at your location, you can use the real information for that device and establish a real VPN connection to AWS. The instructions in this project assume that you do *not* have a VPN gateway device at your location. Instead, you'll configure the VPN connection without activating it.

NOTE ABOUT FEES: Data transfer across a VPN and use of the related gateways incur charges in AWS. To avoid these charges, the steps indicate at what point further progress in the project is optional. If you have plenty of free credit in your account (such as when using an AWS Academy account) or if spending a few dollars is not problematic for you, you might want to complete the project for the sake of the experience. However, these last few steps are not required to earn full credit for the project. Note that if you are using AWS Academy, the free credits provided with your course should cover any incurred costs from completing the entire project, as tested by the author.

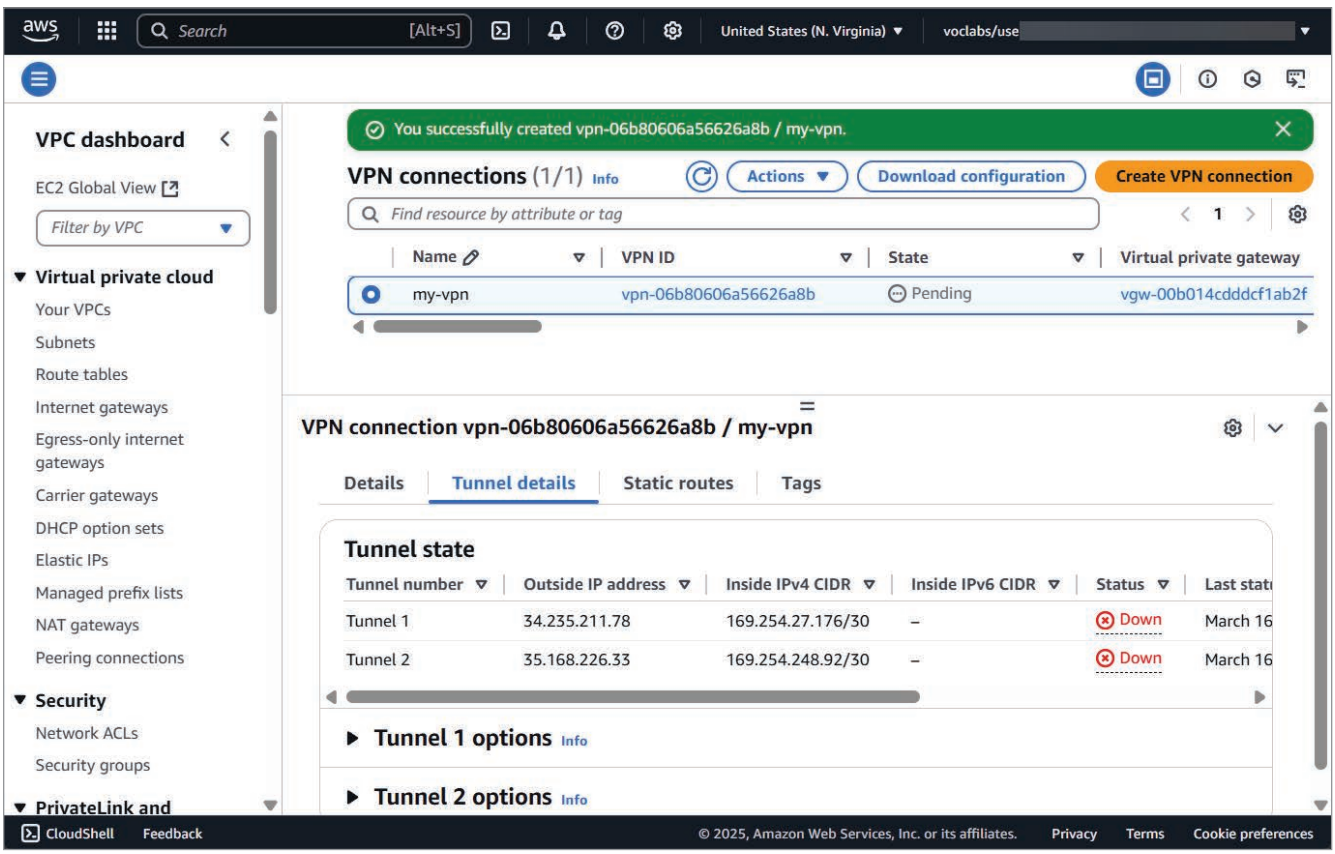
Complete the following steps:

1. In your AWS Management Console, create a VPC, which you learned to do in Project 4-1. What CIDR block did you use?
2. In the list of VPCs, select your new VPC and click its main route table. Give this route table a name that you can easily identify later, such as MyVPN-RT. To add the name, click the pencil icon that appears when you point to the route table's name field. What did you name your route table?
3. Before creating the VPN resources in AWS, you need the public IP address for your local network. Using a search engine, search **what is my ip**. Copy your public IP address from the search results.
4. Now you're ready to create the VPN gateways. In the navigation pane, scroll down and click **Customer gateways**. Recall that the CGW is a VPN appliance in the customer's on-prem network. This process does not create a new resource in AWS—it simply informs AWS of how to find your VPN device in your on-prem network. Click **Create customer gateway**.
5. Give the CGW a name, such as MyCGW, and paste your public IP address in the IP address field. Click **Create customer gateway**. In the listing for your customer gateway, what is your CGW's type?
6. Now you're ready to create a virtual gateway in AWS, which will provide access to the VPN connection on the VPC's side. In the navigation pane, click **Virtual private gateways**, and then click **Create virtual private gateway**.
7. Give the VGW a name, such as MyVGW. Leave the ASN selection as **Amazon default ASN**, and then click **Create virtual private gateway**. What did you name your VGW?
8. Click to select your new VGW. To attach it to your VPC, click **Actions**, and then click **Attach to VPC**. Select the VPC you created in Step 1, and then click **Attach to VPC**.
9. If you were creating this VPN connection for use with a real CGW, you would also need to make some changes to the security group for this VPC. Like allowing traffic through a firewall, this would enable RDP, SSH, and ICMP traffic into the VPC from the on-prem network. You learn more about security groups in a later chapter. For now, because you won't activate the VPN connection in this project, you can skip this step. However, if you're planning to configure the VPN connection on an actual VPN gateway at your physical location, be sure to search online for the changes you need to make to the VPC's security group.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10. You're now ready to create the VPN connection between these two gateways. In the navigation pane, click **Site-to-Site VPN Connections**. Click **Create VPN connection**.
11. Give the VPN a name, such as MyVPN. Select the VGW and CGW that you created earlier in this project. Change the routing option to **Static**. In the Static IP prefixes field that appears, enter a private CIDR range that you might be using in your on-prem network. If you were establishing this VPN connection for real, you would need to ensure that your on-prem network's private CIDR ranges and your VPC's CIDR ranges don't overlap. What private CIDR range did you list for your on-prem network?
12. **WARNING:** If you need to avoid incurring any charges in your AWS account, only read through the following steps, and pick up at Step 20.
    - If you are using AWS Academy and you have available credits, you should be able to complete this project using only a small amount of credit without encountering any permissions restrictions. Therefore, continue with the next step.
    - If you are using a standard AWS account and you must avoid any charges, click **Cancel**, read the next few steps, and complete Step 14. You can read the remaining steps instead and play the video for this project to observe the project's completion. Otherwise, continue with the next step.
13. If you have credits in your account or if you're comfortable with paying a few dollars to complete this project, click **Create VPN connection**.
14. The VPN connection will take a few minutes to reach an available state. In the meantime, select your VPN and click the **Tunnel details** tab in the resource information pane (Figure 5-33). Notice the VPN has two tunnels to provide redundancy. Each of these tunnels can be configured separately when you first create the VPN.

Figure 5-33 Two tunnels in one VPN connection

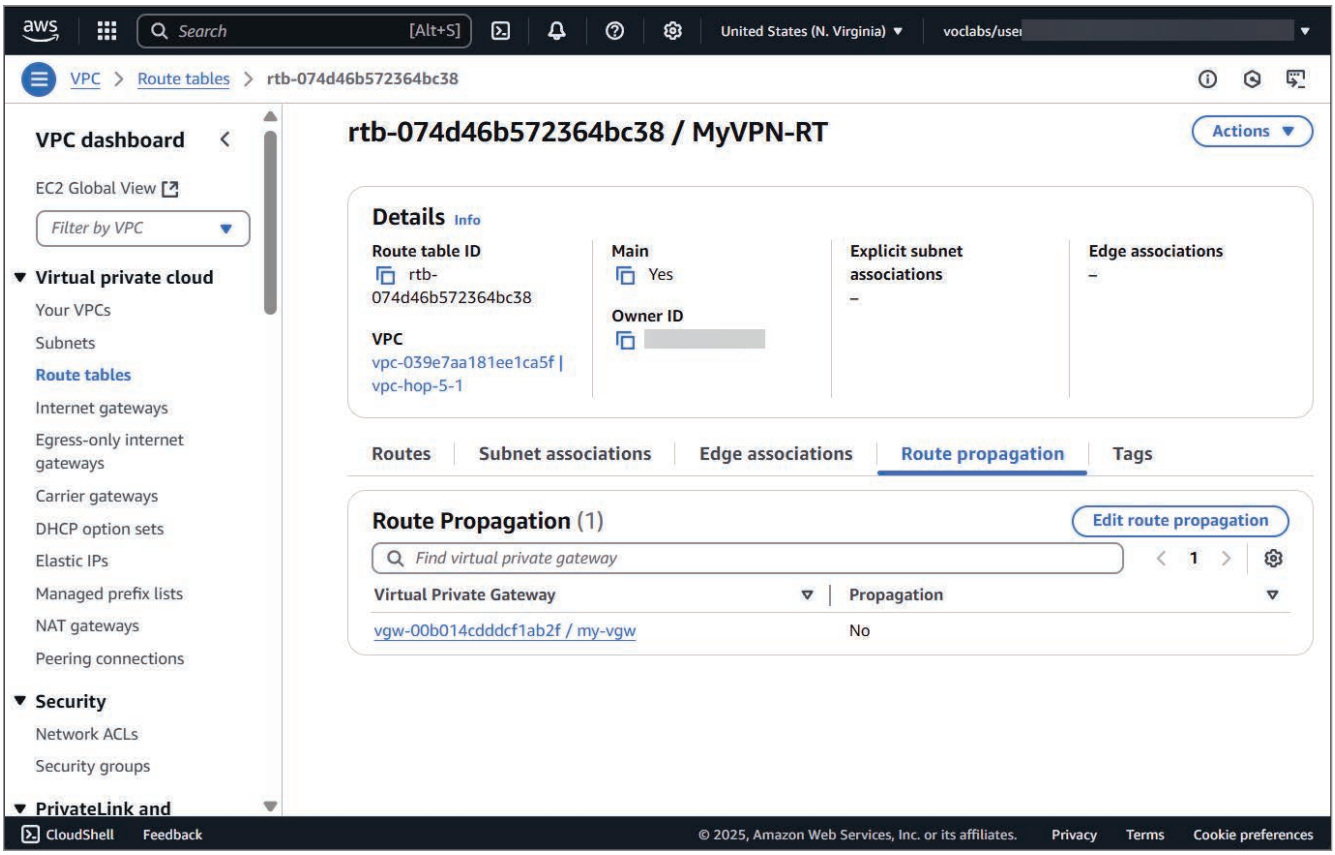


Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

15. Instances in your VPC will need routes in the VPC's route table to find the CGW and to know what IP addresses exist on the other side of the VPN connection. You can configure these routes manually or enable route propagation so this is done automatically. To enable route propagation, in the navigation pane, click **Route tables** and select the route table you identified in Step 2. Click the **Route propagation** tab in the resource information pane (Figure 5-34), and then click **Edit route propagation**. Click the box to enable route propagation, and click **Save**.

Figure 5-34 Route propagation enables routing over the VPN connection

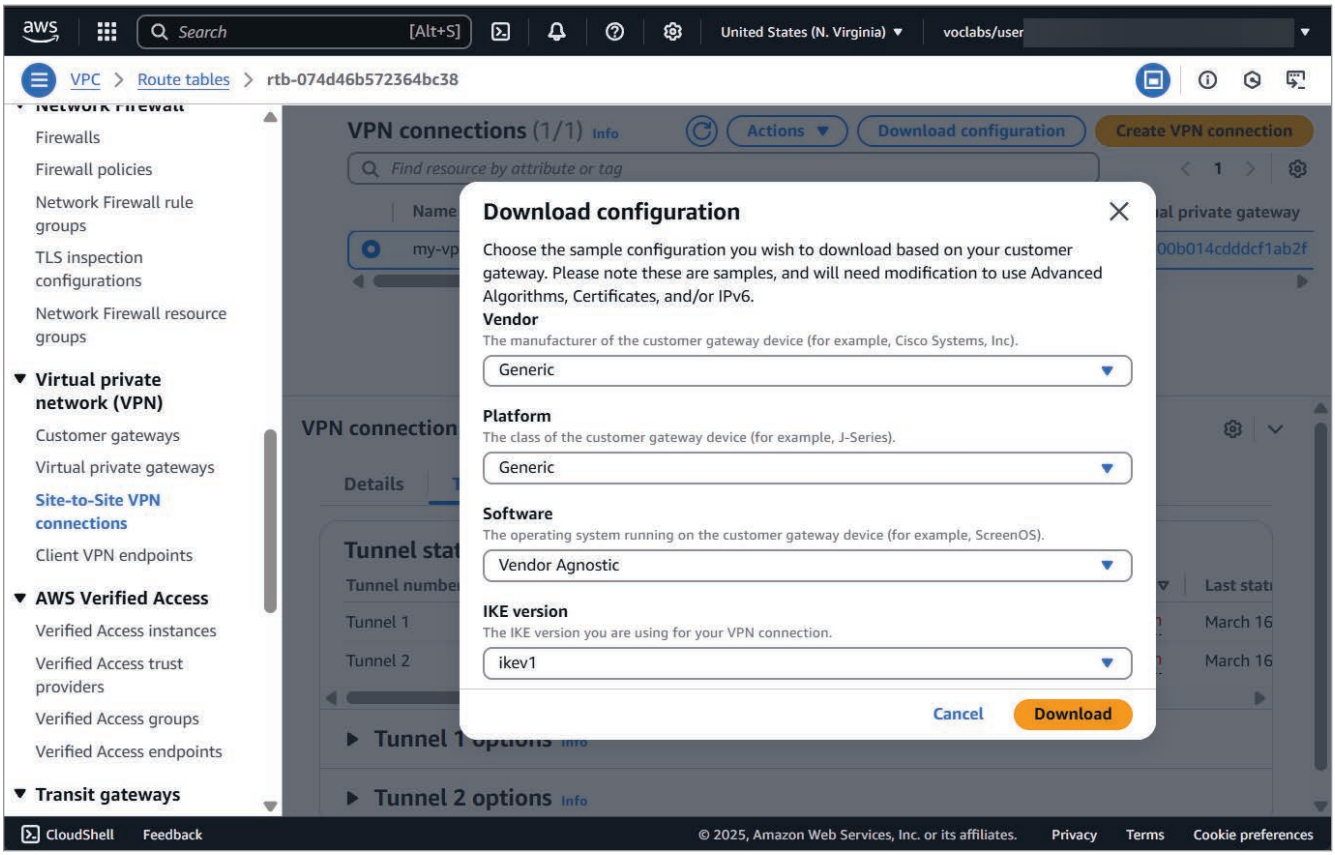


Source: Amazon Web Services, Inc.

16. Return to the **Site-to-Site VPN connections** page. By now, your VPN connection should be available; if not, give it a few more minutes. You might need to refresh the list to display the updated status. Select your VPN and click the **Tunnel details** tab. Notice that both tunnels are still down. To complete the VPN configuration, you'll need to download a configuration file from AWS and use it to configure the customer gateway device in your local network. If you don't have a real VPN gateway, you can still download the configuration file to explore what information it includes.
17. Click **Download configuration**. For the Vendor, select **Generic** or use the correct information for your device if you have a real VPN gateway. Keep the default settings that appear (Figure 5-35), unless you're using an actual device and need to provide more specific information. Which IKE version is selected by default?
18. Click **Download** and save the file in a convenient location on your local computer. After the file downloads, open it and explore the information given.
19. Particularly important pieces of information used to complete the configuration on the customer's side include the preshared key and the outside IP addresses for the CGW and VGW (Figure 5-36). If you have an actual VPN gateway at your location, you can configure it with the provided information from the download and then activate the VPN connection.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-35 Generic settings for a customer gateway device



Source: Amazon Web Services, Inc.

Figure 5-36 Public IP addresses allow VPN gateway devices to find each other

```
The Customer Gateway and Virtual Private Gateway each have two addresses that relate
to this IPSec tunnel. Each contains an outside address, upon which encrypted
traffic is exchanged. Each also contain an inside address associated with
the tunnel interface.

The Customer Gateway outside IP address was provided when the Customer Gateway
was created. Changing the IP address requires the creation of a new
Customer Gateway.

The Customer Gateway inside IP address should be configured on your tunnel
interface.

Outside IP Addresses:
  - Customer Gateway                : [redacted].97.86
  - Virtual Private Gateway          : 3.220.93.162

Inside IP Addresses
  - Customer Gateway                : 169.254.232.178/30
  - Virtual Private Gateway          : 169.254.232.177/30

Configure your tunnel to fragment at the optimal size:
  - Tunnel interface MTU     : 1436 bytes


#4: Static Routing Configuration:

To route traffic between your internal network and your VPC,
you will need a static route added to your router
```

Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

20. Delete all the resources you created in this project, including the VPC, both gateways, and the VPN connection (if you created it). In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
21. Use a drawing app, such as Visio or app.diagrams.net, to **draw a diagram of the resources used to create this VPN connection**. Include the VPC, the on-prem network, both gateways, and the VPN connection (with two tunnels) using official AWS diagram symbols, such as those illustrated in this chapter. Submit this visual with your responses to this project's questions.
22. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 5-10 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 5-2: Explore VPN Options in Azure

Estimated time: 1 hour (includes 30 minutes' wait time at the end of Step 4)
Objective 1.3: Explain cloud networking concepts.
Resources:

- Azure account and subscription
- Drawing app, such as Visio or app.diagrams.net
- Internet access

Description: A significant difference between Azure's VPN options and AWS's VPN services is that Azure allows you to create a VPN between two VNets. While you can do this in AWS, it takes more finagling. And in reality, it's rare that you would have a situation where it would make sense to pay for a VPN connection between VNets or VPCs in the same cloud platform.

In this project, you'll create a VPN connection between two VNets in the same subscription. Although it's possible to connect VNets in different subscriptions, that process is much more complicated.

A VPN connection to a VNet in Azure relies on a special subnet called a gateway subnet, which holds two VM instances that manage traffic through the gateway subnet. In this project, you'll practice configuring and connecting these gateway subnets.

NOTE ABOUT FEES: Unfortunately, charges cannot be avoided to complete any significant portion of this project. However, if you follow the steps carefully and quickly (complete in less than an hour), the total charge should not be more than a few cents.

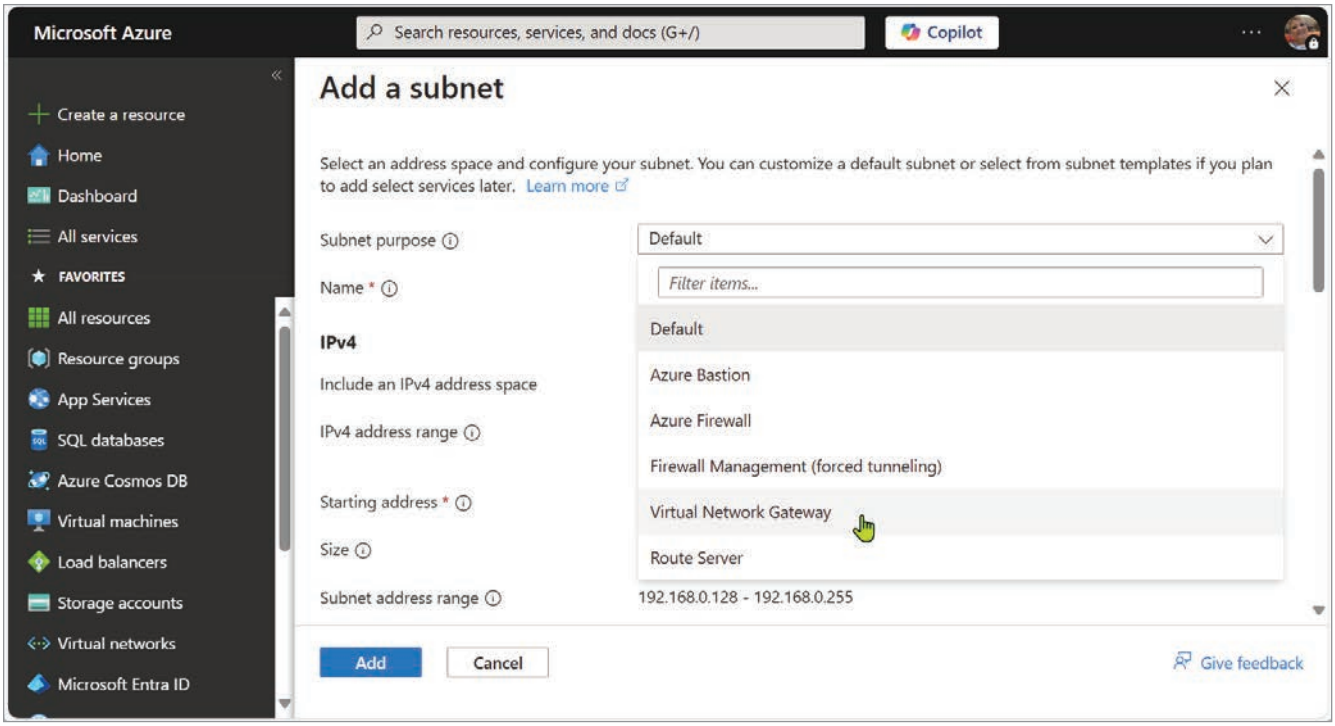
Complete the following steps:

1. In your Azure portal, create two VNets; you learned how to create VNets in Project 4-2. Use the following information for each VNet:
   a. Create a new resource group for each VNet. What did you name your resource groups?
   b. Take note of the region and use the same region for both VNets.
   c. Use a different CIDR block for each VNet that does not overlap each other. What CIDR blocks did you use?
   d. Use a smaller CIDR block for the subnet in each VNet, as you'll need some of the available address range for the gateway subnets. What CIDR block did you use for each subnet?
2. Create a new gateway subnet in each VNet. You learned how to add a second subnet to a VNet in Project 4-2. However, this time you need to create a gateway subnet, not a regular subnet, by selecting

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Virtual Network Gateway for Subnet purpose (Figure 5-37). The rest of the process is similar except that you won't be able to change the name of the subnet. You should also use a /27 CIDR block. Don't change any other default settings. What CIDR block did you use for each gateway subnet?

Figure 5-37 A gateway subnet gives a VNet access to a VPN connection

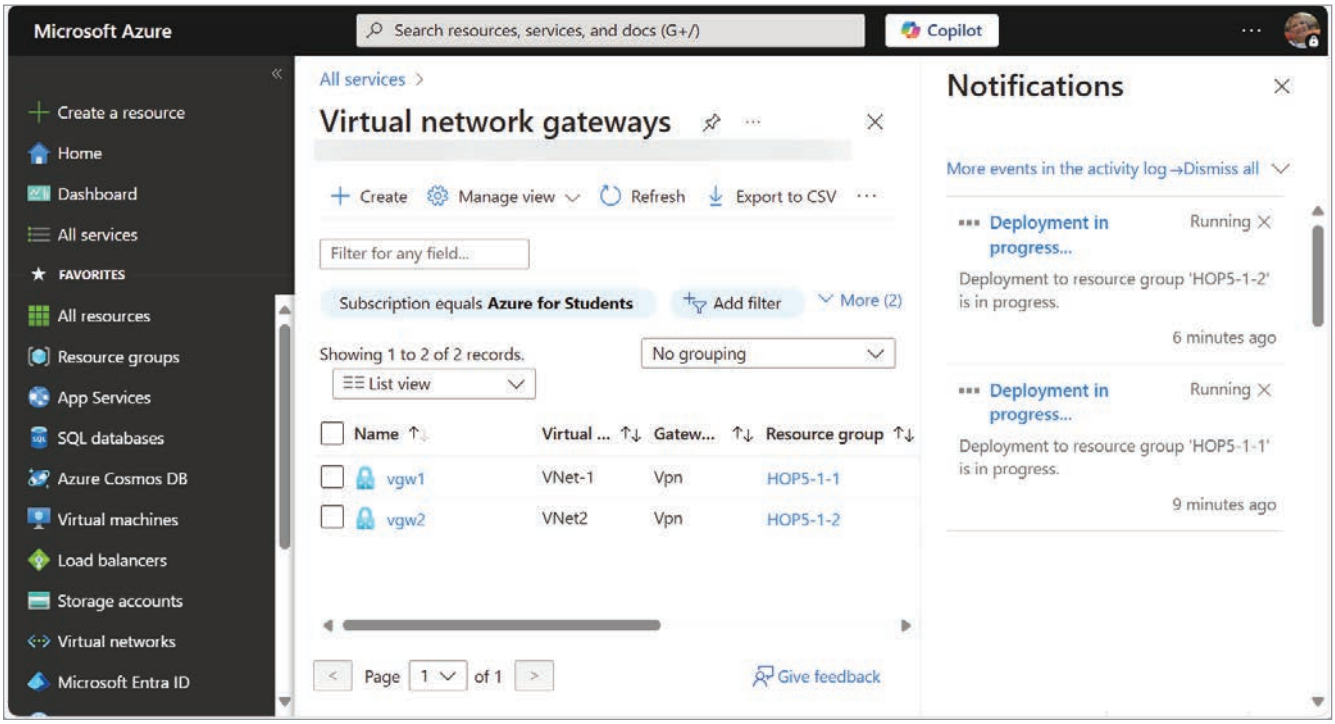


Source: Microsoft

3. Now you're ready to create a virtual network gateway for each VNet. In the portal menu, click **Create a resource**. In the Search services and marketplace field, enter **virtual network gateway**. Click **virtual network gateway** when it appears as a suggestion. On the Virtual network gateway tile, click **Create** and **Virtual network gateway**.
4. Configure the following settings:
   a. Give the VGW a name and make sure you add it to the same region where you created your VNets. What did you name the VGW?
   b. Make sure **VPN** is selected for the gateway type.
   c. Under SKU, choose **VpnGw1**. This basic VPN supports fewer users, less security, and no BGP routing. However, this option is sufficient for this project.
   d. Assign the VGW to a VNet.
   e. Make sure **Create new** is selected for the public IP address, and then give the IP address a name.
   f. Disable active-active mode.
   g. Confirm that BGP is disabled.
   h. Click **Review + create** and then click **Create**.
   i. Repeat Steps 4a through 4h to create a second VGW in your second VNet.
   j. Make sure you've created a VGW for each VNet before continuing with the next step, and make sure all deployments are complete. To monitor deployment progress, click the **Notifications** bell icon to open the Notifications blade (Figure 5-38). It might take quite a while—perhaps 30 minutes or more—for the VGWs to fully deploy. Grab a cup of coffee or tea while you wait.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5-38 Deployment in progress for both VGWs



Source: Microsoft Azure

5. After all deployments have completed, you're ready to create the connection between the VGWs. Go to your list of virtual network gateways. One way to do this is to click **All services** in the portal menu, click **Networking**, and then click **View** on the Virtual network gateways tile. Click the first VGW.
6. Click **Settings** and **Connections**, and then click **Add** to create a VPN connection.
   a. Make sure that **VNet-to-VNet** is selected for the connection type.
   b. Give the connection a descriptive name, such as MyVNet1toMyVNet2. Click **Next: Settings**.
   c. Select each VGW for each end of the connection.
   d. Create a shared key (PSK) for the connection (this is basically a password), and write it down for use in the next step.
   e. Click **Review + create**. Once validated, click **Create**.
7. Create a connection in reverse from the second VGW. Be sure to select the VGWs in the reverse order as the first connection and use the same shared key.
8. Monitor the Connections list for one of the VGWs until the status for each connection changes to Connected (this might take several minutes).
9. Go to the Properties page for each VGW. What public IP addresses are being used at each end of the VPN connection?
10. Delete all the resources you created in this project, including both VNets, both resource groups, all related subnets and gateway subnets, both VGWs, and all connections. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
11. Use a drawing app, such as Visio or app.diagrams.net, to **draw a diagram of the resources used to create this VPN**. Include both VNets, all subnets, both gateways, and the VPN connections using official Azure diagram symbols, such as those illustrated in this chapter. Submit this visual with your responses to this project's questions.
12. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 5-11 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 5-3: Explore VPN Options in GCP

Estimated time: 30 minutes
Objective 1.3: Explain cloud networking concepts.
Resources:

- GCP account
- Drawing app, such as Visio or app.diagrams.net
- Internet access

Description: As with Azure, GCP allows you to create a VPN connection between two VPCs. In this project, you'll create a VPN connection between two VPCs in the same region. Data transfer across a VPN and use of the related gateways incur charges in GCP.

NOTE ABOUT FEES: Unfortunately, these charges cannot be avoided to complete any significant portion of this project. However, if you follow the steps carefully and quickly (complete in less than an hour), the total charge should not be more than a few cents.

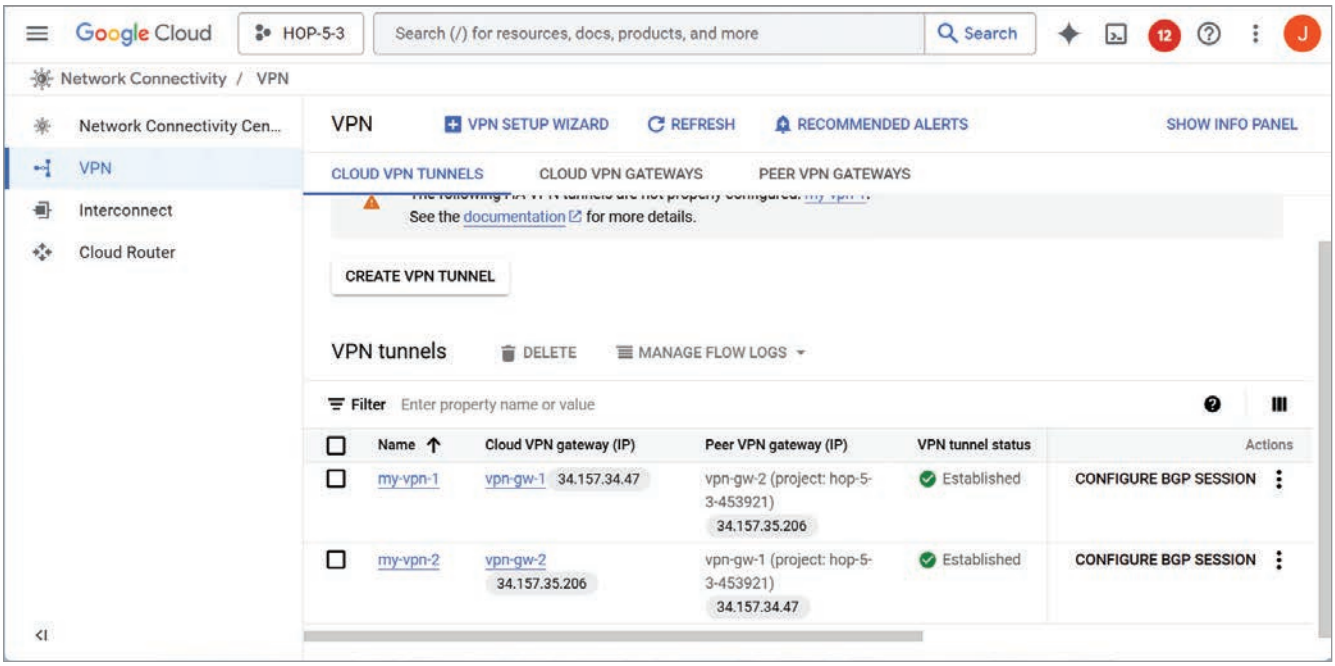
Complete the following steps:

1. To make cleanup easier at the end of this project, first create a new project. What did you name your project?
2. Make sure you're in the new project you just created. Then create two custom-mode VPCs with one subnet each—you might be prompted to enable Compute Engine API first. You learned how to create VPCs in Project 4-3. Place each VPC's subnet in the same region, and make sure the CIDR blocks for the VPCs' subnets don't overlap. What region did you use? What CIDR blocks did you use?
3. In the navigation menu, click **View all products**. Click the **Networking** category and then click **Network Connectivity**. You might need to enable the Network Connectivity API.
4. Click **VPN** and click **CREATE VPN CONNECTION**. Configure two VPN gateways with one in each VPC, using the following information:
   a. Under VPN options, make sure **High-availability (HA) VPN** is selected and click **CONTINUE**.
   b. Give the VPN gateway a name, such as vpn-gw-1.
   c. For Network, select the first VPC you created in Step 2.
   d. Select the VPC's region.
   e. Click **CREATE & CONTINUE**.
   f. You're not ready to create VPN tunnels until after you create the second VPN gateway, so in the navigation pane, click **VPN**.
   g. Click the **CLOUD VPN GATEWAYS** tab, and then click **CREATE VPN GATEWAY** to add a second VPN gateway resource.
   h. Create the second VPN gateway, adding it to the second VPC you created in Step 2.
5. You need a Cloud Router in each VPC to manage communications across the VPN:
   a. In the navigation pane, click **Cloud Router**, and then click **CREATE ROUTER**.
   b. Give the router a name, place it in the first VPC you created in Step 2, and use the VPC's region. For Google ASN (autonomous system number), enter **65400**. Leave the other default settings and click **CREATE**.
   c. Create a second cloud router in the second VPC you created in Step 2 and use Google ASN **65500**.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6. Now you're ready to configure the tunnels:
   a. In the navigation pane, click **VPN**. On the CLOUD VPN TUNNELS tab, click **CREATE VPN TUNNEL**.
   b. Choose the first VPN gateway you created earlier and click **CONTINUE**.
   c. Under Peer VPN gateway, select **Google Cloud VPN Gateway** to choose your other VPC for the VPN connection. Choose the project you created in Step 1, and then choose the second VPN gateway you created in this project.
   d. Select **Create a single VPN tunnel** to minimize expense. Choose the first cloud router you created earlier.
   e. Give the tunnel a name, such as my-vpn-1.
   f. Generate and copy an IKE preshared key. Paste this key into a text document so you can use it again in a moment.
   g. Click **CREATE & CONTINUE**. You do not need to configure BGP sessions for this project, so click **CONFIGURE BGP SESSIONS LATER**.
7. Create a second VPN tunnel from the second VPN gateway to the first one and be sure to use the same IKE preshared key from Step 6f.
8. To use this VPN connection for communication between the VPCs, you would need to create firewall rules on both sides of the VPN to allow traffic through the gateways. You learn how to configure firewall rules in a later chapter, and this step is not necessary for the purposes of this project. For now, return to the **CLOUD VPN TUNNELS** tab and confirm that both tunnels show as Established (Figure 5-39). What IP addresses are assigned to each end of each VPN tunnel?

Figure 5-39 Both VPN tunnels are established



Source: Google LLC

9. On the **CLOUD VPN GATEWAYS** tab, what IP addresses are listed here for each gateway? Why do you think there is a discrepancy between the information on this tab and the information on the CLOUD VPN TUNNELS tab?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 5-12 | VPN tunnels incur a per-hour charge, although this fee is not large. For example, in a test where these tunnels were left running overnight, the total fee was less than $1. Still, you don't want to leave the tunnels running any longer than necessary for this project.

10. Go to the list of projects and shut down the project you created in Step 1. Confirm the project is shut down and scheduled for deletion.
11. Make sure to properly sign out of your account.
12. Use a drawing app, such as Visio or app.diagrams.net, to **draw a diagram of the resources used to create this VPN**. Include both VPCs, both subnets, both gateways, both cloud routers, and both tunnels using official GCP diagram symbols, such as those illustrated in this chapter. Submit this visual with your responses to this project's questions.

Note 5-13 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Challenge Projects

Note 5-14 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 5-1: Configure Cloud Peering

Estimated time: 45 minutes (+20 minutes for group work, if assigned)
Objective 1.3: Explain cloud networking concepts.
Group work: This project includes enhancements when assigned as a group project.
Resources:

- Public cloud account (such as AWS, Azure, or GCP)
- Internet access

Description: VPN connections are the traditional method for connecting networks across the Internet. However, for network connections within the cloud, you have the option to use a more streamlined configuration to connect VPCs or VNets to each other. This service is called peering, which you read about in the chapter. Cloud peering, for what it does, is a significant improvement over using VPNs when connecting virtual networks to each other. Peering is simpler to set up and is typically cheaper than a VPN. Because of your experience at this point with configuring VPNs, peering connections should be easy for you to figure out. In fact, this is an excellent opportunity for you to practice teaching yourself a new skill in the cloud.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In this project, you'll use one or more cloud platforms to practice researching and teaching yourself how to peer virtual networks. Complete the following steps:

1. Choose a cloud platform, such as AWS, Azure, or GCP. You can repeat these steps with a second or third platform later. Check with your instructor on the requirements for this Challenge Project.
2. If you need to create a project or resource group to manage project resources when you're finished, do so now.
3. Research the platform's peering service and respond to the following questions:
   a. How much does peering cost?
   b. Can you peer with a virtual network in another account?
   c. Is peering transitive?
   d. Within which service is peering configured in the cloud platform?
   e. What resources must already exist before peering can be established?
   f. What are the basic steps to create the peering connection?
   g. What significant limitations exist on the peering service for this platform?
4. In your selected cloud platform, create two virtual networks with nonoverlapping CIDR blocks, such as 192.168.0.0/24 and 192.168.1.0/24. Create a peering connection between these virtual networks. **Take a screenshot** showing the peering connection; submit this visual with your responses to this project's questions.
5. In your selected cloud platform, create two virtual networks with overlapping CIDR blocks, such as 192.168.0.0/23 and 192.168.1.0/23. Attempt to create a peering connection between these virtual networks. What happens?
6. Suppose you wanted to peer three virtual networks in a mesh network. List three private /24 CIDR blocks that would work for this scenario.
7. **For group projects:** Each group member should create a VPC or VNet in the same cloud platform. Work together to try to create a peering connection between your VPC or VNet and another group member's VPC or VNet. What information did you have to share with each other? What problems did you run into? Were you successful in establishing an active peering connection?
8. Delete any cloud resources you created for this project. Repeat the project with another cloud platform if desired or if required by your instructor.
9. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 5-15 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Challenge Project 5-2: Configure Network Services

Estimated time: 30 minutes
Objective 6.2: Given a scenario, troubleshoot network issues.
Resources:

- Public cloud account (such as AWS, Azure, or GCP)
- Internet access

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Description: Each cloud platform offers DHCP and DNS services a bit differently. To manage cloud networks, you need to understand where these and related networks services are configured for each cloud you work with. Complete the following steps:

1. Choose a cloud platform, such as AWS, Azure, or GCP. You can repeat these steps with a second or third platform later. Check with your instructor on the requirements for this Challenge Project.
2. If you need to create a project or resource group to manage project resources when you're finished, do so now.
3. Create a VPC or VNet in your selected cloud using your choice of private CIDR range. **Take a screenshot** showing the VPC's or VNet's CIDR settings; submit this visual with your responses to this project's questions.
4. Either explore the settings for your VPC or VNet, or research online to respond to the following questions:
   a. What steps are required to change the VPC's or VNet's CIDR block?
   b. What steps are required to change the VPC's or VNet's DNS servers?
   c. What steps are required to change the VPC's or VNet's NTP servers?
   d. What is the default NTP service for this cloud's VPCs or VNets?
5. After finishing your research, complete the following steps:
   a. Change the VPC's or VNet's CIDR block to **192.168.100.0/24**.
   b. Change the VPC's or VNet's DNS servers to Cloudflare's public DNS servers at **1.1.1.1** and **1.0.0.1**.
   c. **Take a screenshot** showing the updated configurations; submit this visual with your responses to this project's questions.
   d. What problems did you encounter in making these changes? How did you resolve these issues?
6. Delete any cloud resources you created for this project. Repeat the project with another cloud platform if desired or if required by your instructor.
7. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 5-16 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Solutions to Self-Check Questions

Section 5-1: Hybrid Cloud and Multi-Cloud Networking

1. Which of the following technologies represents a physical device in the customer's on-prem network?

   Answer: b. CGW

   Explanation: A **CGW (customer gateway)**, which supports the customer end of a VPN connection, might be a hardware VPN gateway device or a virtual VPN gateway application hosted on a server. A VGW (virtual gateway), which supports the cloud end of a VPN connection, is hosted in a VPC and provides access to resources running within that VPC. A cloud WAN provides a centralized platform for managing a large number of connections in widely dispersed geographical areas, such as across continents and regions. VLANs (virtual LANs) are logical configurations used to segment networks for both security and efficiency purposes.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. Which cloud connection type is the most expensive for the cloud customer?

   Answer: b. Direct connection

   Explanation: The **direct connection** option is more secure and also tends to be the most expensive option. The public Internet is the cheapest option for connecting to a cloud network. VPNs are less expensive than a direct connection and provide increased security when using the public Internet. A VLAN (virtual LAN) can be used on a direct connection to segment network traffic for increased security; adding VLAN segmentation does not typically increase the cost of a direct connection.

3. Which of the following characteristics is increased by a transit gateway?

   Answer: c. Connectivity

   Explanation: A transit gateway serves as the hub in a hub-and-spoke topology, which increases **connectivity** across large numbers of VPCs without adding an unreasonable amount of complexity. Transit gateways typically are designed to minimize latency by reducing the number of connections required to increase connectivity. A transit gateway likely has generous limits on throughput, though there are ways to increase these limitations.

4. Which of the following factors directly affects the cost of a direct connection?

   Answer: a. Bandwidth

   Explanation: Configured **bandwidth** is used to determine much of the cost associated with a direct connection. Routing and segmentation can be configured over a direct connection without affecting its cost. The amount of outgoing data affects cost, but incoming data often is not charged on direct connections.

Section 5-2: Extending Network Services

5. Which of the following devices is least likely to need a static IP address?

   Answer: b. Web client

   Explanation: Devices such as **web clients** can use dynamic IP addresses assigned from a DHCP server because they do not need to be located at the same address over time. Web servers, DNS servers, and default gateways all need static IP addresses so clients can consistently find them at the same address each time.

6. Which routing protocol is used across the Internet?

   Answer: b. BGP

   Explanation: BGP (Border Gateway Protocol) dynamically manages routes between autonomous systems, such as across the Internet. OSPF (Open Shortest Path First), RIPv2 (Routing Information Protocol, version 2), and EIGRP (Enhanced Interior Gateway Routing Protocol) are all interior gateway protocols.

7. Which device helps ensure HA for a web server?

   Answer: d. Load balancer

   Explanation: A **load balancer** distributes traffic over multiple components to optimize performance and fault tolerance, thereby increasing HA (high availability). DNS servers resolve domain names to IP addresses. A DHCP server assigns IP addresses to network hosts from an allocated IP address space. Routers direct traffic across networks.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8. Which type of device can load balance the highest volumes of traffic?

Answer: b. NLB

Explanation: When the workload needs to support high volumes of traffic with high scalability and low latency, an **NLB (network load balancer)** is usually the better choice because it can handle higher volumes of traffic than an ALB (application load balancer). In Azure, Application Gateway is an ALB service. In other contexts, the term refers to a service that manages user authentication to give users remote access to applications without requiring a VPN (virtual private network). A transit gateway connects many VPCs and other network types to each other.

Section 5-3: Extending the Cloud

9. Which feature of CDN files helps ensure customers get current copies of cached data?

Answer: c. TTL

Explanation: The CDN (content delivery network) service manages the **TTL (time to live)** of files cached at edge locations to ensure updated files are requested from the origin point if the files have been sitting in the cache for too long. Providing files in closer proximity to users will reduce latency but does not ensure currency of provided files. The origin point's storage location is irrelevant to data currency because files are provided through cache locations, not directly from the origin. The access point's URL (uniform resource locator) provides a central point of communication to send requests, but it has no effect on how current CDN files are.

10. Which IoT messaging protocol is most known for providing security?

Answer: a. AMQP

Explanation: AMQP (Advanced Messaging Queuing Protocol) is an open-source messaging protocol that is known for providing better security than some of the other options. MQTT (Message Queuing Telemetry Transport) is a particularly lightweight messaging protocol that is not so focused on security and requires another protocol to secure messages. CoAP (Constrained Application Protocol) lacks built-in security features, instead relying on some other protocol to layer on security. LoRaWAN (long-range WAN) is a transmission protocol at layer 1 of the OSI model, not a messaging protocol.

Section 5-4: Troubleshooting Cloud Connectivity

11. Which protocol requires port 22 be open on a firewall?

Answer: b. SSH

Explanation: For **SSH (Secure Shell)** to work, TCP (Transmission Control Protocol) port 22 must be enabled on any intervening firewalls. RDP (Remote Desktop Protocol) uses port 3389. ICMP (Internet Control Message Protocol) and ARP (Address Resolution Protocol) don't use ports in this manner.

12. Which device is used to camouflage internal network resources from the outside world?

Answer: b. Proxy server

Explanation: A **proxy server** acts as an intermediary between external and internal networks, screening all ingress and egress traffic. A load balancer distributes traffic over multiple components to optimize performance and fault tolerance. A DNS server translates domain names to IP addresses. A default gateway provides network hosts with access to resources outside the local network.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 6



Molibdenis-Studio/Shutterstock.com

Securing Cloud Resources

Chapter 6 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Describe common threats to and configurations for cloud security.
2. Secure cloud networks.
3. Secure cloud compute resources.
4. Secure cloud-hosted data.
5. Troubleshoot cloud security issues.

Chapter 6 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 6-1: Cloud Security Configurations
- Threats to Cloud Security
- Vulnerability Management
- Shared Responsibility for Cloud Security

Section 6-2: Virtual Network Security
- Allow and Deny Rules
- Zero Trust Security
- AWS VPCs and Subnets
- Azure VNets
- GCP VPCs
- Securing Hybrid and Multi-Clouds

Section 6-3: Compute Security
- Endpoint Protection
- Application Security
- Pipeline Security

Section 6-4: Data Security
- Encrypting Data
- Key Management

Section 6-5: Troubleshooting Cloud Security
- Security Misconfigurations
- Key Management Issues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 6 Scenario "What's up, dude? Did you catch the game last night?" Malik gives Dianne a fist bump as the two of you walk into the guard station at the front entrance of your school's campus, which also hosts the security department's main office.

"Aw, man, that game was brutal," groans Dianne. "I bet Kendra's giving you a hard time today! But there's still time to make the playoffs." Dianne and Malik, both Cubs fans, show tireless optimism the entire baseball season. Kendra, a Cardinals fan, enjoys bantering with them. But they insist the Cubs' World Series win has fueled their hope for another hundred years, if that's what it takes. You've not yet decided if this is a sign of admirable loyalty or mild delusion.

You spot two large monitors on Dianne's desk, each with several camera feeds that slowly rotate among different cameras throughout campus. "Hey, so this is where the security camera feeds go?" you ask her.

"Yep, this is the grand view of the whole campus," replies Dianne. "We've got cameras in and around every building, even the football field."

"You said you're having a problem with some of them, right?" asks Malik.

"Right," Dianne says. "The new cameras in the science building—we're not getting a feed from those here at this station. When I plug my laptop into the system there in the building, everything appears fine. But back here, we're not getting anything. Any idea what's wrong?"

"We'll look into it and get it fixed for you," you reply with a little more confidence than you actually feel. It seems like security cameras should be such a simple thing, and yet those new cameras were tough to install.

After you and Malik get back to the IT office, the two of you check in with Kendra for a little advice on how to troubleshoot the security cameras before heading to the science building. As she gears up to gloat over the Cardinals' win last night, Malik cuts her off with a question: "You know, I've been wondering about how to keep our stuff safe in the cloud. We keep everything behind a firewall here on campus. What happens when we start moving stuff to the cloud?"

"Ah, yeah. That's a good question," muses Kendra. "Well, it's kind of like security here on campus," she begins. "We have the guard station at the front that is kind of like a firewall. And like with on-prem firewalls, there are ways to control what traffic gets into your cloud network. But a *lot* of people come onto this campus, and not everyone can be closely and always tracked. We need good security throughout campus to alert us if something goes wrong."

"Oh, like the security cameras?" you suggest.

"Right," says Kendra. "And locked doors, and passwords on computers, and security personnel patrolling the campus. At your home, you don't need such dense security because it's a private space—you have more control over who is in your home. But our campus is more communal, like the cloud. There are security measures throughout campus, just like in the cloud. We can add different layers of security that protect our cloud resources in different ways."

"What are some of those other security tools?" you wonder out loud.

"I smell a research project coming on," quips Malik.

"You read my mind!" grins Kendra. "Here ya go—do a little research on these questions," she adds as she scribbles a few notes on her notepad:

- What kinds of threats do cloud users face?
- What techniques can keep cloud networks, VM instances, and data secure?
- What tools can help in troubleshooting problems with cloud security?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 6-1: Cloud Security Configurations

Certification

1.1 Given a scenario, use the appropriate cloud service model.

2.3 Summarize aspects of cloud migration.

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.1 Explain vulnerability management concepts.

4.4 Given a scenario, apply security best practices.

4.6 Given a scenario, monitor suspicious activities to identify common attacks.

6.3 Given a scenario, troubleshoot security issues.

When considering a move to the cloud, technical and nontechnical professionals alike often list security concerns as their number one reason for hesitating. Cloud consumers face both familiar and new threats to their data and other cloud resources. While their concerns are warranted to a degree, the technologies supporting cloud security continue to evolve and mature. By understanding the nature of the relevant threats and the tools used to protect the cloud, IT professionals can make informed and effective decisions as their reliance on cloud technologies continues to increase. This section explores these common threats and popular tools used to mitigate related risks. Later in this chapter, you learn about ways to protect cloud networks, compute resources, and cloud-hosted data.

Threats to Cloud Security

To begin, let's first examine a few of the most common internal and external threats to cloud security:

- **Data breaches**—Any data accessible from the Internet, no matter where it's stored, is inherently at risk. Data stored in the cloud has some unique vulnerabilities, however, including risks due to shared hardware resources and vulnerabilities to CSP personnel or third-party providers. Cloud consumers also have no control over physical security and little influence on configuration of the physical infrastructure supporting their cloud-hosted data. For example, when deleting data in the cloud, the consumer has no way of knowing whether the data continues to exist on hardware outside their control, and this data could be recovered in an attack. The outcome of a data breach, regardless of the cause, can be devastating to a business, its customers, and its employees due to loss of privacy, risk of fraud, and resulting fines, lawsuits, and possibly criminal charges.
- **DoS and DDoS attacks**—A **DoS (denial of service) attack** is any kind of attack that affects the availability or performance of services or data. For example, an economic DoS attack might artificially escalate the demand on a start-up company's auto-scaled cloud resources, increasing the company's bills beyond its ability to pay. Sometimes multiple attackers (or bots) are used in a **DDoS (distributed denial of service) attack**, which is a DoS attack orchestrated from many sources at once. For example, an orchestrated attack from a botnet of IoT (Internet of Things) devices might flood a web server with traffic until the server is no longer able to respond.
- **Malware**—**Malware (malicious software)** is any program or piece of code designed to intrude upon or harm a system or its resources. Included in this category are viruses, Trojan horses, worms, bots, and ransomware. **Ransomware** is malicious software that encrypts data or locks a device until a ransom is paid. The accessible nature of the cloud makes cloud-hosted resources more vulnerable to malware infection. Accessibility via APIs (application programming interfaces), shared hardware, and complex configurations increase the attack surface of cloud resources, requiring added vigilance on the part of cloud administrators.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

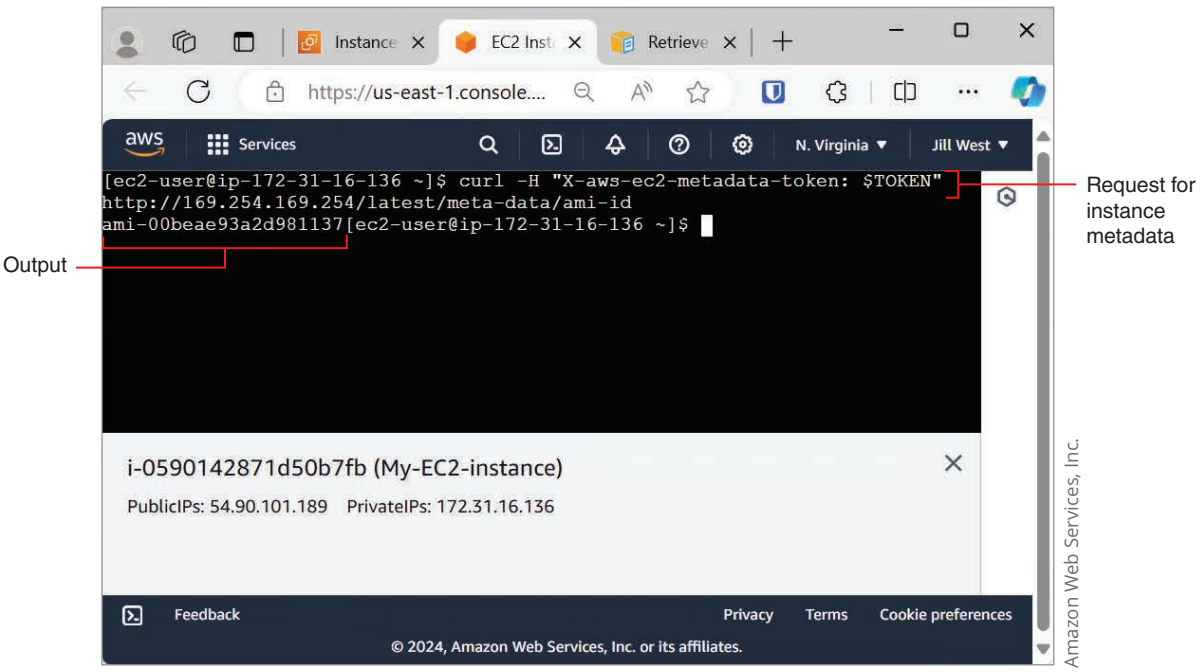
- **Software vulnerabilities**—Software running in the cloud can have security weaknesses and vulnerabilities from misconfigurations, poor access control, insecure APIs, missing updates, data leakage, or other issues. Software is also vulnerable during development.
- **Insecure interfaces**—Whether a UI (user interface) or an API, points of transition from one system to another can leave security loopholes and access points. Designed for access and interactivity, UIs and APIs are both prime targets for hackers attempting to gain entry.
- **Advanced persistent threats**—APTs (advanced persistent threats) are attacks that clear the way for future attacks, essentially gaining a foothold in the target's IT systems and infrastructure. APTs are long-term projects where the infiltration blends with normal network traffic so as not to raise red flags. These efforts are more commonly undertaken by large enterprises or government organizations due to the complexity of the attack, the required resources, and the type of data that is targeted.
- **Account hijacking**—While account hijacking is not a unique cloud computing risk, having a cloud account hijacked can result in some cloud-specific damage due to the software-defined nature of cloud resources, especially when that cloud account has administrative privileges. The cloud environment allows a hacker greater leverage for eavesdropping on activities, manipulating data, redirecting customer traffic, reconfiguring cloud resources, and collecting payment credentials such as credit card numbers.
- **Poor account management**—Poorly managed identity and account systems increase the likelihood of an account being compromised. Weak password rules; lack of MFA (multifactor authentication); improper storage of account credentials; inadequate account life cycle oversight; and sloppy handling of keys, passwords, and certificates all put a company and its users at risk.
- **Social engineering**—**Social engineering** is the strategy of manipulating social relationships to gain access to sensitive information and systems. **Phishing** is a well-known form of social engineering where communications that appear to come from a legitimate source request access or authentication information, such as user passwords. Phishing emails can be extremely effective, especially when augmented by AI (artificial intelligence) technologies to design the emails or manage their transmission. Other types of social engineering attacks include baiting, quid pro quo, and dumpster diving.
- **Human error**—People make mistakes. Whether from lack of knowledge, carelessness, or willful sloppiness, simple mistakes can create huge problems. User carelessness specific to the cloud includes file oversharing, storing protected information in insecure file storage accounts, and designing problematic file names (such as naming a password spreadsheet "passwords"). Any of these situations could create some serious legal and economic consequences for their employer.
- **Insider threats**—Beyond user error, malicious users pose an arguably greater threat. This threat could come from current or former employees, contractors, or business partners who intentionally use company resources beyond acceptable-use standards in a way that damages confidentiality, integrity, or availability of company data and services.
- **Cryptojacking**—**Cryptomining** is the typically legitimate process of running calculations to enable blockchain transactions, which earns cryptocurrency. This process requires a large amount of physical or virtual compute resources. **Cryptojacking** is the illegal use of someone else's compute resources for cryptomining, which financially benefits the attacker. Considering how much compute power is available in the cloud, cloud platforms are a prime target for cryptojacking. Some monitoring services can be used to help identify and defend against cryptojacking. For example, the attacker typically must obtain compromised credentials, elevate access permissions, and establish an attack infrastructure. This might include using the victim's billing details to run the attacker's compute resources or taking control of compute resources within the victim's account (a strategy called "living off the land"). Monitoring systems can evaluate for the access and usage patterns common to these attacks and alert administrators that an attack is suspected or confirmed.
- **Zombie instances**—Cloud resources that are created and then abandoned without being deleted are often called **orphan resources**. These resources continue to cost a company money even though they're not being used because it's so easy to create resources and then forget to clean up what is no longer needed. In addition to being costly, these orphaned resources also present security vulnerabilities—they can be taken over by attackers who use them for their own purposes, such as cryptojacking.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Many of these outdated systems and software are not monitored, analyzed, or regularly updated with needed security patches, leaving them vulnerable to hostile takeover. Additionally, many of these resources are created in the process of testing or contain misconfigurations and, therefore, possibly have elevated privileges that attackers can use for their own benefit. If compromised, these resources are called **zombie resources** or zombie instances. Monitoring tools can help identify unused, untracked, or unpatched resources. Many of these tools offer visualizations that make it easier for human admins to quickly identify outliers to the intended infrastructure. Scripts also can assist in this process by identifying unused or underutilized resources, such as unattached storage volumes; unused VMs, snapshots, or load balancers; security groups with problematic rules (such as remote access ports open to all IP addresses); and user accounts that have not been recently or regularly accessed. Regular auditing and compliance reporting can also help to identify these kinds of risky, costly resources.

- **Metadata**—**Metadata** is data about data; in other words, it's information about resources that describe the resources, how they're used and configured, and what they can do. Some of this information is added automatically when the resource is created, and additional information can be added with tags. For example, when you name a VM instance in the cloud, its name is stored as a tag and is part of its metadata. You can add additional tags, such as identifying a project or billing account the VM is part of, to expand the metadata on each instance. A lot of an instance's metadata, such as the image used to build the instance, is available from within a VM instance. Figure 6-1 shows an example of this metadata retrieved from the instance's CLI. Access to this instance metadata is not authenticated and could be retrieved by anyone with direct access to the instance or possibly by code running on the instance. Many metadata fields in the cloud platform itself, such as tags, are not encrypted and don't require any particular authorization process to allow access to the data. A great deal of metadata can be harvested about cloud-hosted resources, including configuration details (such as the image used), comments documenting configuration changes, company infrastructure or personnel information, timestamps, logged changes, maintenance windows, open ports, and available APIs. While this metadata can be especially helpful for cloud admins and effective monitoring systems, it can also be highly informative for attackers. Sensitive information should never be stored in metadata fields. For example, don't use customer names to label individual cloud resources. Documents should be checked for metadata before they're posted or shared with others. Metadata management software can monitor metadata on files stored in or shared through the cloud. Strict access controls can, in some cases, limit who can view metadata.

Figure 6-1 Instance metadata accessed from the instance's command line



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Data loss with no backups**—Just because data lives in the cloud doesn't mean it's being backed up. With the current trend toward relying on the cloud to store large volumes of data, it's easy to overlook the need to back up cloud-hosted data. Several things can go wrong in this scenario: There could be problems with backup syncing, a cloud service could crash, accounts that access the data could be compromised, or accounts can be deleted for inactivity.
- **CSP changes or outages**—Capabilities and conditions change, and SLAs (service-level agreements) are occasionally adapted to keep up with the changes. However, these changes aren't always made in a cloud customer's favor, which can affect the ways they manage their data and other cloud resources. When relying on a service provider, the organization is susceptible to changes and outages experienced by the CSP. For example, a December 2021 power outage at an AWS data center in the North Virginia region widely affected services for many popular websites and companies, including the workplace communications site Slack and the media streaming site Hulu. Due to the reliance of so many services and websites on a few major CSPs, a single cloud outage can have ripple effects throughout many industries.

Note 6-1 | Most CSPs provide a service monitoring site to check availability reports for their platform. For example, status.aws.amazon.com shows the current status of all services across the AWS platform in all regions. Broad coverage monitoring sites, such as Downdetector, report on outages for hundreds of ISPs, CSPs, SaaS providers, social media platforms, and more. Some of these reporting sites are more accurate or timely than others.

Monitoring for Suspicious Activities

Cloud providers make available a wide variety of tools to monitor for security threats and, in many cases, automate responses to shut down attacks and recover from any outages or other damage caused by an attack. Considering that an attack can do a great deal of damage in a very short period of time, automated responses can greatly decrease the impact of these attacks. Human admins require a significant amount of time to receive notification of a detected attack (especially those companies that cannot afford on-duty security personnel 24 hours a day), collect needed information about the incident, determine the appropriate response, and then execute those tasks. In contrast, automated systems can respond almost immediately. The challenge, however, is that automated systems must be correctly configured to reliably interpret what is happening and what response is appropriate. For this reason, few security monitoring services in the cloud are one-size-fits-all services that cover all security monitoring needs. Instead, security monitoring services tend to be focused on particular types of monitoring. For example, consider the following AWS services to understand the kinds of security monitoring and automated response available in the cloud:

- **GuardDuty**—This service incorporates AI and ML features to continuously monitor a customer's cloud deployment for signs of security threats, such as malware. GuardDuty can monitor deeply into a variety of cloud resources, including storage buckets, container deployments, serverless functions, VM instances, attached storage volumes, VPC flow logs, and databases.
- **Inspector**—This service scans VM instances, containers, and serverless functions for vulnerabilities in their software or network configurations. Inspector assigns vulnerabilities a risk score to assist in prioritizing remediation efforts.
- **Security Hub**—This service collects security alerts from other services to present this information in a centralized dashboard. This kind of activity can result in SOAR (security orchestration, automation, and response) workflows. Security Hub also monitors for certain security best practices, alerting the cloud customer if cloud resources fail these security checks. This kind of activity is called CSPM (cloud security posture management).
- **Detective**—This service collects security data from various services for further analysis and visualization. Detective can be especially helpful in investigating the root cause of an issue.

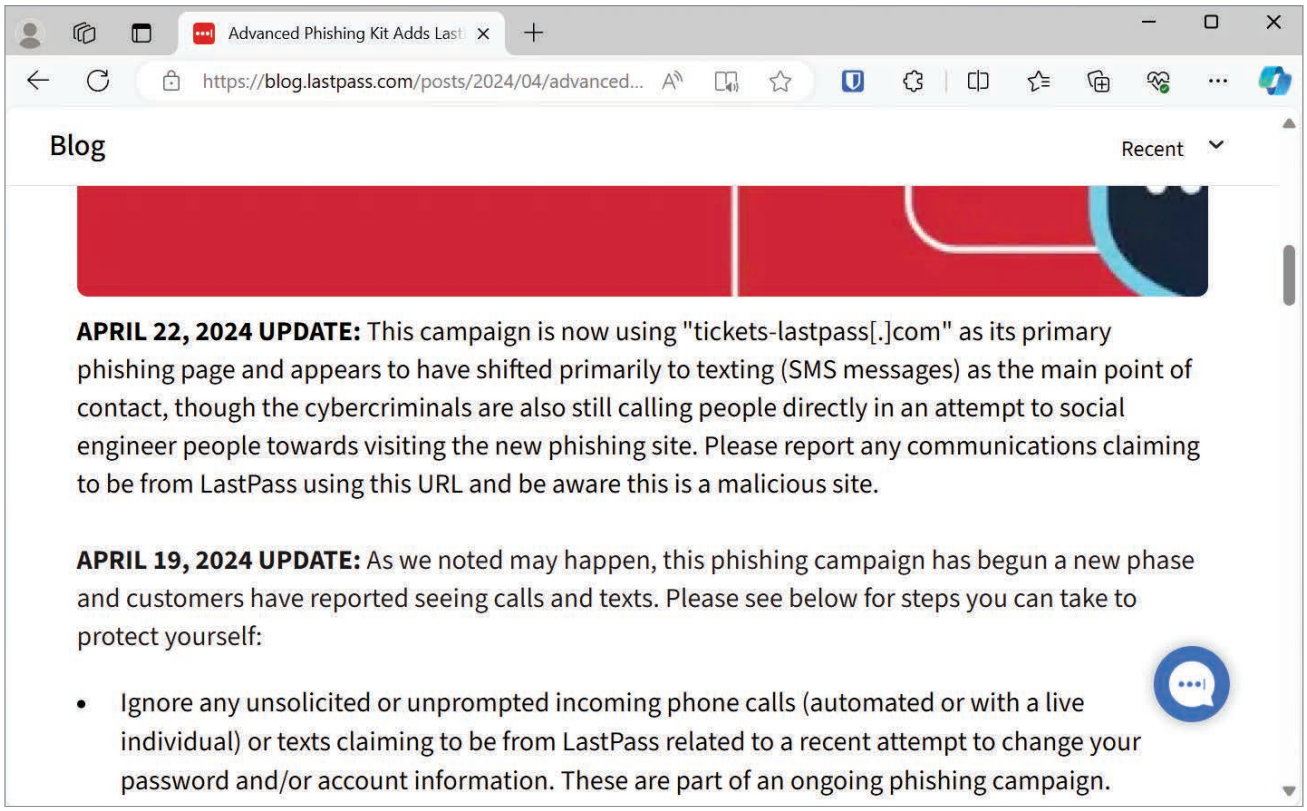
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Security Incident Response**—This service can be used to automate security incident responses and recovery. It collects information from GuardDuty and Security Hub, streamlines security escalation and response, and provides centralized access to a variety of tools needed during a response.

Some kinds of problematic activities are harder to track than others. For example, sophisticated phishing emails can escape detection by automated monitoring tools, especially as generative AI is increasingly used to write and manage these email campaigns. However, phishing email activity can also be tracked via user-reported emails. Your organization might provide a "Report Phishing Email" tool in your email client, which directs suspicious emails to the security team for further evaluation.

Companies also can track phishing activity related to their network or web domains. For example, LastPass tracked a series of phishing campaigns in 2024 where the threat actors created multiple domains that appeared to be related to the official domain of lastpass.com, such as help-lastpass.com. Other companies, such as Okta and Microsoft, were also targeted by similar campaigns. LastPass was able to monitor activity to the spoofed domains, report the domains for misuse, and have them taken down. Continued monitoring revealed additional misleading domains. LastPass conducted their own campaign to inform and educate their customers (Figure 6-2), as well as solicit reports from customers about ongoing phishing activities.

Figure 6-2 LastPass customer updates on phishing campaign



Source: LastPass US LP

Applying Concepts 6-1: Research Data Breaches

Estimated time: 15 minutes
Objective 4.6: Given a scenario, monitor suspicious activities to identify common attacks.
Resources:

- Internet access
- If on a smartphone, a generative AI app, such as ChatGPT or Perplexity

Description: In this chapter and the next, you learn about cloud-based access control methods and some of the risks, concerns, and technologies involved in controlling access to cloud networks, data,

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and other resources. Unfortunately, these methods don't always work as intended, especially when improperly implemented. That's one way that data breaches happen. These breaches can be costly to the organizations involved and can cause significant damage to customers, clients, patients, and users.

Many websites report on major breaches throughout the world. One website in particular allows you to look up your own email address to determine if it's been posted along with other hacked data on the web: Have I Been Pwned (haveibeenpwned.com). "Pwned" is an Internet slang term for having been taken advantage of, and it's pronounced "powned." (The term supposedly came from a frequent mistyping of the word "owned" because of the proximity of the *o* and the *p* on the keyboard.) This website also keeps an ongoing list of the websites whose data has been hacked, along with a brief historical description of the breach. To learn more about recent data breaches, complete the following steps:

1. Visit the webpage at **haveibeenpwned.com/pwnedwebsites** and skim the list of breaches. Choose three of these breaches that most interest you.
2. In your own words, describe what happened in each breach.
3. For each breach you chose, suggest one or more ways similar breaches can be prevented in the future. Give at least one different, thoughtful suggestion for each breach. Provide supporting evidence of why you think this protection measure will be successful in a future, similar situation.
4. Use a GenAI chat tool, such as ChatGPT or Perplexity, to evaluate each of your suggestions for protections. For each breach, give the GenAI tool the information about the breach from the Have I Been Pwned website, and then give the chat tool your protection proposal. Ask the chat tool how effective the protective strategy would be to prevent this type of attack in the future. What response did you get for each proposal?

Vulnerability Management

Vulnerability management is the process of identifying, assessing, and addressing security vulnerabilities in a system. A variety of scans can be used to identify potential vulnerabilities in software, computers, or networks. A **scanning scope** determines the boundaries of a scan, listing the scan's targets, objectives, and depth. This statement can ensure clarity for all involved parties regarding which systems will be scanned, what types of scans will be conducted, and what kind of information is expected to be collected. This statement can also ensure scanning efforts will cover all systems that should be covered so nothing important is missed. For example, a scanning scope might state the investigation will scan all devices in a certain network segment.

As you collect information on the vulnerabilities in your network, you will likely need to track this information in a risk register. A **risk register** is a project management tool that can be used to document cybersecurity and other types of risks, background information on these risks, mitigation strategies specific to each risk, plans for responding should a vulnerability be attacked, and team members with responsibilities related to each risk. Documentation will also include historical data on previous attacks and the related responses and outcomes.

During assessment, risks are rated according to the likelihood of the threat happening, the expected impact to the business, and the cost to fix the problem. This information is also included in the risk register and is used to inform decisions on which issues to address first. Risk ratings are often informed by a standardized catalog of known vulnerabilities called the **CVE (Common Vulnerabilities and Exposures)**. Maintained by the MITRE Corporation, the project historically has been funded by the U.S. Department of Homeland Security. Ironically, at the time of this writing, the organization is working to change this arrangement to eliminate the inherent risks imposed by a single funding stream. The CVE list is free to access, use, and analyze, but only MITRE can make changes to it. While the CVE doesn't contain details of the vulnerabilities themselves, it does provide a tracking system that is used by vulnerability databases. Each vulnerability receives a standardized identifying number, which makes it easier to track vulnerabilities across systems (Figure 6-3). You can learn more about the CVE and download your own copy at cve.org.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

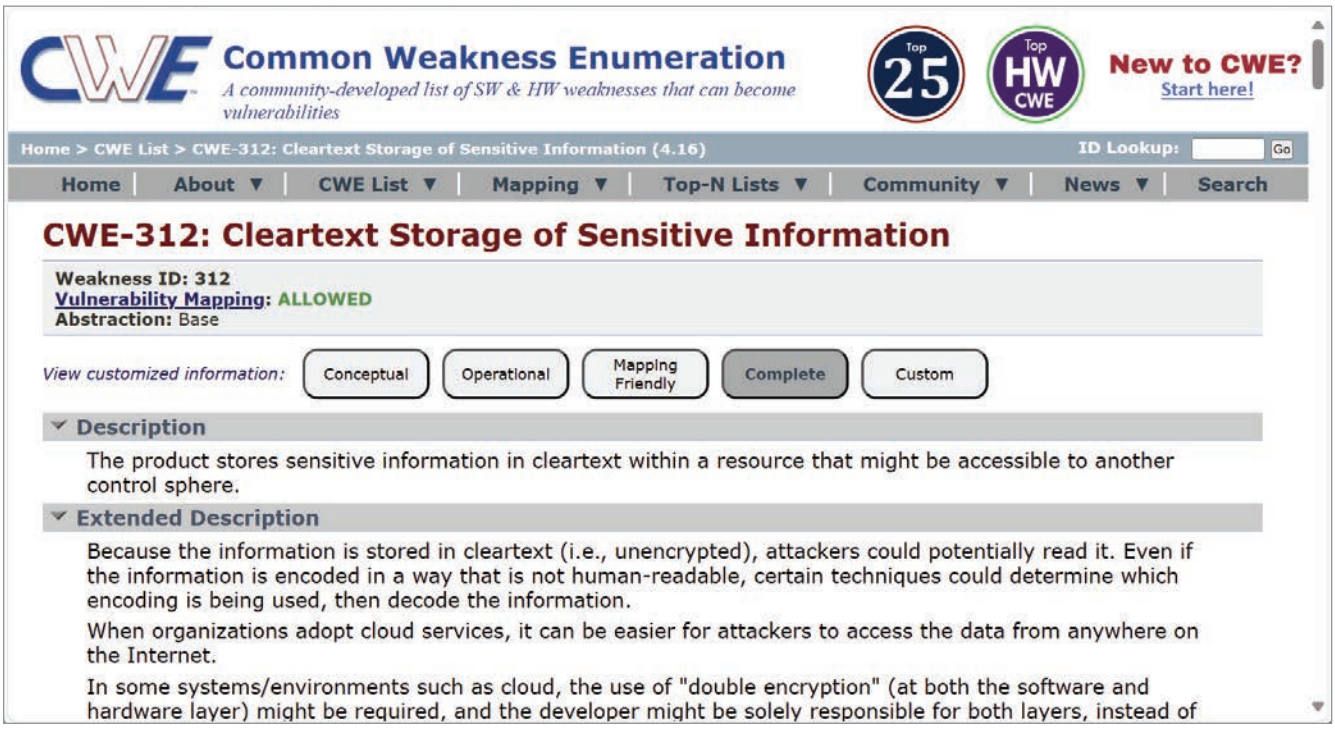
Figure 6-3 An example vulnerability in the CVE



Source: The MITRE Corporation

A similar resource—also maintained by the MITRE Corporation—is the **CWE (Common Weakness Enumeration)**, which is a standardized taxonomy of software, firmware, hardware, and service component weaknesses. A weakness is a condition that is a potential precursor to a vulnerability. For example, storage of sensitive information in plaintext is a weakness (Figure 6-4). A specific mobile app that incorporates this weakness by storing users' credit card numbers in plaintext is a vulnerability in real-world software.

Figure 6-4 An example weakness in the CWE



Source: The MITRE Corporation

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Both vulnerabilities and weaknesses can be rated in terms of severity. The **CVSS (Common Vulnerability Scoring System)** is a standardized framework for rating vulnerabilities on a scale from 0 to 10, with higher scores representing more severe vulnerabilities. Similarly, the **CWSS (Common Weakness Scoring System)** is a standardized framework for rating weaknesses in three metric groups: Base Finding (basic characteristics of the weakness), Attack Surface (options for exploiting the weakness), and Environmental (potential impact on the system's environment or other context, such as business impact).

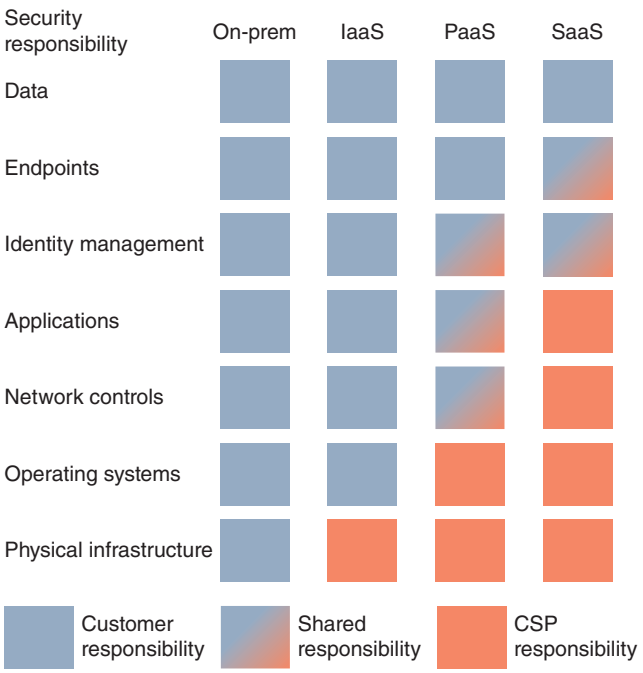
Not all vulnerabilities can be eliminated. In responding to discovered vulnerabilities, you must decide among one of the following options:

- **Remediation**—With this option, the vulnerability is somehow eliminated. For example, you might patch a system, update the code, change system configurations, or replace the system.
- **Mitigation**—With this option, some kind of compensating control is applied. For example, a firewall can block malicious traffic. Employee training can reduce the threat of phishing emails. And access controls can restrict access to sensitive systems or data.
- **Acceptance**—With this option, the risk caused by the vulnerability is low enough or unlikely enough that the organization decides to leave the vulnerability unchanged.

Shared Responsibility for Cloud Security

With new challenges come new innovations. Cloud computing provides a prime example of how security innovations are solving new problems. Cloud security works according to the **shared responsibility model**, meaning the CSP is partially responsible for a cloud's security, and the customer is responsible for the rest of it. However, this division of labor can shift depending on what type of cloud service the customer is using. Figure 6-5 compares this shared responsibility according to the type of cloud deployment.

Figure 6-5 Shared responsibility model



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A theoretical framework borrowed from project management can help further define who is responsible for what. The **RACI (responsible, accountable, consulted, informed) matrix** identifies role and responsibility assignments for team members or groups. Consider the following responsibility types:

- **Responsible**—Responsible parties perform the work.
- **Accountable**—Accountable parties make sure work is done to the required standards.
- **Consulted**—One or more parties are consulted for their opinions or expertise in relation to the work.
- **Informed**—Some parties should be kept informed of progress and notified when the work is complete.

In the context of cybersecurity, consider how these roles can define stakeholder responsibility and interest. For example, the CISO (chief information security officer) might be accountable for the security of a particular resource or system, and another party (such as a cloud security engineer) might be responsible to actually perform the work to secure that system. Similarly, yet another party (such as a cloud security consultant) might be consulted on how best to secure a system, and others (such as auditors or other members of the executive suite) might need to be informed of a system's security status and implementations made to ensure that security.

These roles can also be spread across multiple organizations, such as when using a cloud provider or a managed service provider. For example, consider how AWS is responsible for the physical security of its data centers. And yet, as the cloud customer, you must have access to certain documentation informing you of AWS's security measures in a way that allows your cloud deployment to also pass regulatory or industry security inspections.

As these roles are being defined for various stakeholders, a matrix is developed that visualizes each party's role for various phases or aspects of a project. For example, consider the RACI matrix in Figure 6-6 that illustrates the roles of various stakeholders in a cloud deployment's security.

Figure 6-6 RACI roles define stakeholder responsibility levels

Tasks	Architect	Administrator	Developer	Project manager
Security policies	C	R	I	A
Host security	R	A	I	I
Application security	R	I	A	I
Network security	R	A		I
Incident response	C	R, A		I

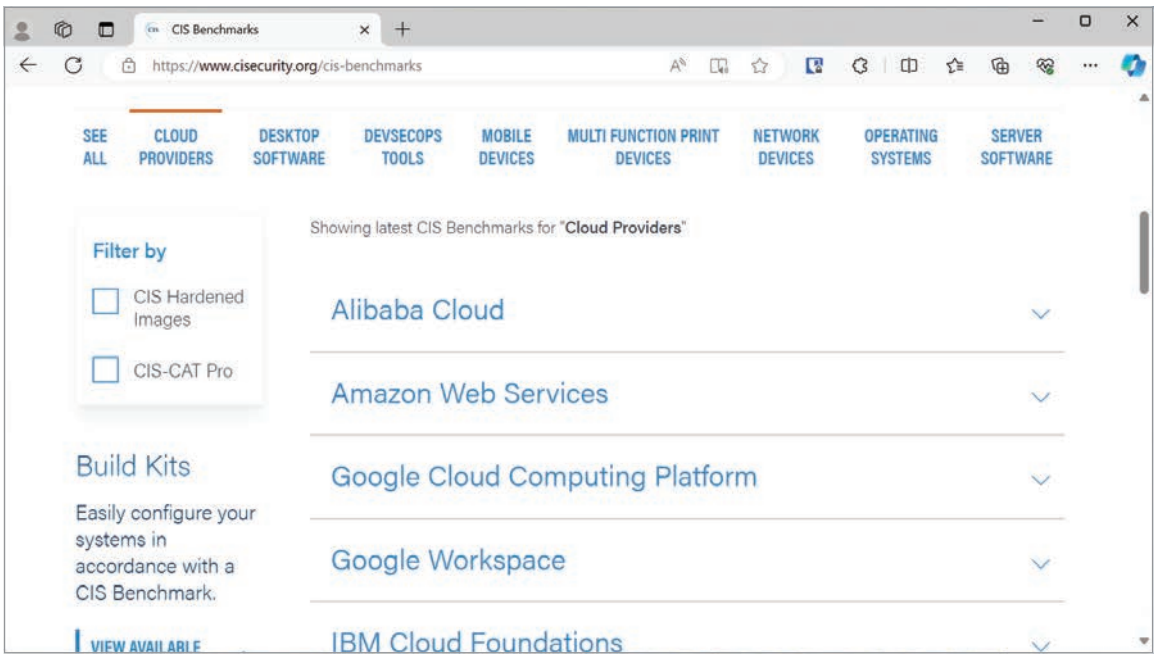
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exam Tip | If you plan to take the CompTIA Cloud+ certification exam, be prepared to answer questions related to how RACI roles can define the shared responsibility relationship between the cloud provider and the cloud customer. For example, consider what combination of RACI roles would reflect shared responsibility for the security of a particular resource or service.

Cloud Security Industry Standards

CIS (Center for Internet Security) is a nonprofit organization that draws on expertise from IT professionals in government, business, and academia to define best practices for cybersecurity. Many of their guidelines are organized into **CIS benchmarks** that specify secure configurations of various technologies categorized by vendors, such as AWS, Microsoft, and GCP (Figure 6-7). For example, one CIS benchmark specifies what hardening configuration changes should be made to Amazon Linux to ensure high confidence in its security, along with instructions on how to make the recommended changes. You can learn more about vendor-specific benchmarks at cisecurity.org/cis-benchmarks.

Figure 6-7 List of CIS benchmarks



Source: Center for Internet Security

Vendors also often post additional information about CIS benchmarks customized for their platforms within their own documentation, and sometimes this information can be monitored through one or more security services. For example, AWS's Security Hub supports monitoring of the CIS AWS Foundations Benchmark. Some of the security controls listed in this benchmark include the following:

- Each AWS account should list the customer's contact information to use during a security incident.
- No inbound or outbound traffic should be allowed in the default security group for a VPC.
- Every VPC should have flow logs enabled.
- No NACL should allow incoming traffic on port 22 or port 3389 from 0.0.0.0/0.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

AI in IT | AI (artificial intelligence) is used extensively to secure cloud-hosted systems and resources. Many cloud security services incorporate AI to detect anomalies that indicate possible threats, analyze those threats, and automate threat response. AI also is used to analyze logs and other extensive data (such as user behavior) to predict attacks, automate compliance, enforce security policies, and improve auditing processes. AI-powered security tools can also dynamically adjust security policies in response to real-time activities.

Self-Check

1. Which of these threats could be mitigated with good backups?
   - **a.** Data loss
   - **b.** Insider threat
   - **c.** Account hijacking
   - **d.** Data breach
2. Which cloud service model places the greatest responsibility on the cloud customer for cloud security?
   - **a.** RaaS
   - **b.** PaaS
   - **c.** IaaS
   - **d.** SaaS
3. Which source would provide an ID for a vulnerability in a specific web app?
   - **a.** CVE
   - **b.** CVSS
   - **c.** CWE
   - **d.** CWSS

Check your answers at the end of this chapter.

Section 6-2: Virtual Network Security

Certification

1.3 Explain cloud networking concepts.

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.4 Given a scenario, apply security best practices.

4.5 Given a scenario, apply security controls in the cloud.

4.6 Given a scenario, monitor suspicious activities to identify common attacks.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hosting resources and services in the cloud presents a slew of new security challenges. In response to these changing needs, many technologies have emerged to help protect cloud resources from external attacks. If you consider cloud resources in layers—platform, networks, instances, applications, and data—think about where DiD (defense-in-depth) strategies can be applied at each of these layers.

There are various perspectives on how best to secure each layer of resources in the cloud. Many organizations end up with a patchwork system that uses pieces of one security tool and pieces of another. This is dangerous in that it can leave gaps in coverage. It's also difficult to monitor feedback from multiple security systems. Often, a better approach is to use a unified, end-to-end cloud security system. This often means relying on a CSP's built-in security tools, which helps eliminate compatibility issues, increases visibility into your cloud infrastructure, collects monitoring activities in a central location, and scales more efficiently as business needs expand. New challenges arise when blending multiple cloud platforms, but even in these circumstances, CSPs and third-party providers are increasingly providing security solutions that bridge the gaps for hybrid and multi-clouds.

In these next few sections, you explore some of the built-in tools for securing cloud resources hosted in AWS, Azure, and GCP. In a later chapter, you learn more about how to secure access to the cloud platform itself through identity and access management. Let's begin with an analysis of how to secure cloud networks.

Allow and Deny Rules

You learned about some layers of VPC (virtual private cloud) security when you studied techniques for controlling a subnet's access to the Internet. For example, in AWS, you learned that a private subnet is logically isolated from the Internet in that it does not have direct access to an Internet gateway. You can also control whether an instance has a public IP address and what routes allow traffic between subnets or other gateways, such as a VPN (virtual private network) gateway. VPNs themselves can control traffic that reaches a subnet and encrypt traffic as it crosses a WAN (wide area network) connection. Additionally, load balancing can filter the traffic allowed to reach a subnet from the open Internet.

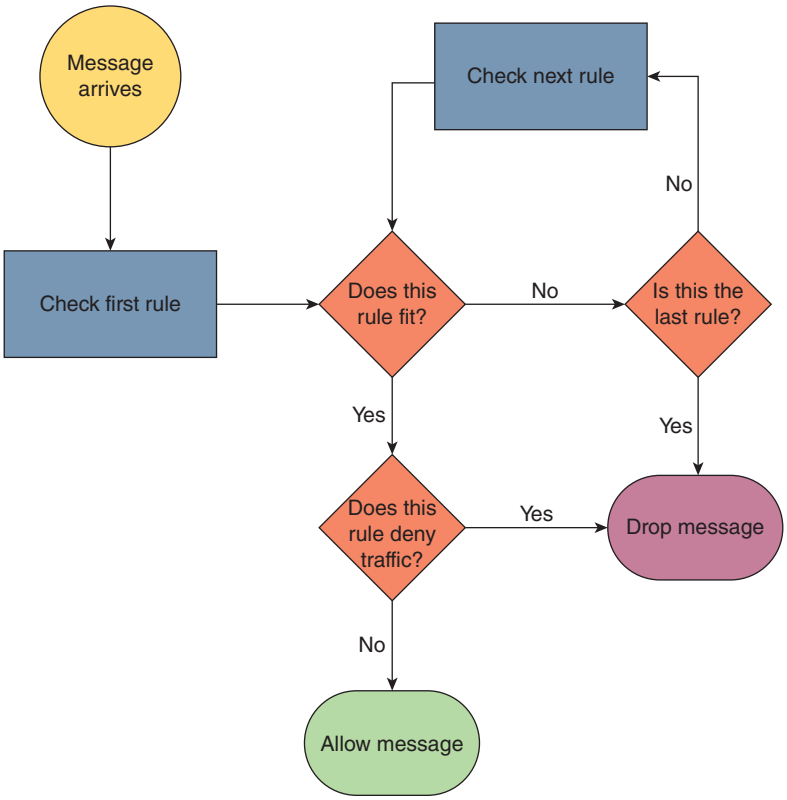
Cloud platforms also typically offer built-in security at the platform level that provides significant protection to your cloud resources. For example, AWS, Azure, and GCP all include always-on DDoS protection at no additional cost that the cloud consumer does not need to configure. This way, all resources on these platforms reside in an environment that is covered by these protective measures. Paid tiers of DDoS protection typically include enhanced features such as attack detection, flow-based monitoring, real-time alerts, customized configurations, cost protection, and highly available customer support during an attack.

At the same time, there are many security techniques that you, as the cloud consumer, must apply, configure, and manage without relying on the CSP to ensure all levels of security. Although not a true perimeter, traffic to and from a virtual network can be managed using various techniques, depending on the platform. A common component used by many of these security techniques is access rules that allow or deny traffic, similar to what you might find on a firewall device. Before digging into the ways each platform configures and uses these rules, it's helpful to understand on a basic level how allow and deny rules work.

When a message crosses a firewall (physical or virtual), the firewall checks its rules to determine whether the message is allowed to pass. Earlier in this course, you learned that routes are applied to a message according to which route in a route table most closely matches the message's destination. With firewalls, however, rules are usually applied in order of priority. The highest-priority rule is checked first. If the message is allowed or denied according to that rule, no further rules are checked. If the rule doesn't apply to that message, the firewall continues to the next rule. This process continues until no rules remain. The diagram in Figure 6-8 illustrates how this process plays out.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-8 Firewall rules are checked in order of priority to determine whether a message is allowed to pass



These rules check transmitted packets for information such as port, protocol, and IP address or CIDR range. If the information in the packets matches a rule's defined parameters, the message is allowed or denied according to the matched rule. An allow rule lets traffic through while a deny rule prevents traffic from passing. Firewall rules are configured differently for inbound and outbound traffic. In other words, there's a different list of rules for inbound traffic, which is typically more restrictive, than there is for outbound traffic. Security rules configured in various cloud services also provide high availability in that you don't have to manage firewall instances or devices, and you don't have to provide backups for those firewalls.

Zero Trust Security

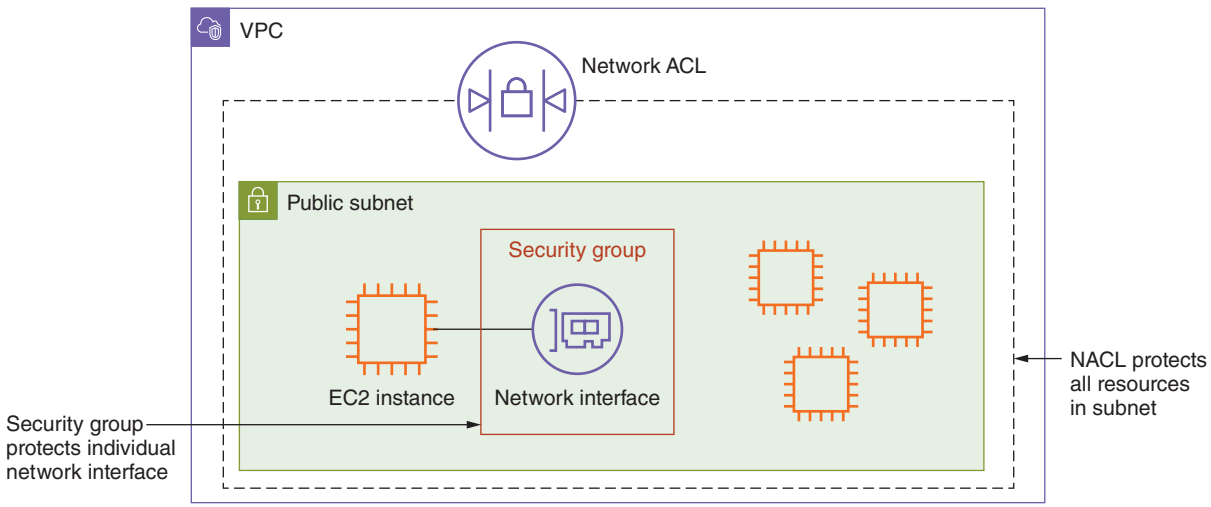
Instead of a binary "inside" versus "outside" approach to security, cloud-native security strategies attach security to individual resources. This is called a **zero trust** security model where everything in or out of the network is considered untrustworthy until proven otherwise. A zero trust model takes a default deny posture toward every entity and every request. To achieve embedded security at this scale, it must be implemented in many, seemingly redundant layers that permeate the network and protect resources from every angle.

In on-prem networks, you primarily secure the perimeter by closely monitoring traffic into and out of the network. This traffic travels in what's called a north–south direction. **Micro-segmentation** techniques allow for more granular control of security for traffic and workflows within the cloud. The flow of traffic between resources within the cloud is said to travel in a lateral or east–west direction. Micro-segmentation helps monitor and restrict this east–west activity and applies security measures specifically to workloads, VMs, or network connections.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

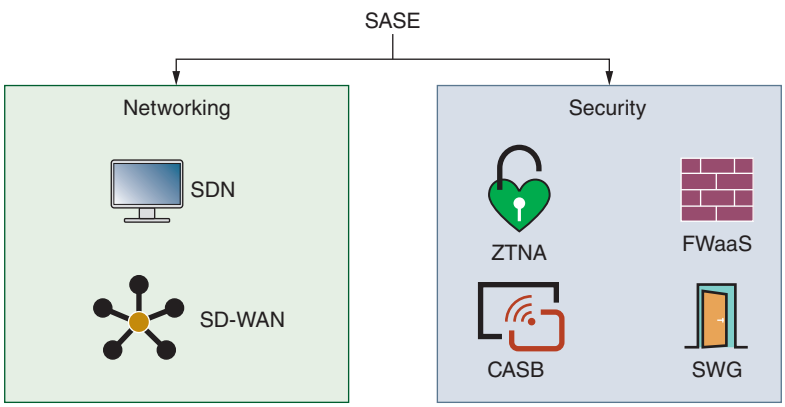
The zero trust security model is implemented in a **ZTA (zero trust architecture)**. This architecture ensures that traffic and access requests are evaluated at the most granular levels available. For example, consider the diagram in Figure 6-9. AWS recommends that firewall type protections on network interfaces—provided by security groups—take precedent over configuring firewall rules for a subnet, which are provided by a NACL (network access control list; informally pronounced "nackle"). Security groups protect individual network interfaces with rules customized to the needs of those interfaces, even to the point where a single VM with multiple network interfaces can have a different security group for each one, based on the needs of the traffic expected to reach that interface.

Figure 6-9 Security groups secure individual interfaces while NACLs secure subnets



ZTA is a component of a larger, all-encompassing approach to modern network security. SASE (secure access service edge), pronounced "sassy," is a theoretical model defined by Gartner in 2019 that merges software-defined networking services and zero trust security principles and services into a single framework to better serve the needs of distributed network resources (Figure 6-10).

Figure 6-10 SASE consists of networking and security technologies



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

On the networking side, SASE relies on SDN (software-defined networking) and SD-WAN (software-defined wide area networking) technologies to connect resources located in widespread locations, such as branch offices and cloud platforms. The security side of the SASE framework—collectively referred to as SSE (security service edge)—incorporates many security services that, together, can protect distributed resources. Although Gartner has identified 25 elements that make up the full SASE stack, the following list explains some of the core components of a SASE-based approach to cybersecurity:

- **ZTNA (zero trust network access)**—Somewhat of a misnomer, ZTNA refers to controlling remote access not to networks but to applications, data, and services independently of the type of network connection used. In practice, ZTNA can replace screened subnets to provide granular access control for all Internet-facing services.
- **CASB (cloud access security broker)**—A CASB (pronounced "caz-b") solution enforces access control and **DLP (data loss prevention)** policies on cloud-hosted resources, especially data. A DLP solution identifies sensitive data on the network and prevents it from being copied (such as downloading to a flash drive) or transmitted off the network (such as emailing or posting to cloud storage).
- **FWaaS (Firewall as a Service)**—Cloud-hosted firewalls provide scalable NGFW (next generation firewall) services for cloud-based network spaces.
- **SWG (secure web gateway)**—Similar to a firewall, an SWG sits between users and the Internet to filter incoming traffic and block risky or unauthorized outgoing traffic.

In the remainder of this section, you learn about the primary security techniques for virtual networks in AWS, Azure, and GCP. In the next section, you explore security techniques for resources hosted within those virtual networks. While there are significant similarities between each platform, there are also differences that will help you better understand the purposes and limitations of these various technologies. Notice how security can be applied at multiple layers, and each CSP has different recommendations for which services provide the best security for different purposes.

AWS VPCs and Subnets

NAC (network access control) balances the need for network access with the demands of network security by employing a set of network policies that determine the level and type of access granted. For example, adding a NAT gateway to a private subnet in AWS allows that subnet's VMs to access the Internet without exposing them directly to Internet traffic. In addition to the segmentation provided by public and private subnets, AWS offers several more techniques for securing VPCs and the resources contained within them, as described next.

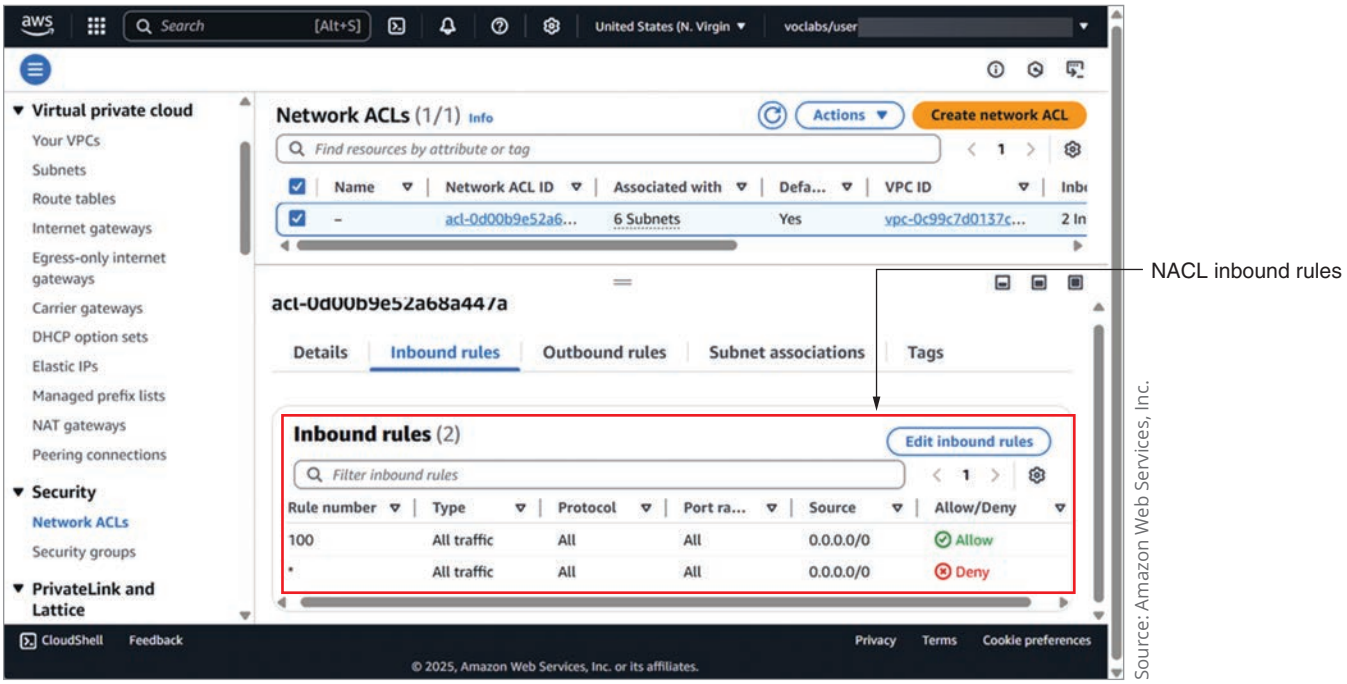
NACL (Network Access Control List)

Essentially a virtual, network firewall, a NACL can restrict or allow traffic into and out of a VPC's subnets. Every subnet in a VPC is initially associated with the default NACL, which by default allows all traffic. You can change these rules or create custom NACLs. You can use one NACL for multiple subnets, but each subnet must have exactly one NACL associated with it. Figure 6-11 shows the default NACL inbound rules near the bottom of the screen. Some specific NACL characteristics include the following:

- NACL rules can either allow or deny traffic.
- NACL rules are numbered for priority, starting with the lowest number. When a rule is found that applies to a message, that rule is enacted immediately without looking at any other rules.
- NACL rules are stateless, meaning inbound and outbound messages are evaluated by separate rules.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

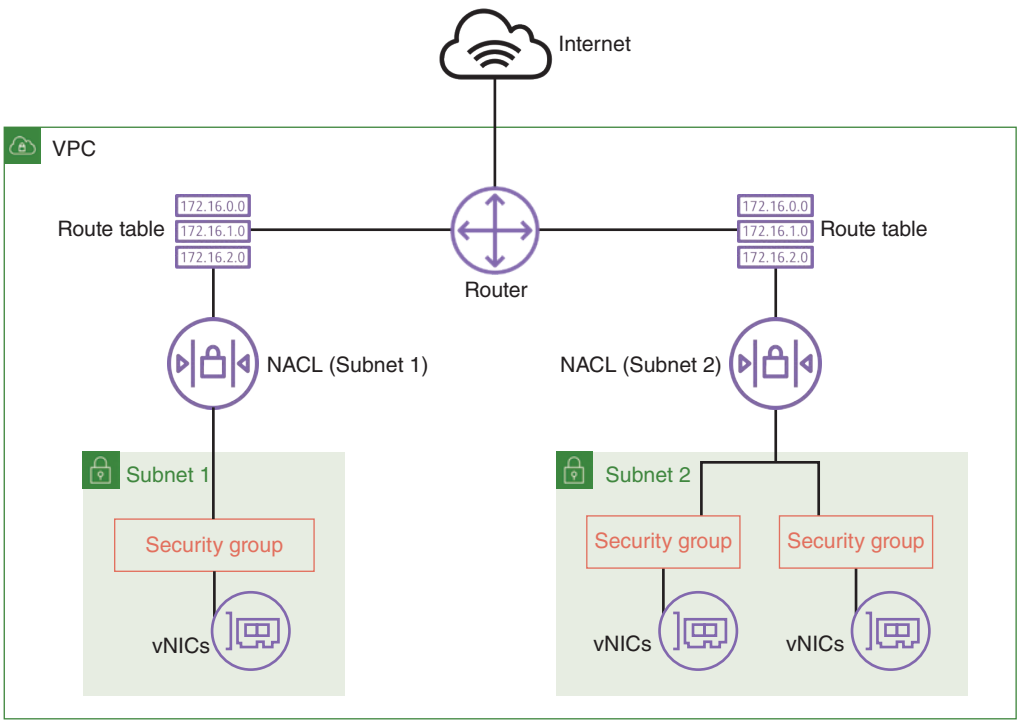
Figure 6-11 The default NACL rule of highest priority allows all traffic to enter the VPC



Security Groups

A security group serves as a virtual firewall at the host level. Like NACLs, security groups have rules that control whether traffic can cross a network interface. Unlike NACLs, security groups protect individual network interfaces within a subnet, not the entire subnet. You can associate up to five security groups per network interface, and interfaces within a subnet don't have to use the same security groups. The diagram in Figure 6-12 illustrates the relationship of NACLs, VPCs, security groups, and network interfaces in AWS.

Figure 6-12 NACLs apply to subnets, and security groups apply to instances



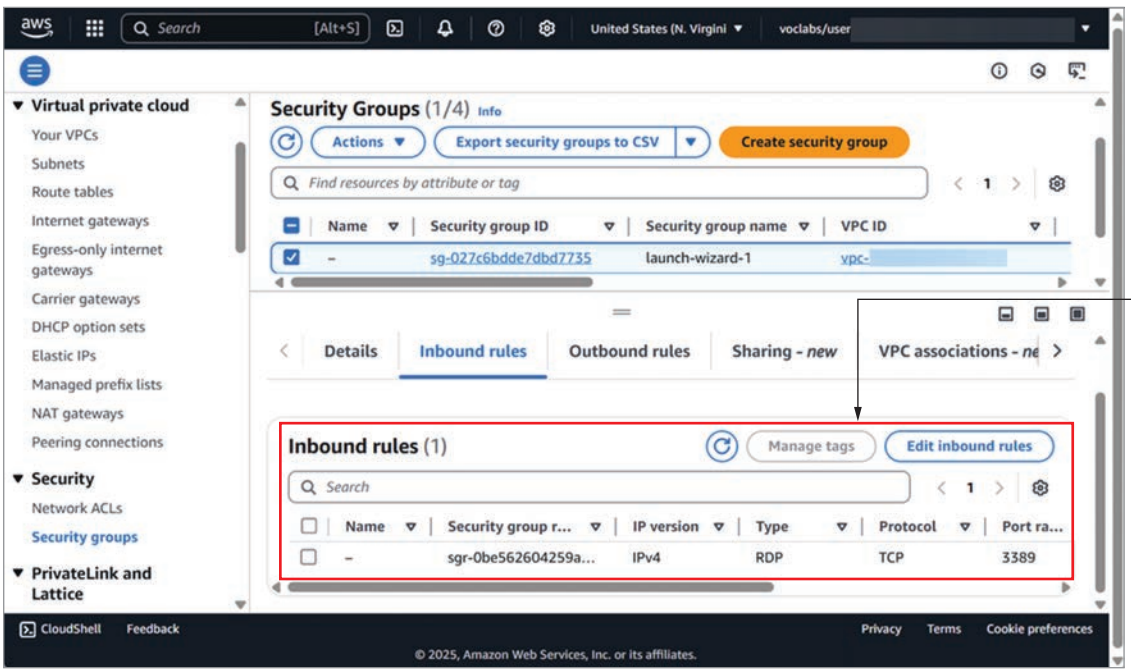
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

AWS best practices state that security groups and NACLs should be used together to secure EC2 instances because each of these security techniques offers different strengths. Some key characteristics of security groups are as follows:

- Security group rules only allow traffic—you can't create deny rules for security groups.
- Security group rules are stateful, meaning that return traffic for an existing connection is allowed regardless of any other security group rules.
- All security group rules are evaluated when determining whether to allow traffic to enter an instance's interface, and the most specific rule to the type of traffic is applied.

Figure 6-13 shows a security group rule allowing RDP traffic.

Figure 6-13 Only RDP traffic is allowed—all other incoming traffic is denied



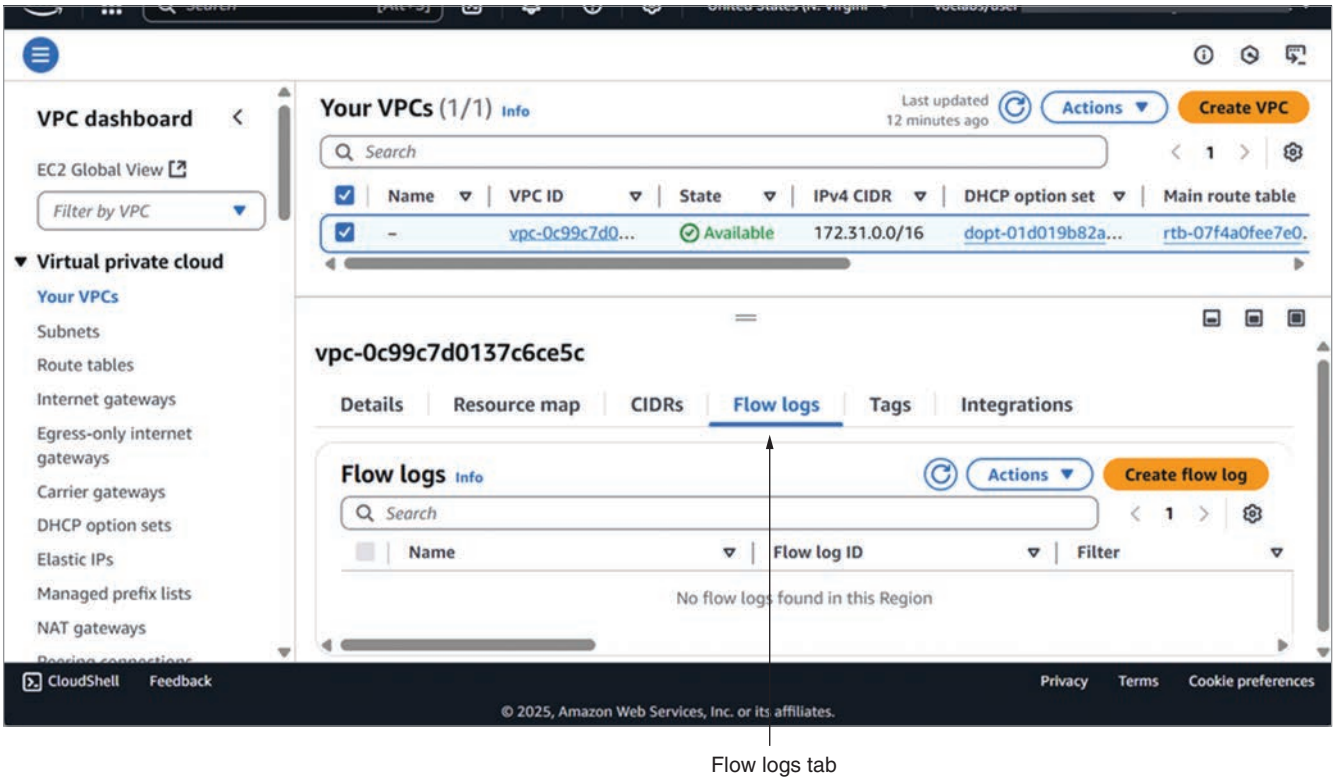
Source: Amazon Web Services, Inc.

Flow Logs

VPC flow logs monitor traffic that crosses an individual network interface. When you create a flow log for a VPC or subnet, all interfaces within that VPC or subnet are monitored. Log data is directed to the CloudWatch Logs service, to an Amazon S3 (Simple Storage Service) storage bucket, or to a service called Data Firehose that can deliver streaming data to any of several other destinations for further analysis. This activity accrues charges to your account. Fields within flow log records include information such as source and destination IP addresses (IPv4 or IPv6), source and destination ports, number of packets and bytes, start and end times, and whether the traffic was accepted or rejected. Because flow logs collect traffic data for a period of time (such as 15 minutes) before publishing each new record, there is a delay from when the traffic occurs until that information is available for review. Figure 6-14 shows where to find flow logs in a VPC's configuration information, and Figure 6-15 shows some of the configuration options when creating a flow log.

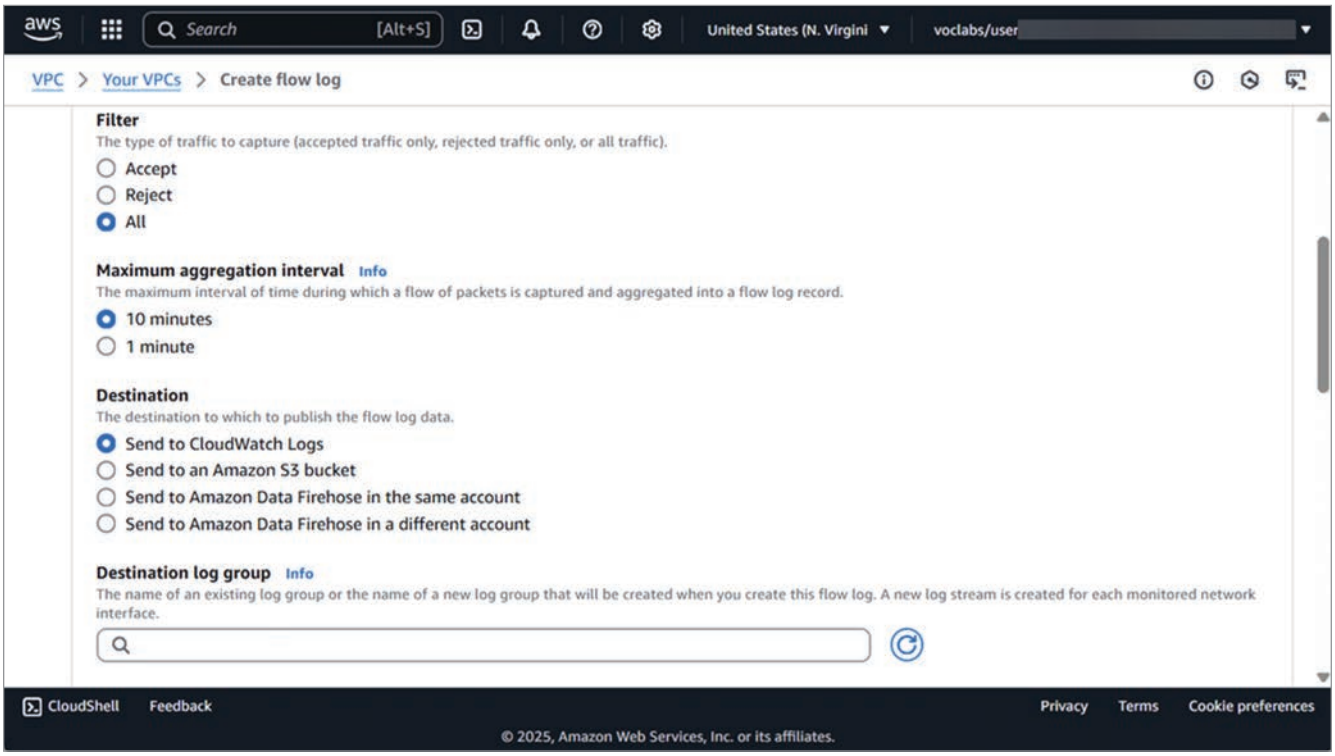
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-14 Find a VPC's flow logs



Source: Amazon Web Services, Inc.

Figure 6-15 Flow log configuration options



Source: Amazon Web Services, Inc.

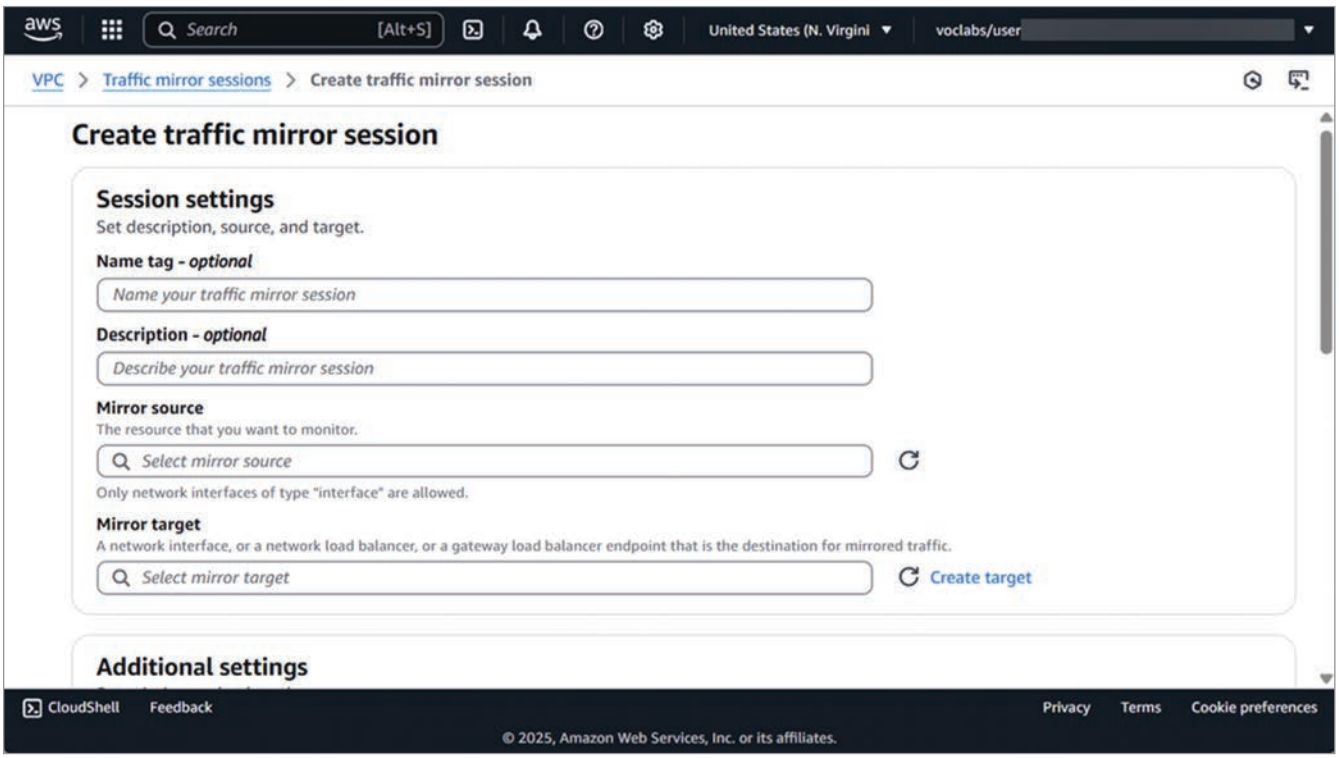
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Traffic Mirroring

Traffic mirroring in AWS provides deeper insight into captured traffic than what flow logs offer, and it's similar to the packet capture and analysis functions you can do with Wireshark. For example, flow logs gather statistics about allowed or denied traffic, but with traffic mirroring, you can also examine the content of denied packets. This information can be used to reverse-engineer a network attack and identify the vulnerabilities that allowed the attack to happen. You can also use this information to identify the root cause of a performance issue. Captured packets can also be exported to out-of-band monitoring and packet analyzer tools for filtering and examination.

To use traffic mirroring, create a source network interface (which can be attached to an EC2 instance or some other resource), a target (such as another network interface or a network load balancer), and a traffic mirroring session. Traffic mirroring sessions incur a small fee per hour per interface. Figure 6-16 shows the traffic mirror session settings requesting a mirror source and a mirror target.

Figure 6-16 Create a traffic mirroring session



Source: Amazon Web Services, Inc.

Note 6-2 | AWS NACLs, security groups, flow logs, and traffic mirroring sessions don't filter or monitor traffic to or from link-local addresses (169.254.0.0/16) or AWS-reserved IP addresses. These addresses are used to support DNS, DHCP, instance metadata, license management services, and subnet routing.

Network Firewalls

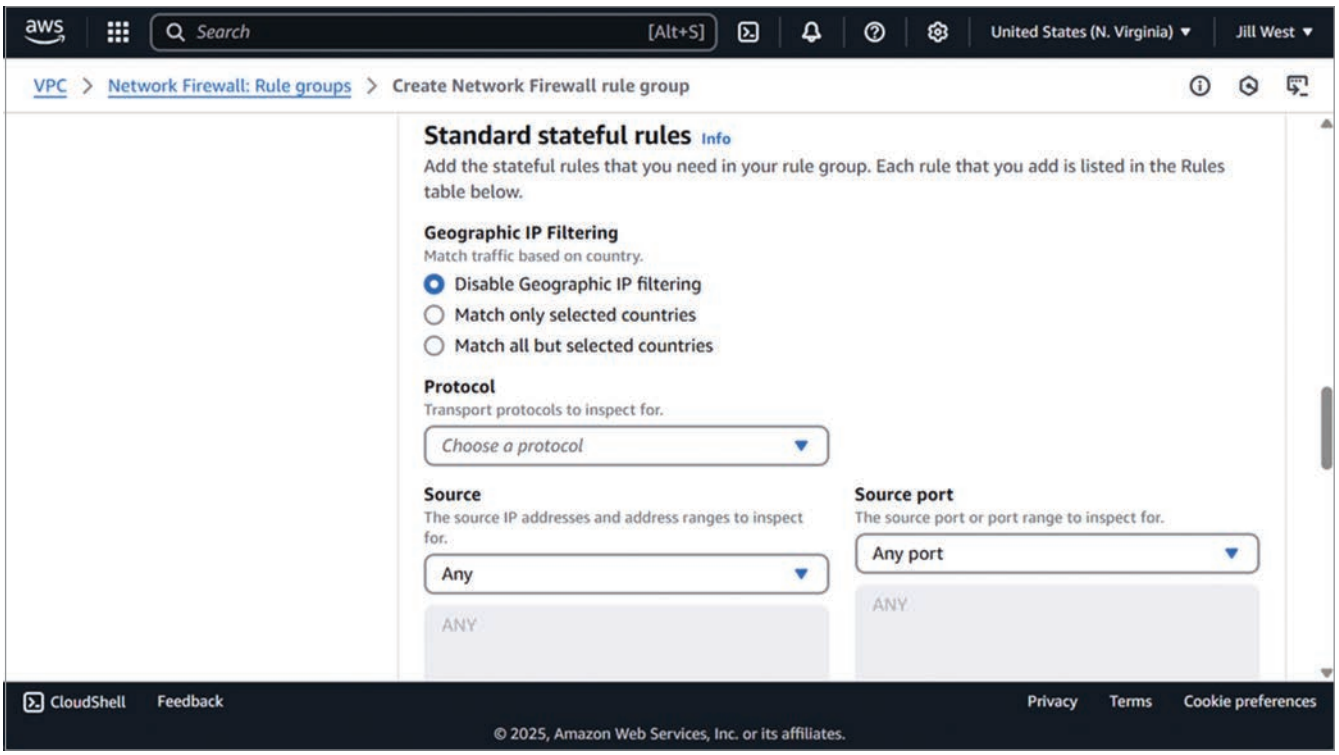
While each of the tools you've read about so far incorporates firewall rules such as allowing or denying traffic, you also have the option to add more robust, virtual firewall devices to your cloud networks. For this purpose, AWS offers Network Firewall.

AWS Network Firewall is a managed firewall, which is a PaaS resource that includes built-in high availability and unrestricted scalability. You might choose to incorporate a managed firewall to gain fine-tuned

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

control of traffic flows and greater visibility of network traffic across your entire account and even across multiple accounts. For example, recall that NACL rules are stateless and are evaluated in order of priority. In contrast, security group rules are stateful and are applied according to which rule provides the best match. However, AWS Firewall policies can be either stateful or stateless. Stateless rules are evaluated in order of priority. Stateful rules can be formatted and applied with more granular control, such as with geographic IP filtering (Figure 6-17).

Figure 6-17 Detailed filtering on stateful rules



Source: Amazon Web Services, Inc.

AWS Firewall also includes **IPS (intrusion prevention system)** capabilities that inspect traffic in real time at the network and application OSI layers. Recall that an IPS is a physical or virtual device or service that stands in-line between an attacker and the targeted network or host where it can prevent malicious traffic from reaching the target. A similar technology, **IDS (intrusion detection system)**, monitors network traffic, generating alerts about suspicious activity, but can do nothing to prevent suspicious traffic from reaching its target.

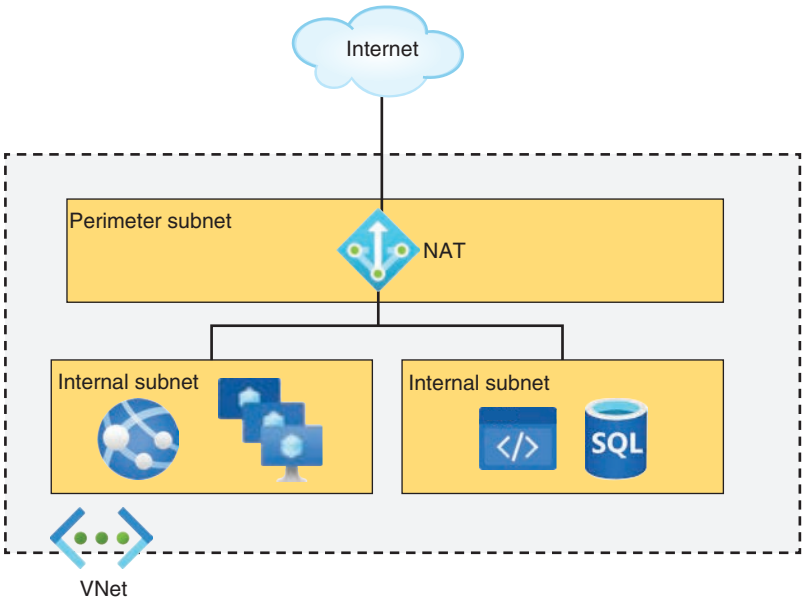
Azure VNets

Microsoft recommends that you establish a perimeter network in your Azure cloud to restrict and manage Internet traffic to your protected resources. A perimeter network is a subnet within a VNet that functions as a **screened subnet** (also called a DMZ, or demilitarized zone) between your private resources and the open Internet by standing between the Internet and your other subnets. The perimeter network in Azure is somewhat similar in concept to a public subnet in AWS. Figure 6-18 illustrates how this architecture is laid out conceptually.

Additionally, Azure offers several security devices you can configure to provide protection on and within your VNets, as described next.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-18 The perimeter network provides a level of isolation for internal subnets from the Internet



NSG (Network Security Group)

Similar to AWS's NACL, the Azure NSG (network security group) provides firewall-style protection at the subnet level. Unlike AWS's NACL, an NSG also can be applied to a single interface on a VM instance. In this sense, an NSG functions more like AWS's security group. However, Microsoft recommends using an NSG for an entire subnet instead. This arrangement applies the NSG to all instances in the subnet rather than creating complexity from using a different NSG for every instance.

An NSG can be created automatically when you create a new VM instance. Azure adds six default rules for every NSG: three for inbound traffic (Table 6-1) and three for outbound traffic (Table 6-2). These rules cannot be removed or altered, but you can override them by creating higher-priority rules. Lower numbers represent higher priorities. For example, the first rule in Table 6-1 has a priority of 65,000 and allows all traffic within the VNet. The last rule in Table 6-1 has a priority of 65,500 and denies all traffic, regardless of where it comes from or where it's going. But this rule has the lowest priority rating (that is, 65,500), which is a much lower priority than the first rule listed (that is, 65,000). If traffic matches the AllowVNetInBound rule, meaning it comes from a resource within the VNet and is sent to a resource also inside the VNet, the traffic is allowed, even though a lower-priority rule would deny that traffic.

Priority ratings for custom rules that you create can range from 100 to 4096. This means, for example, you could create a rule with a priority rating of 4000 that denies HTTP (Hypertext Transfer Protocol) traffic between VNet resources. This rule would trump the lower-priority rule, AllowVNetInBound whose priority rating is 65,000, which would have allowed that traffic.

Table 6-1 Default inbound rules in an Azure NSG

Name	Priority	Source	Destination	Access
AllowVNetInBound	65,000	VirtualNetwork	VirtualNetwork	Allow
AllowAzureLoadBalancerIn-Bound	65,001	AzureLoadBalancer	0.0.0.0/0	Allow
DenyAllInBound	65,500	0.0.0.0/0	0.0.0.0/0	Deny

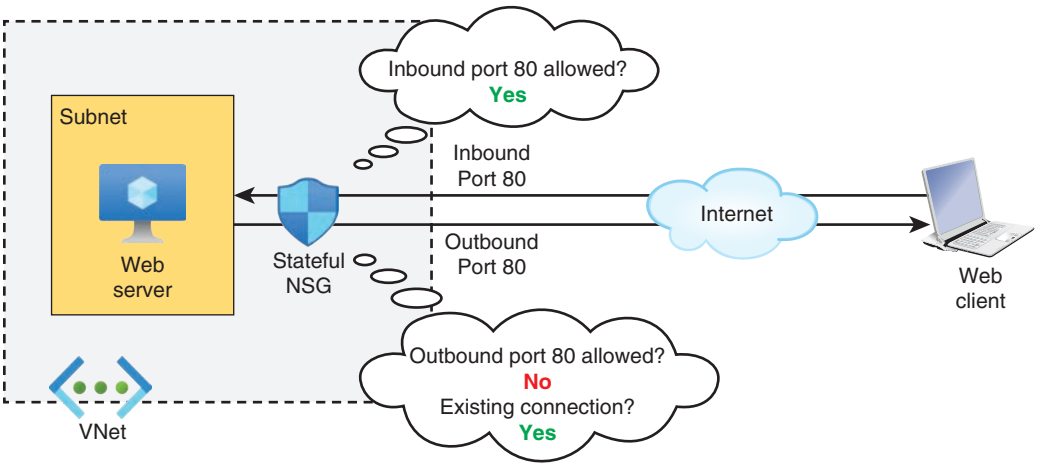
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 6-2 Default outbound rules in an Azure NSG

Name	Priority	Source	Destination	Access
AllowVNetOutBound	65,000	VirtualNetwork	VirtualNetwork	Allow
AllowInternetOutBound	65,001	0.0.0.0/0	Internet	Allow
DenyAllOutBound	65,500	0.0.0.0/0	0.0.0.0/0	Deny

NSG rules are stateful, meaning that once a message is allowed to cross the NSG, later messages in that conversation continue to be allowed. For example, imagine an incoming HTTP message on port 80 requesting webpage information is allowed through to a subnet that holds a web server. The web server can then reply to the request even if outgoing rules don't specify that traffic is allowed on port 80. The same is true in reverse. If a subnet's NSG allows an instance inside the subnet to send traffic on port 80 to external websites, inbound responses from those websites are allowed even if port 80 is not explicitly allowed by an inbound NSG rule. Figure 6-19 illustrates how this works to support a connection between a web client and a web server.

Figure 6-19 Return traffic on an existing connection is allowed by the NSG



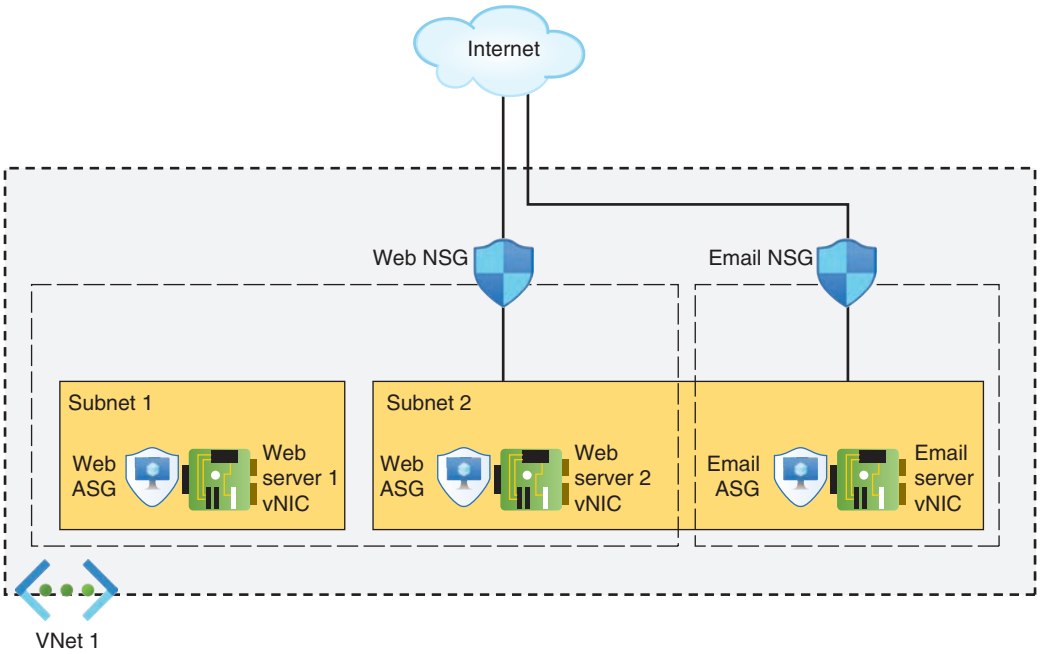
ASG (Application Security Group)

An ASG (application security group) allows for micro-segmentation of your security policies according to application and workload type within the VNet. You can associate one or more ASGs to each network interface. For example, you might configure one group of VM interfaces to allow access to an application's database, while another ASG in the VNet might not have this access. ASGs can also be identified as a source or destination for NSG rules.

Note that ASGs are not themselves firewall-type security like NSGs in Azure and NACLs in AWS are. Instead, an ASG is a logical grouping of vNICs that can collectively receive security configurations, such as the application of an NSG (Figure 6-20). It's still the NSG or other resources that perform the actual traffic filtering.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-20 vNICs in an ASG



Other Azure Security Tools

Azure offers many other tools for securing cloud traffic, including the following:

- Azure Firewall is a firewall as a service with stateful inspection for highly available threat protection.
- Azure Network Watcher captures and monitors traffic logs for troubleshooting.
- MS Defender for Cloud provides anti-malware not just for VM instances but also for containers, code pipelines, storage resources, databases, network traffic, and other kinds of workloads.
- VNet TAP (Terminal Access Point) streams network traffic for capture and analysis.

GCP VPCs

Segmentation of GCP resources is first achieved at the project level. Resources that are logically related, such as the various components of an application or the resources used by a specific team, typically should be contained in the same project. However, resources that have no need to communicate with each other, such as the HR database and the sales team's CRM (customer relationship management) database, should be hosted in different projects.

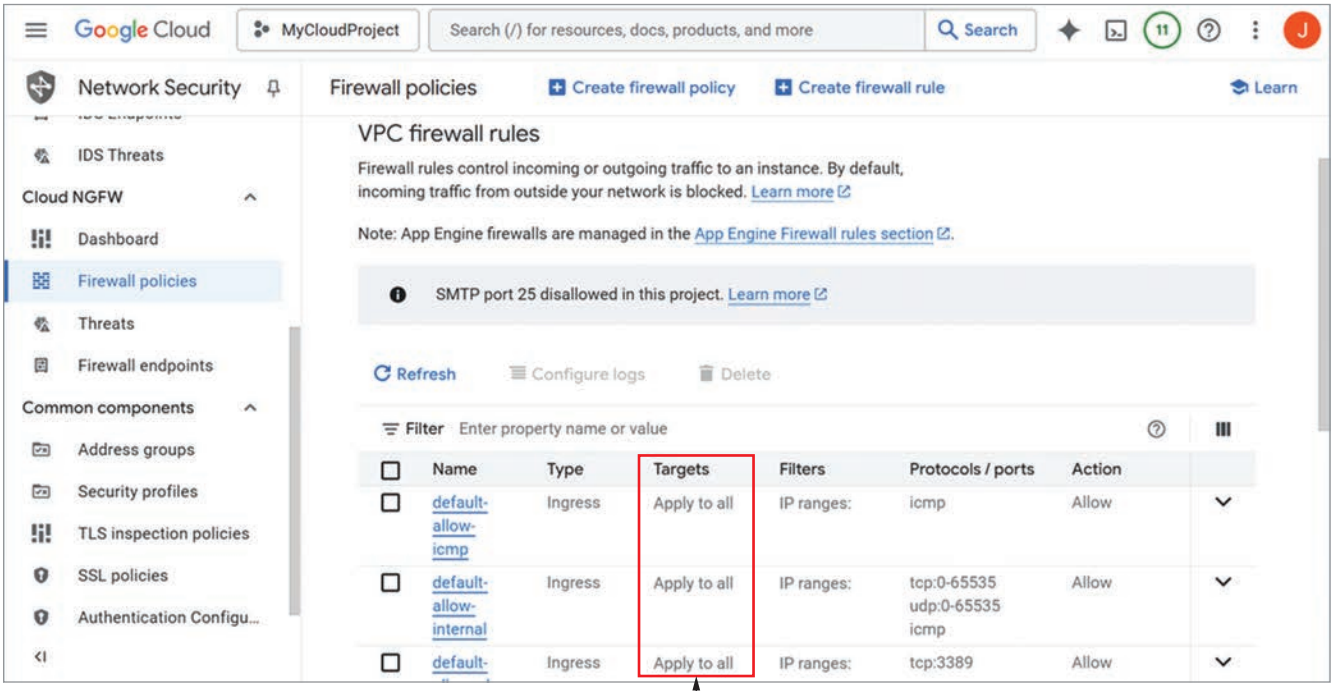
Other techniques and tools are available for managing and monitoring traffic into and out of a VPC, including VPC firewall rules, VPC flow logs, and packet mirroring, described next.

VPC Firewall Rules

GCP's VPC firewall rules (Figure 6-21) are set at the network level to manage traffic into and out of the VPC. Additionally, firewall rules work within the VPC between the VPC's own resources. These rules can be set to target specific protocols, ports, sources, or destinations. However, each rule can contain either IPv4 ranges or IPv6 ranges—not both—and IPv6 is supported only if enabled in the VPC. Some traffic within a VPC is always blocked, such as incoming DHCP offers, regardless of what the firewall rules say. Other traffic is always allowed, such as traffic in conversations with the Google Cloud metadata server that provides the VPC with DHCP and DNS services. Set firewall rules to allow or deny specific types of traffic. Consider carefully which traffic needs to be allowed so you can minimize the traffic allowed into, out of, or through the VPC.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-21 Firewall rules can apply to one, some, or all instances within a VPC



Source: Google LLC

Firewall rules are applied first according to priority; if two rules of the same priority equally apply to traffic, a deny rule takes precedence. Firewall rules are also stateful, meaning that traffic allowed in one direction automatically allows traffic in the other direction for an active connection. As long as there is at least one message going in either direction within 10 minutes, the connection is kept active. You can enable firewall logging to monitor the effects of specific rules on your VPC traffic.

VPC Flow Logs

The VPC flow logs service is enabled at the subnet level (Figure 6-22). Once enabled, flow logs track TCP, UDP, ICMP, and some other types of traffic for all instances in that subnet. Each log record contains two types of fields:

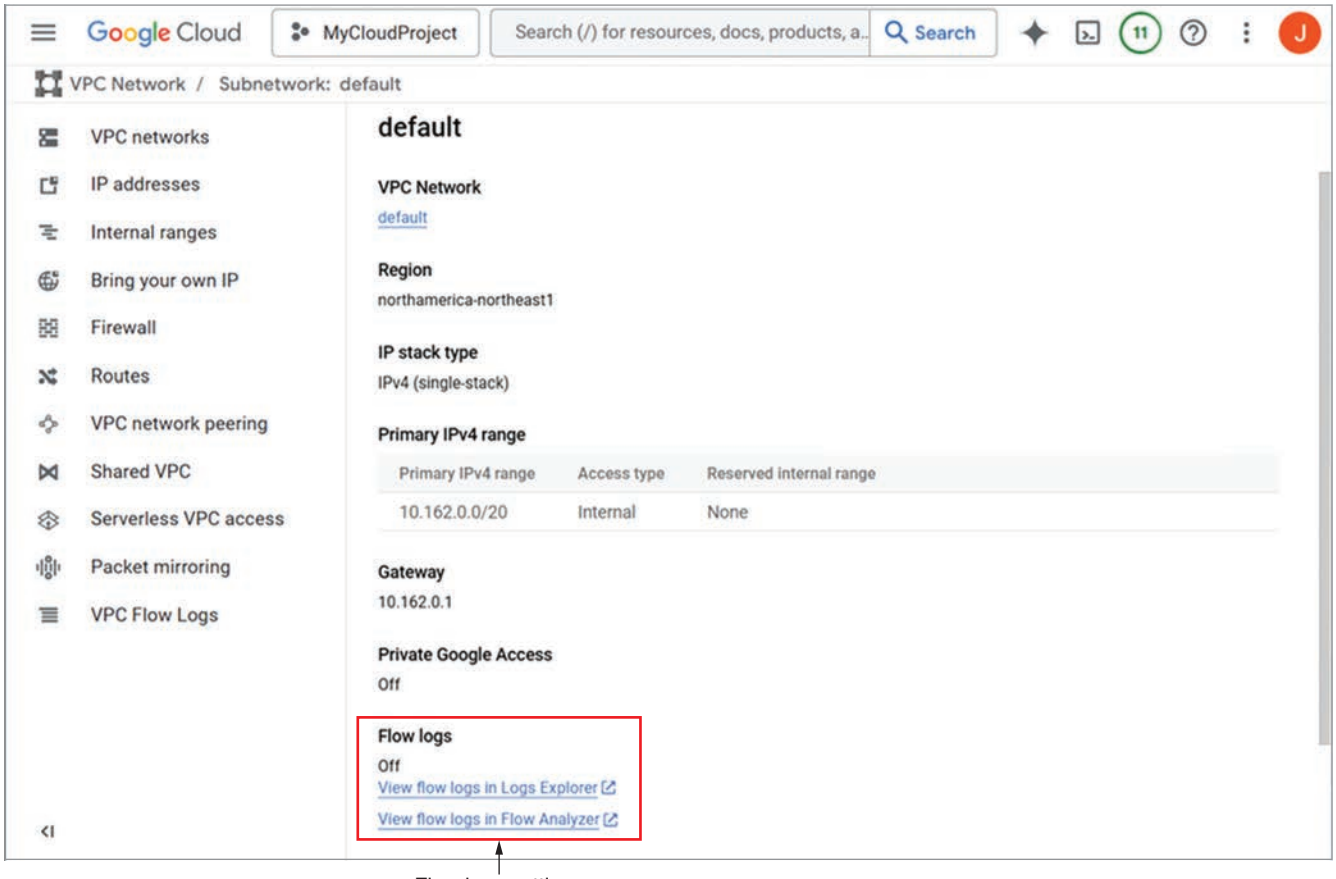
- Base fields are common to all log records, such as connection type, start time, and bytes sent.
- Metadata fields are specific to the type of log record and, for VPC flow logs, contain information such as source and destination instance and source and destination location. To save storage space (which reduces cost), metadata fields can be excluded from flow log records.

Sampled messages are collected into a single flow log record at various time intervals as short as 5 seconds (the default configuration) or as long as 15 minutes. These records are stored in a GCP storage service, such as Cloud Logging, which incurs charges after a basic, free allotment. While not every message crossing a VM's interfaces is captured in a flow log record, log record captures still take up a great deal of storage space.

Flow logs are helpful for troubleshooting connectivity issues and monitoring network performance. They're also used for network forensics when investigating an incident or for real-time security monitoring using third-party solutions.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-22 Enable flow logs in the subnet settings



Packet Mirroring

GCP also offers a Packet Mirroring service that copies traffic involving specified VM instances for further analysis—this capture can include incoming traffic, outgoing traffic, or both. The captures also include message headers and payloads, similar to what Wireshark can do. Messages can then be analyzed by security software or evaluated for performance issues.

Securing Hybrid and Multi-Clouds

Securing resources across a hybrid or multi-cloud presents the challenge of needing to blend different security tools into a unified system for both simplicity and effectiveness. Cloud security techniques must auto-scale, while on-prem security tools must have deep access within the network perimeter. Further complicating this situation, cloud security tools in each platform can vary widely in technique, applicable layer, and output types.

Many of the same best practices apply for hybrid and multi-clouds as for other cloud structures, such as encrypting data, monitoring activity on the network, and using segmentation and micro-segmentation to isolate certain resources. Some tools have been developed specifically to help bridge the security gap between the on-prem and cloud portions of an organization's networks or between different cloud platforms. One of the most commonly used is a CASB solution, which you read about earlier in this chapter.

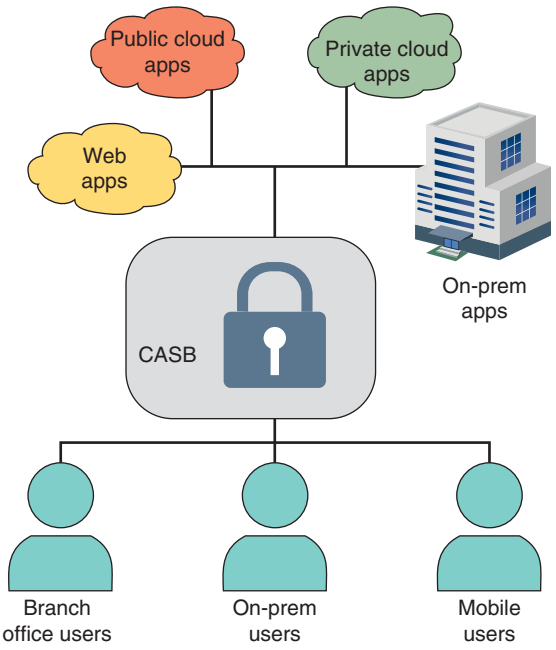
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

It's designed to detect applications and other resources running within a domain—even resources IT staff don't know about. It can then monitor those resources according to policies set by the organization. Examples include Bitglass's Zero-day CASB or Masergy's Managed CASB.

A CASB product increases transparency throughout an organization's domain, both on-prem and in the cloud. The CASB provides a central dashboard, sometimes called a "single pane of glass," for managing security across the cloud and on-prem environments. Gartner defines a CASB's purposes as providing visibility, compliance, data security, and threat protection.

The reason for a CASB's increased reach is because it functions between users and the services provided to them (Figure 6-23). This is true regardless of whether that service is located on-prem or in any cloud platform, even web apps such as video conferencing tools and office productivity tools. This also provides adaptability to a variety of user devices, even if those devices are not monitored by the protected organization. A CASB can also enforce security at layers beyond network services, such as preventing data loss, enforcing password policies, and applying access policies.

Figure 6-23 CASB placement



Self-Check

4. A stateless rule applies to _____.
   a. incoming traffic
   b. outgoing traffic
   c. traffic in one direction
   d. traffic in both directions

5. What cloud security tool functions like Wireshark?
   a. Virtual firewalls
   b. Traffic mirroring
   c. Flow logs
   d. Security groups

6. Security at which of these levels provides the most fine-tuned traffic filtering?
   a. Subnet
   b. Interface
   c. Instance
   d. Virtual network

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 6-3: Compute Security

Certification

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.3 Given a scenario, implement identity and access management.

4.4 Given a scenario, apply security best practices.

4.5 Given a scenario, apply security controls in the cloud.

5.2 Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.

6.3 Given a scenario, troubleshoot security issues.

Typical cloud compute resources include IaaS resources (such as VM instances), PaaS resources (such as code), and SaaS resources (such as applications). While you've already learned about some security functions that apply to network interfaces on VM instances (such as AWS security groups) in conjunction with network security, other techniques help secure instances from within the instance. Similarly, security functions must be configured on containers, code, and applications running in the cloud. This section explores compute security for instances and applications.

Endpoint Protection

VM instances, like physical workstations, require patches, updates, and anti-malware measures to protect them from external threats. Firmware upgrades and OS upgrades play an important role in keeping on-prem systems up to date and securely hardened against potential attack vectors. The standard methods for securing a physical or virtual machine on-prem similarly apply to VM instances running in the cloud. For example, access and privileges should be restricted to the minimum required for each instance, and deviations in performance should be monitored. The cloud offers additional security options, depending on the specific tools available in a cloud platform. Consider the following techniques:

- **Plan subnet placement**—Carefully choose the subnet for each instance so that it has only the necessary access to the outside world. For example, a public subnet in AWS should only hold VM instances that need direct access to the Internet.
- **Disable unnecessary ports and services**—Just like with a physical machine, only needed ports and services should be enabled. For example, don't allow RDP (Remote Desktop Protocol at port 3389) or SSH (Secure Shell at port 22) traffic from all Internet addresses using the wildcard 0.0.0.0/0 except during initial testing. To remote into VMs in a production environment, use VPN or direct connect options instead, or at the very least, specify the public IP address for the computer you're working from. Many cloud platforms additionally provide services designed for more secure remote access to these instances, such as using a bastion host located in a public subnet that gives in-line access to instances located in a private subnet.
- **Enforce account management policies**—Linux distros are configured with default user accounts, which you can use to remote into VMs. To further protect these virtual systems, deactivate any default user accounts and require other, best-practice account management policies such as password length and least privilege access.
- **Install anti-malware software and keep it updated**—Windows Server 2016 and beyond comes with Microsoft Defender Antivirus preinstalled, which provides robust anti-malware protection for the VM itself. For other OSs, make sure sufficient anti-malware software is installed and kept updated. This

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

might be accomplished through the VM's OS, or it might be available as a service from the cloud platform. For example, the Microsoft Antimalware for Azure extension service can be added to a VM to provide real-time protection, remediation, and reporting. Microsoft Antimalware provides an excellent application of this technology for VMs running OSs besides Windows Server.

- **Install host-based/software firewalls and IDS/IPS**—While your virtual network should already have firewall rules protecting VMs within all its subnets, each VM should also have a host-based, software firewall installed on it. Again, this might be accomplished through the VM's OS or via a cloud service, such as AWS's security groups. Another option is to add an HIPS (host-based IPS) or an HIDS (host-based IDS), which monitors a VM for signs of malicious behavior and, in the case of an HIPS solution, can activate protective measures in response.
- **Create single-function resources**—In the cloud, especially considering the flexibility and elasticity in choosing appropriately sized resources, you can create single-function resources that are optimally designed to perform only a specific task. This limited scope offers the ability to create a hardened baseline specific to each function and leaves a much smaller attack surface for each resource.

Considering how convenient it is to create, or "spin up," new instances, keep in mind how easy it is to forget about unused instances as well. These running systems, if they're not monitored and maintained, eventually become outdated and vulnerable to infiltration. From there, hackers can access other resources in your network. Consider using an **EDR (endpoint detection and response)** solution that detects and monitors endpoints in your cloud to increase visibility of attack surfaces from forgotten or outdated instances.

Container Security

Containers present a different type of endpoint that requires specialized approaches to security. There are many configuration options you can set on a system hosting containers. For example, you can determine what permissions the container is given on the host and how the container accesses the information it needs, as follows:

- **Container permissions on host**—One aspect of container security is the permissions given to the container itself in relation to its host (Figure 6-24). A container is assigned a UID (user ID) and a GID (group ID), which determine these permissions. A **privileged container** is associated with the host's root user account and therefore is given some level of root access to the host. This is inherently risky for the host and any other containers or applications running on it. A better option is to configure **unprivileged containers**, which are assigned non-root user accounts on the host and therefore do not have root access to the host. This configuration is safer by design. Many cloud platforms, such as AWS, set containers to unprivileged by default and show pop-up warnings if you try to change this setting. Any special permissions the container truly needs can be granted using other methods that provide more granular control of container permissions on the host.

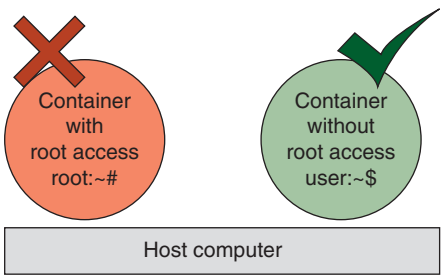


Figure 6-24 Containers receive access permissions from their host

- **File access permissions**—Although containers generally are not expected to hold data for very long, they might at times need access to data storage resources. In this case, consider the container's file access permissions. When configuring containers in the cloud, be sure to assign a role to containers that provide exactly the privileges they need and no more.

Caution | Both containers and VM instances rely on images when they're created. Using images posted by the cloud provider tends to be relatively safe. However, many images can be obtained through a cloud-hosted marketplace or from other parties. Be especially careful of using images from other sources. Make sure you trust the source, and check the image for any signs of vulnerabilities or other compromised configurations.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Application Security

The security of compute resources additionally can be placed at risk by the applications that run on them. Some applications are run on VM instances in services such as AWS's EC2. Further, application security is required for SaaS applications and for the PaaS components used to build those applications. A policy called an approved application list (also known as whitelisting) can restrict which applications can run on an endpoint, whether physical or virtual. This policy can cause problems with users, however, if they rely on a wide variety of applications to do their jobs. A deny-by-default approach increases security but can also increase user frustration and slow their work progress. Consider the trade-offs carefully, and thoughtfully incorporate application restrictions that prioritize company goals while minimizing risks from unauthorized software.

WAF (Web Application Firewall)

A special kind of firewall to protect SaaS applications is a **WAF (web application firewall)**. WAFs can be installed either on-prem or in the cloud and are designed specifically to protect web apps, which are applications accessible over the Internet. (Google Docs is an example of a web app.) A WAF can inspect, monitor, and filter messages coming into a web app to protect it from attack. Many cloud platforms offer WAF services, or you can use a third-party product such as Cloudflare WAF or Imperva WAF.

Pipeline Security

Application security is no less important during application development. Application code development and other automations are managed through a series of life cycle processes that make up a **pipeline**. Consider the various security risks faced by development pipelines: compromised code, unauthorized access, interference with security testing, and exposure of sensitive data or secrets (such as access keys or encryption keys). Ideally, pipeline security should be automated as much as possible to minimize the risk of human error. Continuous monitoring can capture thorough logs, analyze these logs, and detect incidents in real time. The following systems can be used to monitor for and mitigate security issues:

- Secure code:
  - **SCA (source composition analysis)**—A pipeline might import code from third-party sources, such as open-source code repositories. An SCA tool can scan this code for vulnerabilities. It relies on vulnerability databases to identify issues.
  - **SAST (static application security testing)**—A SAST tool scans code written in-house. In contrast to an SCA tool, a SAST tool analyzes the code directly rather than relying on vulnerability databases. It can identify issues such as software security vulnerabilities and coding errors.
- Secure access:
  - **Principle of least privilege**—Access to repositories, code, image registries, environment configurations, and automation configurations should be provided only as needed, conforming to the principle of least privilege. Access control mechanisms include authentication and authorization mechanisms, which should integrate MFA and RBAC (role-based access control) features.
  - **Secrets management**—Containers often need access to secret information, such as account credentials. Consider the sensitivity of secrets such as access keys, encryption keys, and passwords. An **access key** is a string of characters used to authenticate to a system, similar in concept to a password. An **encryption key** is a string of characters combined with a block of data during encryption to obfuscate the data. Encryption keys, access keys, and passwords should never be included directly in an application's code, a configuration file, or a template used to deploy resources, and they should never be stored in plaintext. Instead, this information can be stored in and managed with a secrets management tool, such as AWS's KMS (Key Management Service). Services like KMS can automatically rotate keys, manage access to the keys, and log key access for auditing and compliance requirements.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- Secure environment:
  - **Container security**—The containers used to run code require their own layers of security. Make sure to apply best practices for container security, as described earlier.
  - **Runtime security**—Other techniques you're already familiar with (such as firewalls and monitoring) or will learn about later in this course should be used to protect and secure runtime environments, including testing, staging, beta, and production environments. Runtime security monitors application behavior to detect indications of compromise.

Self-Check

7. Which security technique will increase the number of running resources while decreasing the attack surface?
   a. Disable unnecessary ports and services.
   b. Create single-function resources.
   c. Install host-based firewalls.
   d. Plan subnet placement.

8. Which security tool can you use to find security vulnerabilities in forgotten EC2 instances?
   a. SCA
   b. HIDS
   c. EDR
   d. SAST

Check your answers at the end of this chapter.

Section 6-4: Data Security

Certification

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.3 Given a scenario, implement identity and access management.

4.4 Give a scenario, apply security best practices.

4.5 Given a scenario, apply security controls in the cloud.

6.2 Given a scenario, troubleshoot network issues.

6.3 Given a scenario, troubleshoot security issues.

As you've already learned, data sitting in the cloud is not protected by a secure perimeter the same way data in the on-prem network can be. Security measures need to travel with the data, no matter where it's stored or what hardware it transits. When designing data security measures, keep in mind that data can exist in three states:

- **At rest**—Data at rest, whether stored on a local device or in the cloud, includes data in a storage service or data on a virtual or physical storage device.
- **In motion**—Data in motion, sometimes called "in flight," is being transported between locations, whether that's within a single server along its motherboard's bus lines, between devices on a single network, or between networks and possibly across the Internet.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **In use**—Data in use is temporarily stored in memory or is being used in calculations by a processor. The data might be visible to a user, who might be making changes to it. Or the data might be in use by an underlying compute process.

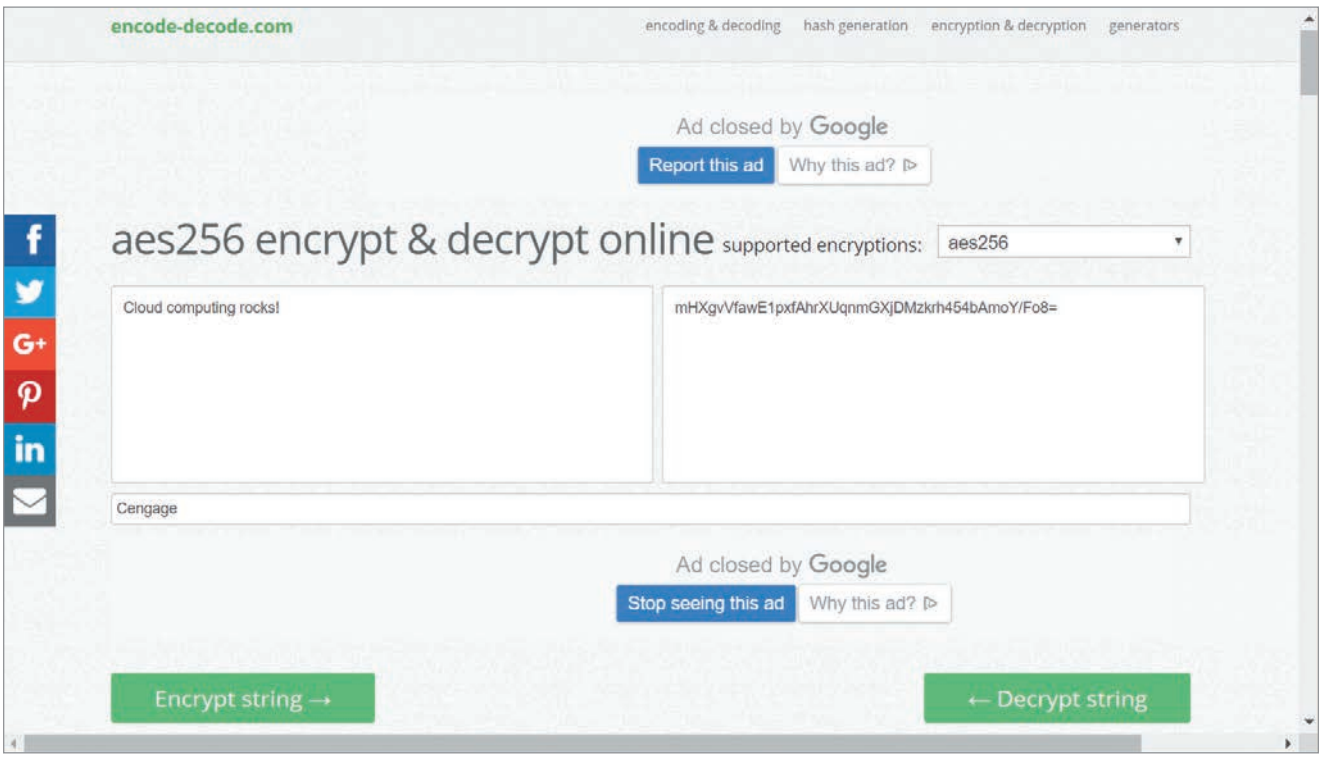
Robust cloud security can reliably protect data in all three of these states—*if* the security measures are correctly configured. In fact, DLP solutions, such as CASBs and SWGs (secure web gateways), specifically target data in all three states to provide comprehensive protection.

Encrypting Data

Unfortunately, a slight network or endpoint misconfiguration can undermine even the most advanced cloud security technologies, resulting in unauthorized access to sensitive data. However, if the data is encrypted, it can still be protected so that damage caused by a breach is minimized. **Encryption** is the use of an algorithm to scramble data into a format that can be read only by reversing the algorithm—that is, by decrypting the data. For example, consider what would happen if you stored sensitive data on a laptop and the laptop was stolen. If that data is encrypted, such as by using an SED (self-encrypting drive) enabled by TPM (Trusted Platform Module) hardware authentication, the data might still be safe from discovery, as the thief would not be able to read the data. Many of the worst cloud-based data breaches could have been significantly reduced or even avoided had the organizations involved used reliable encryption technologies.

Data that has not been encrypted is called plaintext and can be read without any special decoding. In contrast, encrypted data looks like gibberish. For example, Figure 6-25 shows the phrase "Cloud computing rocks!" encrypted using AES (Advanced Encryption Standard) encryption and the key "Cengage." The encrypted output in the right box doesn't reveal any pattern that would indicate the original text in the left box.

Figure 6-25 AES encryption is considered secure



Source: Encode-decode.com

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Encryption of data stored in the cloud can be performed by the company who owns the data before that data leaves the company's network, by the CSP where the data is stored, or by a third-party encryption service. These options add new layers of complexity to encryption of cloud data beyond what you would encounter in on-prem data encryption processes. The following tasks must be completed to deploy encryption in the cloud:

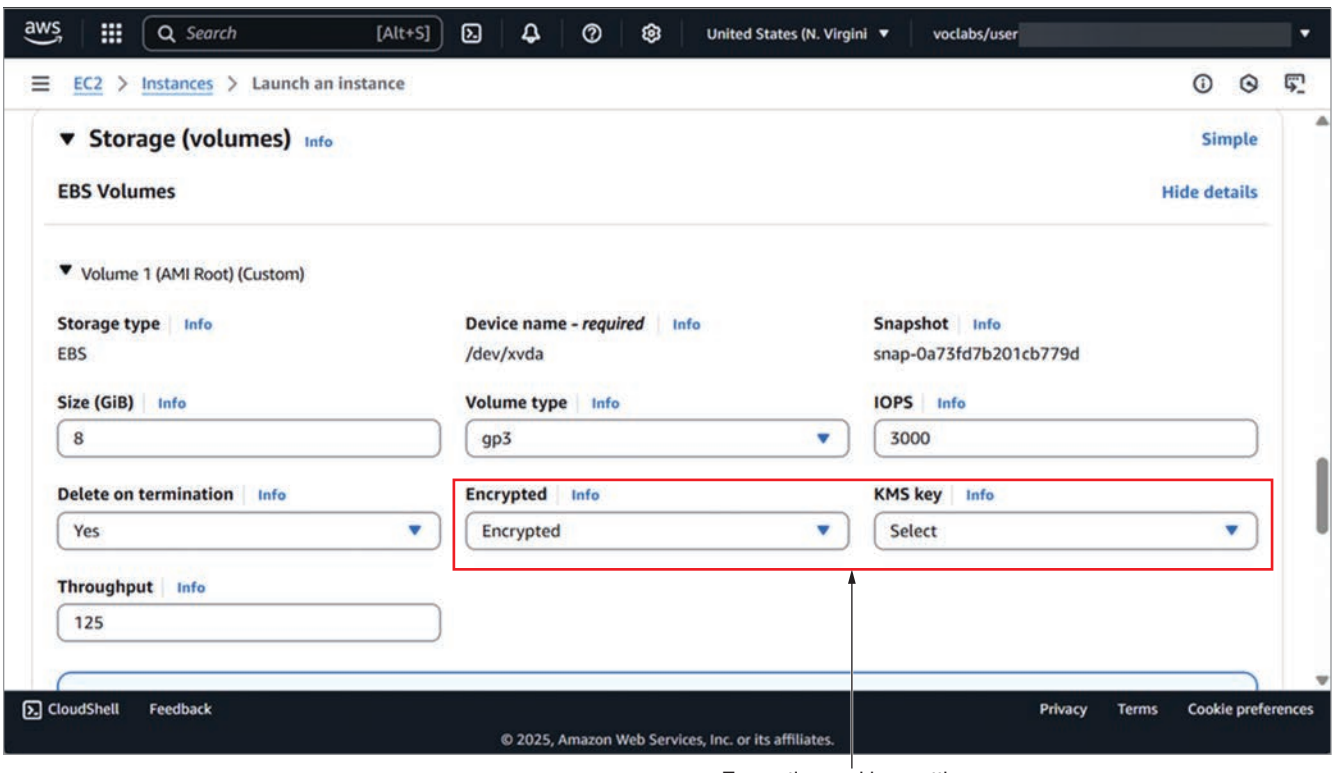
- **Training**—Learn about and work with the CSP's encryption capabilities, policies, and techniques.
- **Key location**—Determine where to store the encryption keys. Storing keys in the same location as the data might be more convenient—both for you and for any hackers. Therefore, in most cases, keys should be stored in a location that is isolated from the data those keys encrypt.
- **Costs**—Consider the costs involved with each of the available options. Performing encryption in-house before sending data to the cloud might be more cost-effective as well as more secure if this works for a particular situation. There are trade-offs in functionality, however, such as not being able to search or edit files stored in the cloud without first downloading and decrypting them. (However, some services are beginning to offer the ability to search encrypted files.)

Encryption Techniques

Encryption can be used to protect data in all three of its primary states, such as in the following examples:

- **At rest**—Storage volumes attached to a VM instance can be encrypted, either in part or in whole. For example, in AWS, you can encrypt an EBS (Elastic Block Storage) volume attached to an EC2 instance. Figure 6-26 shows the option to encrypt a new instance's boot volume (that is, the volume on which the OS files are stored) during EC2 instance creation. AES and RSA (Rivest Shamir Adleman), an algorithm named for its three inventors, are often used to encrypt data at rest.

Figure 6-26 Encrypt an instance's root storage volume



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **In motion**—Data in transit across the Internet or other networks can be encrypted with TLS (Transport Layer Security), SSL (Secure Sockets Layer), or IPsec (Internet Protocol security). For example, TLS or SSL can secure HTTP to make HTTPS (HTTP Secure), and IPsec is used on VPN connections. When building cloud solutions, APIs enable communication between various resources that work together for an application, network, or other system. Data carried by these API messages can be encrypted by HTTPS to protect information such as passwords, private keys, and credit card information. For example, AWS's API Gateway is a service to create, manage, and secure API endpoints. The service requires that API endpoints accept only HTTPS traffic, which ensures API communications are encrypted between endpoints.
- **In use**—Rather than encrypting data only after it is stored, some cloud applications can also encrypt data while it is being used by implementing advanced encryption technologies. Additionally, it's possible to create an isolated environment, called a TEE (trusted execution environment), where sensitive data can be read and processed without exposing it to users or systems outside of that environment, even to the host OS. This technology is a component of confidential computing. Increasingly, cloud platforms are offering services, such as AWS's Nitro Enclaves and Azure's Confidential Computing, to implement these kinds of protected environments and process sensitive data at scale. Securing data in use, however, usually relies more on endpoint security, access control, and user training.

To protect data at rest, in motion, and in use, encryption methods are primarily evaluated by three benchmarks. Together, these principles form the standard security model called the **CIA (confidentiality, integrity, and availability) triad**, as follows:

- **Confidentiality**—Data can only be viewed by its intended recipient or at its intended destination.
- **Integrity**—Data is not modified in the time after the sender transmits it and before the receiver picks it up.
- **Availability**—Data is accessible to the intended recipient when needed, meaning the sender is accountable for successful delivery of the data.

Note 6-3 | **Hashing** means to transform data through an algorithm that is mathematically irreversible. Hashing is not the same as encryption, though it's often listed as a type of encryption and does, in a similar manner, transform data from plaintext to seeming gibberish. Encrypted data can be decrypted, but hashed data cannot. Hashing is mostly used to ensure data integrity—that is, to verify data has not been altered, which is similar to the purpose of a checksum. However, hashes can play a critical role in data security.

If a secure algorithm is used, hashing is realistically impossible to reverse. Instead, you can take known data, hash it using the same hashing function, and compare the new hash with the stored, hashed data. If the hashes match, this indicates the known data is exactly the same as the original data. If the output does not match, this indicates the data has likely been altered. In fact, this is often the most secure way to store and use passwords.

Securing Protocols

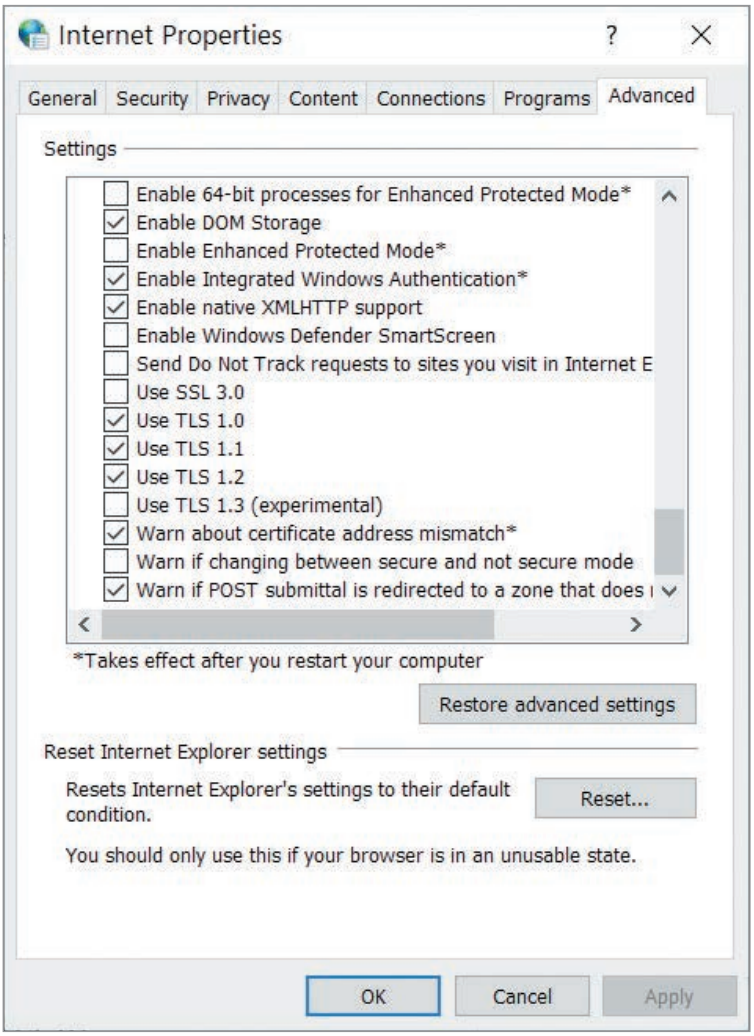
SSL and TLS are two public key encryption protocols commonly used to encrypt communications across the Internet. The two protocols can work side by side and are widely known collectively as SSL/TLS or TLS/SSL. All browsers today support SSL/TLS to secure transmissions of website traffic, which is what makes the difference between HTTP traffic and the more secure HTTPS traffic, or the difference between FTP (File Transfer Protocol) and FTPS (FTP Secure or FTP over SSL). Note, however, that SFTP (Secure FTP) relies on SSH, not SSL/TLS.

SSL was originally developed by Netscape and operates in the application layer. Since that time, the IETF (Internet Engineering Task Force), which is an organization of volunteers who help develop Internet standards, has standardized the similar TLS protocol. TLS operates in the transport layer and uses slightly

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

different encryption algorithms from SSL but, otherwise, it's essentially the updated version of SSL. SSL has now been deprecated and should be disabled whenever possible (Figure 6-27), leaving the more secure TLS to provide protection. In reality, you'll often find them both enabled for backward compatibility.

Figure 6-27 TLS is a more secure option for encryption



Note 6-4 | When requiring the more secure versions of network protocols, you might run into compatibility issues with older technology that does not support newer forms of encryption. When facing the problem of unsupported, secure protocols, consider upgrading the older equipment rather than compromising network security.

Key Management

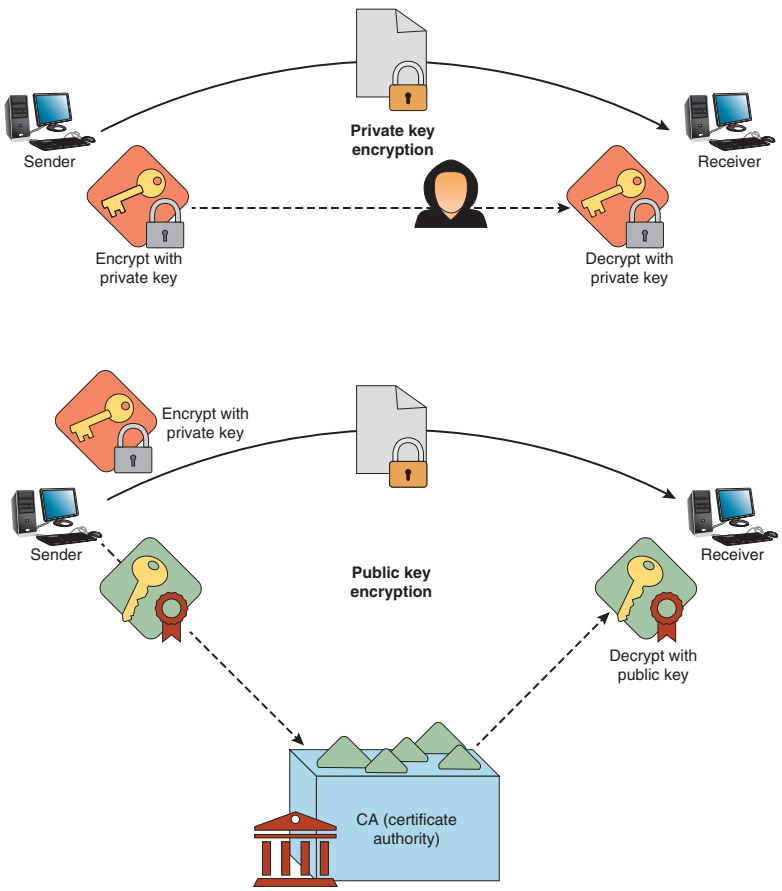
Many forms of encryption exist, with some being more secure than others. Even as new forms of encryption are developed, new ways of cracking their codes emerge, too. The most secure forms of encryption encode the original data's bits using a long and complex key—sometimes several times in different sequences—to scramble the data and, from it, generate a unique and consistently sized data block called ciphertext. The key is created and used according to a specific set of rules, or algorithms.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

You've used keys before, most commonly on your Wi-Fi network. WPA-Personal and WPA2-Personal Wi-Fi security rely on a PSK (preshared key) that is generated from your Wi-Fi passphrase, the name of your network, and other information so that each device has its own unique key. More complex and randomized keys can be generated by techniques such as ECC (elliptic curve cryptography), which uses elliptic curves like what you might have worked with in algebra, to generate a public and private key pair. The public key and the private key in the key pair are mathematically related in a way that is impossible to predict.

Key encryption can be divided into two categories (Figure 6-28).

Figure 6-28 Private key encryption versus public key encryption



- **Private key encryption**—Data is encrypted using a single private key that only the sender and receiver know. Private key encryption is also known as symmetric key encryption because the same key is used during both the encryption and decryption processes. A potential problem with private key encryption is that the sender must somehow share the key with the recipient without it being intercepted.
- **Public key encryption**—Data is encrypted with one key in the pair, and it's decrypted with a mathematically related key. If the encryption key is the private key and the decryption key is the public key, this ensures data integrity, as the sender's public key will only work if the data has not been tampered with or if the data came from the identified source. Alternatively (and more commonly), data can be encrypted with the public key and then can only be decrypted with the matching private key. This ensures data confidentiality, as only the intended recipient (the owner of the keys) can decrypt the data. Because public key encryption requires the use of two different keys, one to encrypt and the other to decrypt, it is also known as asymmetric encryption.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exam Tip | On the CompTIA Cloud+ exam, you might be given a scenario where you must choose whether it would be better to use a public key or a private key to encrypt data, and then which key (public or private) would be used to decrypt that data. In short, when performing asymmetric or public key encryption, you use one key to encrypt and the other to decrypt. Choosing which key is which depends on what you want to accomplish: data integrity or data confidentiality.

More specifically, if the owner of the keys is encrypting the data so everyone will know who performed the encryption, the owner will use their private key to encrypt, and everyone else could use the owner's public key to decrypt. Only the key owner's public key will successfully decrypt that data. This confirms the data owner's identity as well as the assurance the data has not been altered.

If, instead, someone else is encrypting data that only the owner of the keys should be able to read, that person would use the key owner's public key to encrypt the data. Then, only the key owner's private key could decrypt the data, ensuring its confidentiality. This last example is more common.

With the abundance of private and public key pairs, not to mention the number of places where each may be kept, organizations need simple and secure key management. On-prem, this is often handled by a KMS (key management system), such as Key Orchestration KMS by Fornetix or qCrypt by Quintessence Labs. Many cloud providers offer KMSs as well, such as AWS's KMS, Azure's Key Vault, and GCP's Cloud KMS.

Note 6-5 | If you'd like more information about how key encryption works, use your Cengage Unlimited subscription to go to *CompTIA Network+ Guide to Networks*, 10th edition, Chapter 4, and read the section titled "Encryption."

If you don't have a Cengage Unlimited subscription, you can find more information at cengage.com/unlimited.

Self-Check

9. Which of the following data states can be secured with encryption?
   - **a.** Data in use
   - **b.** Data in transit
   - **c.** Data at rest
   - **d.** All of the above

10. How many keys are required for asymmetric encryption?
   - **a.** None
   - **b.** One
   - **c.** Two
   - **d.** Three

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 6-5: Troubleshooting Cloud Security

Certification

2.5 Given a set of requirements, provision the appropriate cloud resources.

6.3 Given a scenario, troubleshoot security issues.

Security issues carry significant risks to a company, both from a technological standpoint and from financial, reputational, and legal perspectives. This section discusses some common security troubleshooting scenarios you might face when managing cloud resources.

Security Misconfigurations

Clearly, security technologies must adapt quickly to the demands of cloud computing and the cloud environment. The public in general and IT professionals in particular were long skeptical of the potential for securing data and workloads in the cloud. However, the development of new security techniques and the adaptation of tried-and-true security technologies has, in many cases, made the cloud more secure than other environments rather than less so.

Still, all these cloud security techniques rely on proper configurations and consistent maintenance. When traffic fails to arrive at its intended destination, often the problem is one of the following:

- A misconfigured or misapplied security policy or ACL (access control list) in a firewall or network security group
- A misconfigured or failed security appliance (such as an IPS, IDS, or WAF)
- Incorrectly administered micro-segmentation

Similarly, these misconfigurations can result in access that should not exist, thereby potentially opening a resource to attack. These problems are common vulnerabilities that inadvertently allow data exploits.

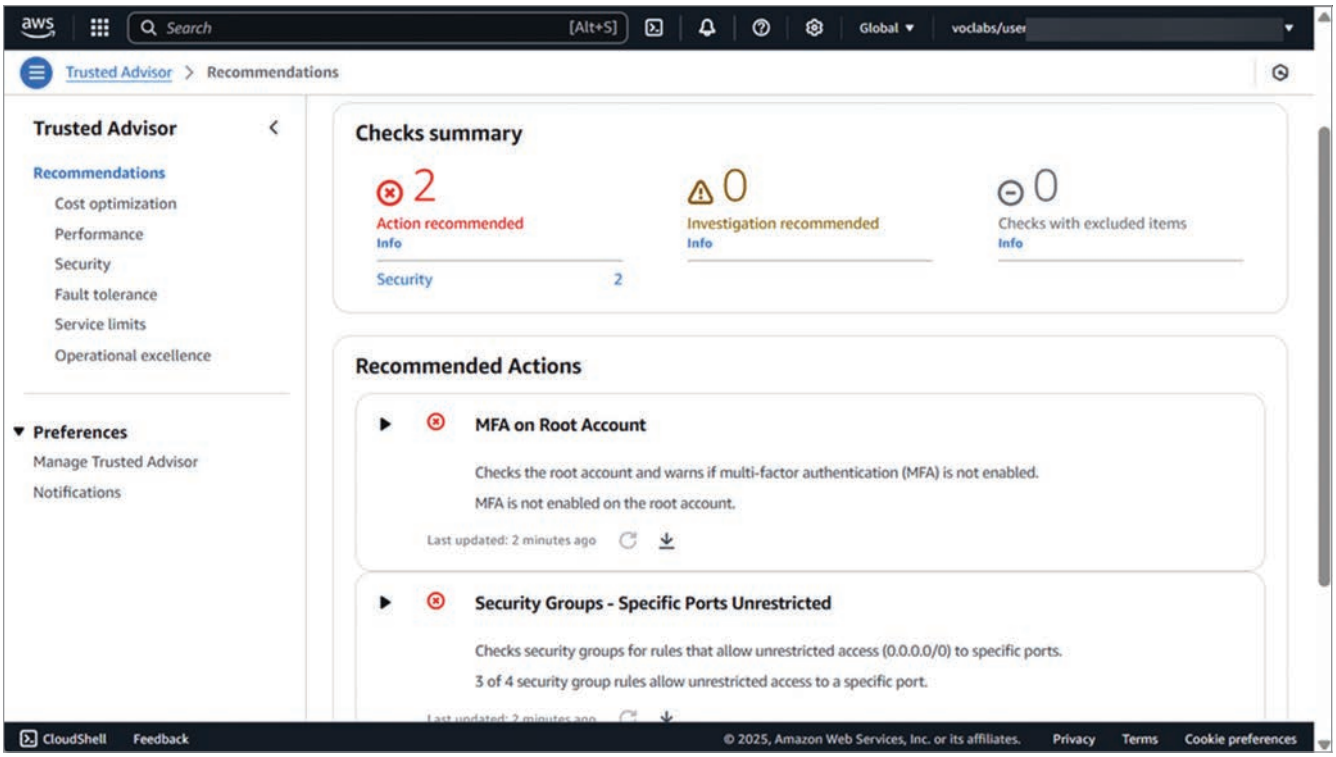
Tools within cloud platforms can help to protect your cloud resources by confirming best practices are applied to resource configurations. For example, AWS offers Trusted Advisor, which checks resources for security and availability configurations that do not conform to best-practice or compliance standards. It also checks for performance issues, cost optimization configurations, and relevant service limits. Figure 6-29 shows recommended actions in a student's AWS account hosted in an AWS Academy Learner Lab environment. Following the recommended actions would help improve security for the student's AWS resources.

A huge majority of security failures in the cloud result from user or administrative error. With this reality in mind, the following is a list of common security weak spots you'll want to be mindful of:

- **Unencrypted data and unencrypted communications**—Breaches happen. When they do, encrypted data can mean the difference between a learning opportunity and a career-breaking crisis. Make sure that all sensitive data, whether at rest, in motion, or in use, is securely encrypted using reliable encryption protocols.
- **Weak or obsolete security technologies**—As an IT professional, keep your own knowledge up to date on the latest developments in security techniques relative to your work. Every so often, cloud providers deprecate outdated protocols, features, or services. Although the CSP sends several notifications of what changes to expect and when, it's up to the cloud customer to ensure their cloud deployments are kept up to date with CSP security recommendations.
- **Incorrect hardening settings**—Most default system hardening settings at the instance level are effective. For example, Windows Server on a VM benefits from built-in defenses that typically don't need adjusting from their default settings. However, hardening settings within the cloud platform

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-29 Trusted Advisor recommended actions



Source: Amazon Web Services, Inc.

can be problematic. Oftentimes, default settings allow for maximized access while you're setting up your resources and testing your connections. Heed any warnings in your console indicating you need to tighten security settings, and be wary of areas where cloud admins tend to leave openings or backdoors so you can close these gaps.

- **Security device failure**—Security devices do fail sometimes. Failovers and redundant systems can minimize the impact of these events.
- **Insufficient security controls and processes**—Poorly designed networks can leave gaps where intruders can enter, or data might be exposed to the open Internet. When a security event occurs, insufficient logging can prevent you from learning what went wrong and how to fix the problem. Consider network design carefully, and employ logging, monitoring, and alerts where needed. You'll learn more about monitoring and alerts later in this course.
- **Unauthorized physical access**—Although the cloud consumer is not responsible for managing physical access to hardware supporting a public cloud, you and the other employees at your organization still need to be aware and vigilant of risks involving an attacker's physical access to the organization's trusted endpoint devices, such as workstations and mobile devices, and an attacker's ability to use social engineering to collect sensitive information that could be used to illegally access and manipulate cloud resources.

Key Management Issues

Good key management can make encrypting and decrypting data more convenient. However, problems arise if the keys are too convenient, such as being stored in the same location as the encrypted data or when storing keys temporarily in memory while they're being used. Keys should be kept in a protected, isolated environment with current backups stored securely in yet another location, and keys should be regularly rotated. (Don't forget to update backups when rotating keys.)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To support public key encryption in the cloud, you can store private keys either on-prem (and provide access to those keys where needed) or in the cloud through a key management service. While cloud-based key stores are highly available and resilient, you might still encounter certain issues, such as:

- Unavailable keys
- Failed connections
- Encryption operation failure

To fix common key store problems, confirm that needed keys are configured as available instead of unavailable or deleted. Also check the configuration of resources attempting to access stored keys and the configuration of network connections between these resources and the key store. These resources need permission to access any keys required to perform tasks assigned to those resources.

Self-Check

11. What security technique can protect inappropriately exposed data?

a. Device hardening
b. Encryption
c. Device failovers
d. Physical access controls

12. Which of the following security issues does the public cloud customer rarely need to worry about?

a. Incorrect hardening settings
b. Unauthorized physical access
c. Unencrypted data
d. Obsolete security technologies

Check your answers at the end of this chapter.

Summary

Section 6-1: Cloud Security Configurations

- Cloud providers make available a wide variety of tools to monitor for security threats (both internal and external) and, in many cases, automate responses to shut down attacks and recover from any outages or other damage caused by an attack.
- Vulnerability management is the process of identifying, assessing, and addressing security vulnerabilities in a system.
- Cloud security works according to the shared responsibility model, meaning the CSP is partially responsible for a cloud's security, and the customer is responsible for the rest of it.

Section 6-2: Virtual Network Security

- When a message crosses a firewall (physical or virtual), the firewall checks its rules to determine whether the message is allowed to pass, according to information such as port, protocol, and IP address or CIDR range.
- The zero trust security model is implemented in a ZTA, which ensures that traffic and access requests are evaluated at the most granular levels available.
- AWS offers NACLs, security groups, flow logs, traffic mirroring, and network firewalls to secure VPCs and the resources contained within them.
- In addition to NSGs and ASGs, Microsoft recommends that you establish a perimeter network in your Azure cloud

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to restrict and manage Internet traffic to your protected resources.

- Segmentation of GCP resources is first achieved at the project level, and then other techniques and tools are available for managing and monitoring traffic into and out of a VPC, including VPC firewall rules, VPC flow logs, and packet mirroring.
- Many of the same best practices apply for hybrid and multiclouds as for other cloud structures, such as encrypting data, monitoring activity on the network, and using segmentation and micro-segmentation to isolate certain resources; additional tools have been developed specifically to help bridge the security gap between the on-prem and cloud portions of an organization's networks or between different cloud platforms.

Section 6-3: Compute Security

- In addition to patches, updates, and anti-malware measures to protect VM instances from external threats, the cloud offers additional security options, depending on the specific tools available in a cloud platform.
- A special kind of firewall to protect SaaS applications is a WAF, which can be installed either on-prem or in the cloud and is designed specifically to protect web apps.
- Pipeline security should be automated as much as possible to protect development pipelines from security risks such as compromised code, unauthorized access, interference with security testing, and exposure of sensitive data or secrets (like access keys or encryption keys).

Section 6-4: Data Security

- Encryption is the use of an algorithm to scramble data into a format that can be read only by reversing the algorithm—that is, by decrypting the data.
- The most secure forms of encryption encode the original data's bits using a long and complex key—sometimes several times in different sequences—to scramble the data and, from it, generate a unique and consistently sized data block called ciphertext.

Section 6-5: Troubleshooting Cloud Security

- Tools within cloud platforms can help to protect your cloud resources by confirming best practices are applied to resource configurations.
- Keys should be kept in a protected, isolated environment with current backups stored securely in yet another location, and keys should be regularly rotated.

Key Terms

For definitions of key terms, check the Glossary.

access key
CIA (confidentiality, integrity, and availability) triad
CIS (Center for Internet Security)
CIS benchmark
cryptojacking
cryptomining
CVE (Common Vulnerabilities and Exposures)
CVSS (Common Vulnerability Scoring System)
CWE (Common Weakness Enumeration)
CWSS (Common Weakness Scoring System)
DDoS (distributed denial of service) attack
DLP (data loss prevention)
DoS (denial of service) attack
EDR (endpoint detection and response)
encryption
encryption key
hashing
IDS (intrusion detection system)
IPS (intrusion prevention system)
malware (malicious software)
metadata
micro-segmentation
NAC (network access control)
orphan resource
phishing
pipeline
privileged container
RACI (responsible, accountable, consulted, informed) matrix
ransomware
risk register
scanning scope
screened subnet
shared responsibility model
social engineering
unprivileged container
vulnerability management
WAF (web application firewall)
zero trust
zombie resource
ZTA (zero trust architecture)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

ACL access control list
AES Advanced Encryption Standard
CIS Center for Internet Security
CVE Common Vulnerabilities and Exposures
CVSS Common Vulnerability Scoring System
CWE Common Weakness Enumeration
CWSS Common Weakness Scoring System
DDoS distributed denial of service
DLP data loss prevention
IDS intrusion detection system
IPS intrusion prevention system
RACI responsible, accountable, consulted, informed
SSL Secure Sockets Layer
TLS Transport Layer Security
WAF web application firewall

Review Questions

1. How should you decide which security configurations to apply in a cloud deployment?
   a. Apply security first to the virtual network level, and then to the instance level.
   b. Apply security first to the instance level, and then to the virtual network level.
   c. Check CSP best practice recommendations.
   d. Use all available security tools for each layer.
2. What security technique can you use to control traffic to a web server separately from traffic control for the website database?
   a. NACL
   b. Micro-segmentation
   c. Traffic mirroring
   d. Screened subnet
3. Which of the following security rules would allow an incoming Echo Reply on a stateful firewall?
   a. Outbound rule: ICMP Allow to 0.0.0.0/0
   b. Inbound rule: HTTP Allow from 0.0.0.0/0
   c. Outbound rule: HTTPS Allow to 0.0.0.0/0
   d. Inbound rule: ICMP Deny from 0.0.0.0/0
4. Which security appliance can be used to secure an LMS (learning management system) application with some components deployed in a public cloud platform and some components deployed on-prem?
   a. CASB
   b. WAF
   c. APT
   d. EDR
5. Which of the following service types can automate key rotation?
   a. WAF
   b. NAC
   c. TLS
   d. KMS
6. Suppose you are running an application secured by client-side encryption and you receive an error stating CLUSTER_NOT_FOUND. Which of the following problems is most likely the source of the issue?
   a. Deleted key store
   b. Invalid credentials
   c. Encryption failures
   d. Connection misconfiguration
7. When considering the roles represented on a RACI chart, which role is the decision-maker if team members disagree on how to solve a problem?
   a. Responsible
   b. Accountable
   c. Consulted
   d. Informed
8. Which resource has secure access to application files on the host without requiring unnecessary vulnerabilities?
   a. Public container image
   b. Privileged container
   c. Private container image
   d. Unprivileged container

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9. Which tool analyzes custom code as it's written to identify vulnerabilities and misconfigurations?
   a. SAST
   b. SCA
   c. KMS
   d. WAF

10. Which of the following technologies can replace a screened subnet?
   a. CASB
   b. FWaaS
   c. ZTNA
   d. SWG

Scenario-Based Questions

Question 6-1

Krista runs a website that reports on current events in IT security. She recently noticed a huge spike in traffic that has caused several new web server instances to spin up in her AWS cloud. What kind of attack is Krista most likely experiencing?

a. User error
b. Data breach
c. Account hijacking
d. DDoS attack

Question 6-2

Roxie's credentialed scan revealed a known vulnerability in a recently updated application that is widely used across her company. Where can she find information on her company's policies for how to reconfigure the application to close the vulnerability?

a. Risk register
b. CVE
c. CISO
d. RACI matrix

Question 6-3

Zane is troubleshooting a problematic connection between a web server and a load balancer. The load balancer had reported the server was shut down, but when Zane checked it, the server instance was running fine. He suspects there's a routing problem or possibly an interface configuration problem. Where can he search for more information about the traffic that has been crossing this connection before the problem occurred?

a. Firewall rules
b. Flow logs
c. Security group settings
d. Load balancer configuration options

Question 6-4

Miguel's monthly cloud bill shows unexpectedly high charges from running VM instances. Further analysis reveals no new instances have been created in the past three months, but traffic was found that is consistent with cryptomining activities. What vulnerability was most likely taken advantage of for the attacker to accumulate charges against his account?

a. Cryptojacking
b. Metadata
c. Zombie instances
d. Phishing

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hands-On Projects

Note 6-6 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Project 6-1: Configure Security in AWS

Estimated time: 45 minutes
Objective 4.5: Given a scenario, apply security controls in the cloud.
Resources:

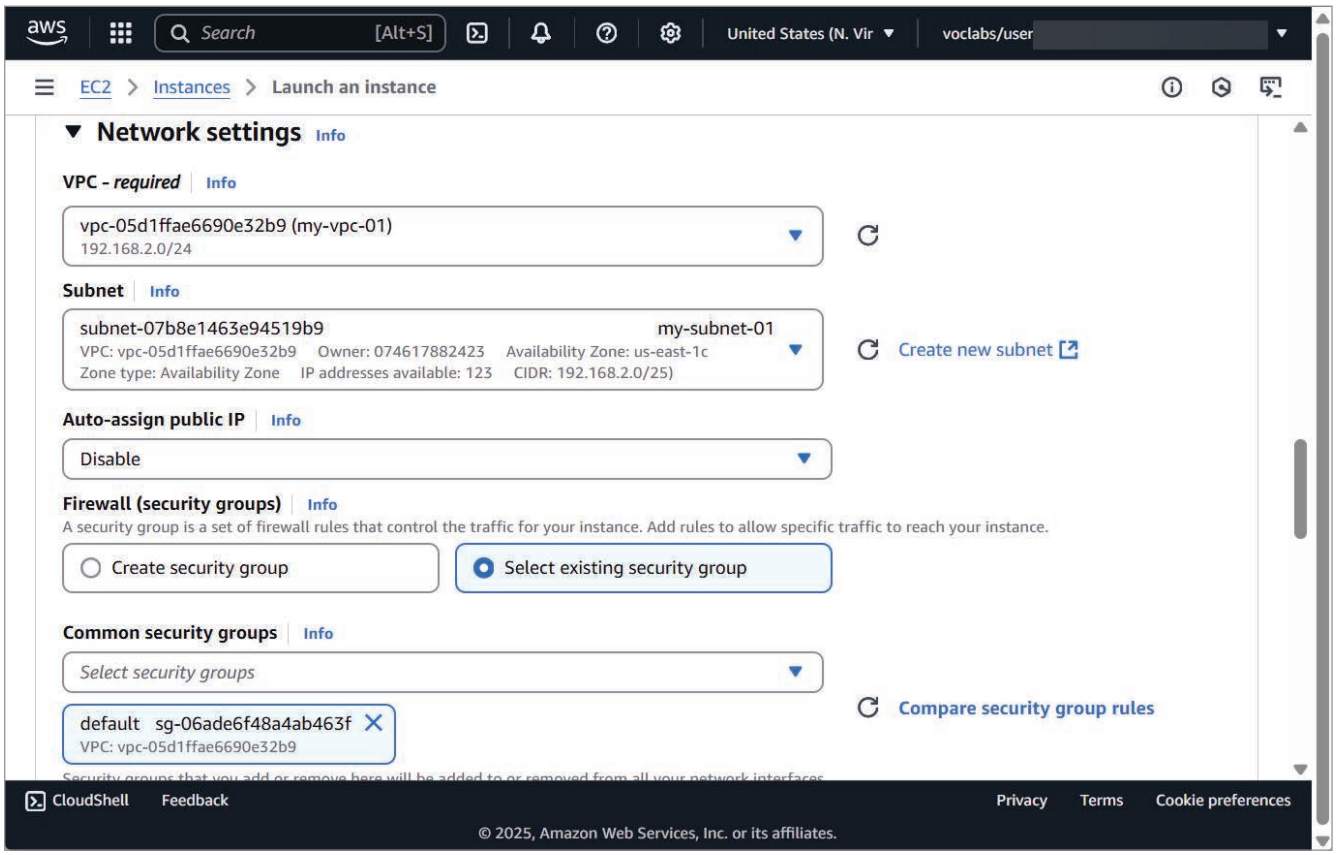
- AWS account
- Internet access

Description: AWS uses NACLs and security groups to manage traffic to and from resources in a VPC. In this project, you'll explore how to configure these tools to manage SSH and RDP traffic. Complete the following steps:

1. In your AWS Management Console, go to the **VPC** dashboard. In the navigation pane, scroll down and click **Network ACLs** in the Security section. You should have at least one NACL listed here as a default. Select that NACL (or if you have multiple NACLs listed, choose the one with active subnet associations).
2. Click the **Inbound rules** tab for that NACL. Two rules should be listed here, one that allows all traffic and one that denies all traffic. Considering that these rules seem to accomplish exactly opposite purposes, what net effect do you think these rules have on traffic into this VPC? Why do you think this is?
3. Click the **Outbound rules** tab and notice that it also shows two similar rules with similar configurations.
4. Click the **Subnet associations** tab. How many subnets are listed here?
5. In the navigation pane, click **Security Groups**. Unless you've been deleting security groups along the way, you should have a few groups listed here. Select a default security group and examine the inbound and outbound rules. Notice there is no setting for "Allow" or "Deny." Any traffic referenced by a rule is allowed. Traffic not referenced by a rule is not allowed. What traffic is allowed in this security group?
6. Deselect the default security group and select a security group created by a launch wizard. Launch-wizard security groups are created when you create a new VM. You've probably deleted the relevant VMs, but the security groups are still present unless you've been deleting these along the way. The traffic that is allowed in a launch-wizard security group is determined by the permission settings you configured on that VM, such as allowing SSH or RDP for remoting into the VM. What inbound traffic type is allowed for this launch-wizard security group? Based on this information, what OS do you think the VM had that this security group was created for?
7. Create a new VPC, a subnet within that VPC, and a VM instance in that subnet. Check the following details as you create the VM:
   - You will not need the VM's key pair.
   - Under Network settings, make sure to associate the VM with your new VPC and its subnet, not your account's default VPC.
   - Also under Network settings, choose to select an existing security group instead of creating a new security group. Select the default security group for your new VPC (Figure 6-30).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6-30 Network settings for VM



Source: Amazon Web Services, Inc.

8. While the VM is initializing, return to the VPC dashboard and check the rules created by default for your new subnet's NACL and the new VPC's default security group. It might be helpful to write down the last few digits of the subnet ID and VPC ID to identify the correct NACL and security group later. What traffic is currently allowed to this VPC, its subnet, and its VM instance?

Now you're ready to create a higher-priority NACL rule that will specifically allow inbound RDP or SSH traffic, depending on the OS you chose for your VM. Complete the following steps:

9. Select the NACL assigned to the VPC you created earlier in this project. On the NACL's Inbound rules tab, note the priority (indicated by its rule number) for the existing rule, and then click **Edit inbound rules**. Click **Add new rule**. For the new rule, choose a higher priority rating than what the existing rules have—remember that higher priorities use lower numbers. What priority did you choose?
10. Under Type, what are two protocols you recognize in this list?
11. Choose **SSH** or **RDP**, depending on the OS you chose for your instance. Default settings should accept traffic from any source (0.0.0.0/0) and will allow this traffic. Save your changes.

Next, you're ready to create a new security group that only allows inbound RDP or SSH traffic, depending on the OS you chose for your VM. Complete the following steps:

12. On the Security Groups page, click **Create security group**. Give the security group a name, such as MySG-1, and a description, such as "RDP traffic only" or "SSH traffic only." Assign the security group to the VPC you created earlier in this project.
13. Under Inbound rules, click **Add rule**. Under Type, select SSH or RDP, depending on your instance's OS. Under Source type, select **My IP**. What happens?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14. Give the rule a description that would help you identify its purpose in a long list of security rules. This is good practice for later when you'll be managing many NACL and security rules, and you won't necessarily remember why you created any specific rule. What description did you give the rule?
15. Finish creating your security group.
16. Return to the EC2 dashboard, and check the detailed information for your instance, which should be running by now. What is the instance's private IP address?
17. Does the instance have a public IP address? Thinking back to what you've learned over the past few chapters, why do you think this is?
18. Change the instance's security group to the one you just created. To do this, click **Actions**, click **Security**, and click **Change security groups**. What security group is currently assigned to the VM?
19. Select and add the new security group. Are you able to assign both security groups to this one VM?
20. Save your changes. **Take a screenshot** showing both security groups listed on the VM's Security tab; submit this visual with your responses to this project's questions.
21. When two security groups are assigned to a VM, both groups' rules apply to determine what traffic is allowed to reach the VM. Considering the current configuration, respond to the following questions:
    - What traffic is now allowed to reach the instance?
    - What would you need to change to allow *only* RDP or SSH traffic to reach the instance and no other traffic?
22. Scroll through the EC2 service's navigation pane. Of the security resources you used in this project, which one can you manage from the EC2 Dashboard?
23. Delete all the resources you created in this project—including the new VM, the VPC you created (***not your default VPC!***), its subnet, its NACL, and both its security groups. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
24. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 6-7 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 6-2: Configure Security in Azure

Estimated time: 45 minutes
Objective 4.5: Given a scenario, apply security controls in the cloud.
Resources:

- Azure account and subscription
- Internet access

Description: Azure uses NSGs and ASGs to manage traffic to and from resources in a VNet. In this project, you'll explore how these tools relate to each other and how to configure NSG rules to allow SSH traffic. Complete the following steps:

1. In your Azure portal, create a VNet using the following details:
   - Create a new resource group.
   - Give the VNet a /24 CIDR range, such as 192.168.0.0/24.
   - Give the default subnet a /26 CIDR range, such as 192.168.0.0/26.
2. In the search box, type **network security groups**, and then click **Network security groups** in the search results.
3. Click **Create network security group**. Place the NSG in your project's resource group that you created in Step 1, and then finish creating the NSG.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. After the NSG is created, go to the resource to find information about it. Respond to the following questions:
   - What inbound and outbound security rules are listed in your NSG?
   - How many network interfaces are associated with this NSG?
   - What subnets are currently associated with this NSG? Why do you think this is?

Notice that NSGs are associated with either network interfaces or subnets. Microsoft recommends using subnets rather than network interfaces as a general rule, so now you're ready to associate your subnet with this NSG. Complete the following steps:

5. On the NSG's subnets list, click **Associate**. Make sure your VNet and its subnet are selected. Click **OK**.

Note 6-8 | If your VNet is not at first available in the Virtual network dropdown list, wait a few minutes and refresh the page, and then try again.

6. Return to the Virtual networks page and click your VNet. Click **Settings** and then click **Subnets**. Scroll to the right to display all the information about your subnet. Is your NSG listed here?
7. In the search box, type **application security groups**, and then click **Application security groups** in the search results. Click **Create**.
8. Place the ASG in your project's resource group that you created in Step 1 and create the ASG.
9. Create a new Linux VM using the following details:
   - Add the VM to the resource group you created for this project.
   - Use the **Standard_B1s** size.
   - Use a password authentication type for simplicity in this project.
   - Under Inbound port rules, make sure **Allow selected ports** is selected, and make sure **SSH (22)** is selected.
   - Confirm the VM will be added to your project VNet and subnet.
   - On the Networking tab in the NIC network security group section, what does Azure say about the VM's NSG?
   - Confirm **None** is selected for NIC network security group so the existing NSG on the subnet will protect the new VM's NIC instead of creating a new NSG.
   - Review and create the VM.
10. Open the VM's information page. Click **Networking** on the left and then click **Network settings**. Scroll through this information and respond to the following questions:
    - What NSG is protecting this VM?
    - How many inbound port rules are currently listed?
    - Why do you think this is?
11. At the top of this page, click **View topology**. What is the NSG connected to?
12. **Take a screenshot** of the topology; submit this visual with your responses to this project's questions.
13. Close the Topology blade. On the VM's Networking blade, click **Application security groups** in the left pane. Click **Add application security groups**. Select the ASG you created in Step 8 and save your changes. Respond to the following questions:
    - Based on this configuration, what is the ASG associated with?
    - In contrast, what is the NSG associated with?
14. Return to the details page for your NSG, click **Inbound security rules**, and then click **Add**. Configure a new inbound security rule using the following details:
    - Allow traffic from any source.
    - Leave the Source port ranges as * because port filtering is typically performed on the destination port in Azure, not the source port. If you're feeling adventurous, change the Source port ranges field to 22,

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and read the caution message at the bottom of the blade. Then change the Source port ranges field back to * before proceeding.
- For Destination, notice that you can choose an IP address, service tag, or ASG as the destination for traffic covered by this rule. Select **Application security group** and select the ASG you created in Step 8.
- Change the Service to **SSH**.
- Allow traffic with priority **100** and name **SSH_traffic**.
- Click **Add**.

15. Return to the details page for your VM, click **Network settings**, and then click **Refresh**. How many inbound port rules are listed now?
16. Click the **SSH_traffic** rule. What warning does Azure give you about this rule?
17. Delete all the resources you created in this project, including the VNet, subnet, NSG, ASG, VM, and resource group. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
18. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 6-9 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 6-3: Configure Security in GCP

Estimated time: 45 minutes (+15 minutes for group work, if assigned)
Objective 4.5: Given a scenario, apply security controls in the cloud.
Group work: This project includes enhancements when assigned as a group project.
Resources:

- GCP account
- Internet access

Description: GCP uses firewall rules to manage traffic to and from resources in a VPC. In this project, you'll explore how to configure these tools to allow you to connect to resources in a custom VPC and ping between those resources. Complete the following steps:

1. In your GCP console, create a new project. In that project, create a new, custom-mode VPC and one subnet within that VPC.
2. Create a new Linux VM in the VPC you just created, using the following information:
   - Use the same region as the VPC you created in Step 1.
   - If you need to avoid accruing charges, use the e2-micro machine type in one of the free regions. (Search online for a current list of regions that support free instances in GCP.)
   - During the creation process, on the **Networking** page under Network interfaces, be sure to expand the configuration section to edit the network interface and select the VPC you created in Step 1.
   - Again, to help avoid charges, select the Standard network service tier, if available in the region you selected.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. Once the VM is running, attempt to console into it by clicking **SSH** in the list of instances. What happens? Why do you think this is?
4. In the navigation menu, point to **VPC Network** and click **Firewall**. What ingress rules are listed here, what are their purposes, and what network do they apply to?
5. Click **Create firewall rule**. Configure the following details:
   - Give the rule a name, such as ssh-myvpc, and a good description.
   - Select your network that you created in Step 1.
   - Leave the priority at **1000**, the direction of traffic as **Ingress**, and the action as **Allow**.
   - Under Targets, select **All instances in the network**.
   - For Source IPv4 ranges, enter **0.0.0.0/0**, which is a wildcard that allows traffic from any source.
   - Under Protocols and ports, select **Specified protocols and ports**, select **TCP**, and enter **22**. Why do you think 22 is the appropriate number for this field?
   - Click **Create**.
6. Return to your VM instance and try again to console into it. What happens this time? Why do you think that is?
7. Leave this connection open and create a second VM in the same region, VPC, and subnet. Attempt to console into the second VM. Does it work? Why do you think this is?
8. Attempt to ping from one VM to the other. Wait several seconds, then press **Ctrl+C** to display the results. What happens? Why do you think this is?
9. Create a new firewall rule that allows ICMP in your custom VPC. ICMP does not have a TCP or UDP port, so you'll have to enter **icmp** in the "Other" field when specifying protocols and ports. Otherwise, use the same settings as in Step 5.
10. Try again to ping from one VM to the other. Does it work? Why do you think this is?
11. Position each of the VM SSH connection windows to show both active connections and the successful ping results in one window. **Take a screenshot** showing the two windows; submit this visual with your responses to this project's questions.
12. **For group projects:** Share the public IP address of one of your VMs with a member of your group. When you receive the public IP address of your group member's VM, attempt to ping that address from one of your VMs. Each group member should respond to the following questions:
    - Does the ping to your teammate's VM work? Why do you think this is?
    - What would you need to change in your firewall rules so your VMs could ping each other, but sources on the Internet could not ping your VMs?
    - Make those edits to your firewall rule and have your teammate try again to ping your VM. Does it work? Can your VMs still ping each other?
    - Troubleshoot as needed to edit the rule so your VMs can ping each other, but your teammate cannot ping your VMs.
13. Go to the list of projects and shut down the project you created in Step 1. Confirm the project is shut down and scheduled for deletion.
14. Make sure to properly sign out of your account.

Note 6-10 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Challenge Projects

Note 6-11 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 6-1: Configure Security Across VPCs

Estimated time: 1 hour
Objective 4.5: Given a scenario, apply security controls in the cloud.
Resources:

- One or two public cloud accounts (This project can be completed entirely in one public cloud platform, or the instructor can require that students set up each VM on a different public cloud platform and establish communication across clouds.)
- Internet access

Description: In the Hands-On Projects for this chapter, you created security rules that allowed traffic from any source to help with initial testing. In this project, you will restrict the traffic source to a single VM instance.

Imagine you work for a small company that is building its first website hosted in the cloud. The company will have one VM as a web server and a second VM in a different VPC or VNet as a database server. The web server should be accessible from the open Internet, but the database server should only allow communication with the web server.

Before opening ports for web traffic and database traffic, you want to ensure you have your security rules set appropriately. You decide to start with pings, which use ICMP and are easy to test. You'll need a source VM (your web server instance) and a target VM (your database server instance) in two separate VPCs or VNets. You need to prove that you can ping your target VM from your source VM but not from your local computer. While there are more secure ways to do this, in this scenario, you will place the target VM in a public subnet so that you can determine how to use security rules to control traffic to the target VM.

Complete the following steps:

1. If you need to create a project or resource group to manage project resources when you're finished, do so now.
2. Create two VPCs with one subnet each, either in the same cloud platform or in two different cloud platforms. (Your instructor might require that you use the same cloud platform for this project or two cloud platforms—check with your instructor for specific requirements.) Make sure both subnets have access to the Internet. For example, in AWS, you'll need to add an Internet gateway to each subnet and add an Internet route to each subnet's route table.

Note 6-12 | Recall that in some cloud platforms, you need a route to a gateway so that resources inside a subnet can access the Internet. For example, in AWS, a public subnet requires a route in its route table to an IG (Internet Gateway). You learned how to do this in Chapter 4, "Cloud Networking." Return to that chapter or research online for more detailed steps.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. Create a VM instance in each subnet. Make sure each VM receives a public IP address. What public IP address is assigned to each of your VMs?
4. Configure security rules that allow ICMP traffic from any source to your target VM so you can confirm the target VM will respond successfully to pings. Get a working ping from your local computer to your target VM before proceeding.
5. Configure security rules that allow a SSH or RDP connection (depending on your VM's OS) to your source VM. Remote into the source VM.
6. Configure security rules that only allow ICMP traffic from your source VM to your target VM. What rules did you add? What effect do you expect each rule to have on traffic to and from each VM?
7. Run a ping from the source VM to the target VM that shows ICMP traffic is reaching the target VM. **Take a screenshot** showing the output; include this visual with your responses to this project's questions.
8. Run a ping from your local computer to your target VM that shows ICMP traffic from other sources cannot reach the target VM. **Take a screenshot** showing the output; include this visual with your responses to this project's questions.
9. Delete all resources you created in this Challenge project. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 6-13 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Solutions to Self-Check Questions

Section 6-1: Cloud Security Configurations

1. Which of these threats could be mitigated with good backups?
   Answer: a. Data loss

   Explanation: Data loss can generally be avoided by keeping good backups. Backups won't address the problems caused by an insider threat, account hijacking, or a data breach.

2. Which cloud service model places the greatest responsibility on the cloud customer for cloud security?
   Answer: c. IaaS

   Explanation: With an **IaaS (Infrastructure as a Service)** cloud service model, the customer maintains a higher degree of responsibility for the management, configuration, and security of a cloud deployment. With a PaaS (Platform as a Service) model, more of the responsibility is shifted to the cloud provider, and with a SaaS (Software as a Service) model, even more of the responsibility moves away from the cloud customer. RaaS (Ransomware as a Service) is a business model for cybercriminals.

3. Which source would provide an ID for a vulnerability in a specific web app?
   Answer: a. CVE

   Explanation: The **CVE (Common Vulnerabilities and Exposures)** provides a tracking system that assigns an ID to each identified vulnerability. The CVSS (Common Vulnerability Scoring System) is used to rate vulnerabilities on a scale from 0 to 10. The CWE (Common Weakness Enumeration) is a standardized taxonomy of weaknesses, not vulnerabilities. The CWSS (Common Weakness Scoring System) is used to rate weaknesses, not vulnerabilities.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 6-2: Virtual Network Security

4. A stateless rule applies to _____.

Answer: c. traffic in one direction

Explanation: Stateless security rules apply to **traffic in one direction** at a time—either incoming traffic or outgoing traffic—but not both at the same time.

5. What cloud security tool functions like Wireshark?

Answer: b. Traffic mirroring

Explanation: Traffic mirroring or packet mirroring provides deeper insight into captured traffic than what flow logs offer, and it's similar to the packet capture and analysis functions that Wireshark offers. Virtual firewalls and security groups provide security by filtering traffic according to security rules.

6. Security at which of these levels provides the most fine-tuned traffic filtering?

Answer: b. Interface

Explanation: Virtual networks can contain several subnets, each of which can contain several instances, each of which can contain several interfaces. Security configured at the **interface** level provides the most fine-tuned traffic filtering.

Section 6-3: Compute Security

7. Which security technique will increase the number of running resources while decreasing the attack surface?

Answer: b. Create single-function resources

Explanation: Creating single-function resources increases the number of smaller-sized resources while leaving a much smaller attack surface for each resource. Disabling unnecessary ports and services, installing host-based firewalls, and planning subnet placement can all increase security but do not affect the number of running resources.

8. Which security tool can you use to find security vulnerabilities in forgotten EC2 instances?

Answer: c. EDR

Explanation: EDR (endpoint detection and response) solutions detect and monitor endpoints in the cloud to increase visibility of attack surfaces from forgotten or outdated instances. An SCA (source composition analysis) tool scans imported code for vulnerabilities. HIDS (host-based intrusion detection service) monitors its host VM for signs of malicious behavior, but it requires updates in order to run and function correctly, which would not be available on a forgotten EC2 instance. A SAST (static application security testing) tool scans code written in-house for vulnerabilities.

Section 6-4: Data Security

9. Which of the following data states can be secured with encryption?

Answer: d. All of the above

Explanation: Encryption can be used to secure data in **all of these states**. Encryption is often used to protect data at rest and data in transit. Special forms of encryption can also be used to secure data in use, though this is uncommon.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10. How many keys are required for asymmetric encryption?

Answer: c. Two

Explanation: Asymmetric encryption, also called public key encryption, requires the use of **two** different but mathematically related keys, one to encrypt and the other to decrypt.

Section 6-5: Troubleshooting Cloud Security

11. What security technique can protect inappropriately exposed data?

Answer: b. Encryption

Explanation: Encrypting data can mean the difference between a learning opportunity and a career-breaking crisis. Device hardening and physical access controls can help prevent the exposure of data but will not help once data has been exposed. Device failover helps in the event of an outage or failure.

12. Which of the following security issues does the public cloud customer rarely need to worry about?

Answer: b. Unauthorized physical access

Explanation: The cloud consumer is not responsible for preventing **unauthorized physical access** to hardware supporting a public cloud, although they still need to be vigilant of risks related to endpoint devices or social engineering. All the other issues—incorrect hardening settings, unencrypted data, and obsolete security technologies—continue to be relevant concerns for public cloud customers, who retain a shared responsibility for the security of their cloud-hosted resources.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Chapter 7

Identity and Access Management

Chapter 7 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Describe how accounts are managed in the cloud.
2. Explain authentication processes used in the cloud.
3. Apply IAM (identity and access management) policies in the cloud.
4. Explain technologies used for hybrid cloud IAM.
5. Identify common IAM issues.

Chapter 7 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 7-1: Cloud Accounts
- Identity
- Account Types
- Account Management

Section 7-2: Authentication
- Authentication Processes
- Password Policies
- MFA (Multifactor Authentication)
- Certificate-Based Authentication
- Programmatic Access
- Securing Secrets

Section 7-3: Authorization
- Authorization Processes
- AWS IAM
- Azure IAM
- GCP IAM

Section 7-4: IAM for Hybrid Clouds
- SSO (Single Sign-On)
- AWS IAM in a Hybrid Cloud
- Azure IAM in a Hybrid Cloud
- GCP IAM in a Hybrid Cloud

Section 7-5: Troubleshooting Cloud IAM
- Common IAM Issues
- Troubleshooting Privilege Issues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 7 Scenario "Malik, did you put these coffee beans on my desk?" Mika, the cybersecurity specialist, walks out of her office carrying a small package of chocolate-covered coffee beans.

"Nope, wasn't me," replies Malik. "And you know *I'm* not your secret pal because *I* know you don't like coffee."

"Well, I did list chocolate as a favorite snack on my questionnaire. Maybe I should have specified coffee-less chocolate." Mika hands Malik the coffee beans and says, "Here, you can have them," as her phone begins to ring and she darts back into her office.

The departments in your building are running a secret pal event this month, and you drew Mika's name. You're aware she doesn't like coffee, but you're hoping the coffee beans will throw her off your trail because she was beginning to suspect her secret pal was someone in IT. She's a former police detective, and her nose for finding clues is legendary throughout campus. No one wants Mika's name for a secret pal because no one has successfully kept the secret from her for more than 16 days. You're determined to set a new record, maybe even get through the whole month.

Malik glances over at you and says, "She knows."

"She doesn't know!" you whisper back. "How could she know? She thought it was you!"

"No, she wants you to *think* she thought it was me," says Malik. "She's playing you."

"Whatever," you reply. "We're only eight days into the month. There's no way she could know already."

"She knows," repeats Malik as he pops a coffee bean into his mouth. "You gave it away when you gave her those *Star Wars* stickers. Only us IT geeks know she likes *Star Wars* and she knows I'm a full-blood Trekkie. You're so busted."

Just then, Kendra walks in and says, "Okay, team. I've set up accounts in two public clouds, and I'm going to give each of you an admin user account so you can play around with this stuff. What thoughts do you have on where you'd like to start?"

"I want to spin up some VMs," Malik volunteers. "I've not worked much yet with Windows Server."

"I'd like to look at the networking services," you reply, "and maybe figure out how to set up users, too. How does that work in the cloud? Do users sign into their workstations like they do now, or do they go to AWS or Azure or whatever and sign in there?"

"That's a good question," says Kendra. "It really depends on what the user needs to be able to do."

"How do we know who is who?" asks Malik. "Do we have, maybe, an HR account that all of HR uses, and they share the password? Or do we have a different account for each person?"

"Each person typically has their own account, but it's a special kind of account," explains Kendra. "Also, we won't use only passwords. That's not secure enough. I'll check with Mika on what kind of authentication we'll want to set up." Then she adds, "Each user gets a different kind of role, depending on their job duties and the kinds of information they need. It's kind of like our secret pal thing—it's almost like you can tell who a person is based on what they know about you."

Malik gives you "the look," and you try not to roll your eyes. Kendra pretends not to notice.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

"Let's find out a little more about how identity management works in the cloud," Kendra suggests. "Here are some more questions for you two," as you make some notes on your notepad:

- What kinds of accounts does the cloud offer?
- How can users prove their identity in the cloud?
- How do admins control what resources users can access?
- How is access control managed across a hybrid cloud?
- What kinds of problems are typical when managing cloud accounts?

Section 7-1: Cloud Accounts

Certification

2.5 Given a set of requirements, provision the appropriate cloud resources.

4.4 Given a scenario, apply security best practices.

When you first create a new cloud account, this account takes the role of the **root** user, which has full administrative privileges throughout all services in the account. For example, if you created an account in AWS in Chapter 1, "Introduction to Cloud Computing," the account you created was configured as a root account. As you've practiced performing tasks in AWS in later chapters, you've likely performed most of these tasks in the root account. From a security perspective, this is a bad idea. It's okay for the practice activities you've been doing in this course, but you don't want to use this approach in the work world.

When working for an organization, you'll instead use less privileged user accounts. These user accounts can also allow other users in your organization to access and interact with cloud resources. You can set up these accounts using **IAM (identity and access management)** services built into most major cloud platforms. AWS, Azure, and GCP all offer many IAM services. Once these user accounts are established, the root account should not be used again except when absolutely necessary to perform tasks that only the root user can complete.

Note 7-1 | Keep in mind that each VM (virtual machine) instance also has one or more default accounts in its OS, including privileged root accounts. As you create new VMs, you'll need to deactivate default accounts in each VM's OS or, at the very least, change default credentials to use secure authentication methods.

IAM works at multiple levels to control users' access to resources: identification, authentication, authorization, and accounting. You read about each of these levels of access management throughout this chapter. Let's begin with identification.

Identity

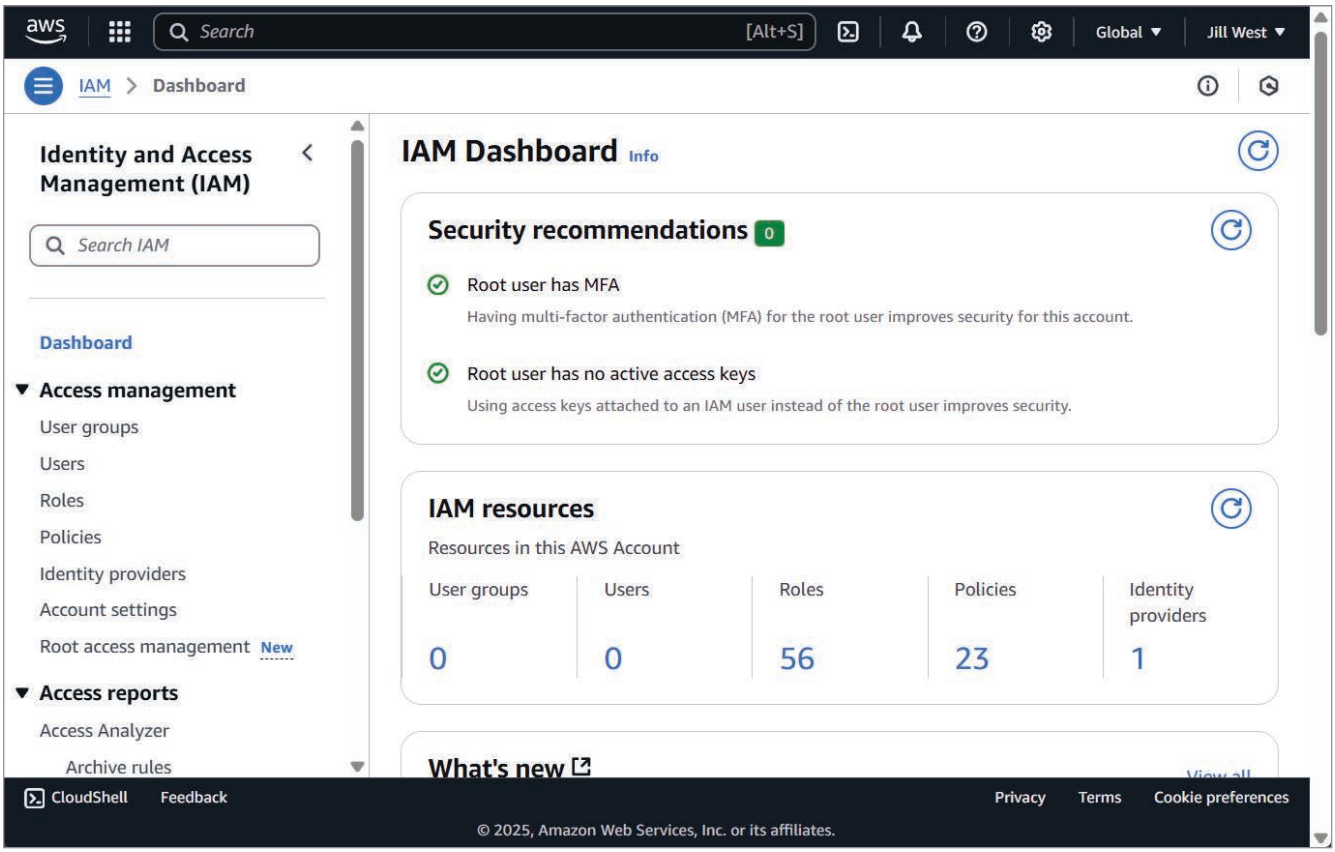
IAM policies and tools provide directory and identity services that are applied broadly to all users who are given any kind of access to resources, both on-prem and in the cloud. Access control is managed through permissions granted to identities. Users and resources (such as an application) both can have an **identity**, which is a digital entity to which you can grant permissions. In contrast, a resource such as a server, device, or application might need an identity, for example, to access information in a database or to access data stored in logs. Accounts assigned to resources (such as a server instance) are sometimes called **service accounts**, while a human user is given a **user account**.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

You can create accounts through the IAM dashboard in your cloud platform (Figure 7-1). Here, you set the username, password, and permissions or create some other kind of credentials such as access keys or a certificate. You might require that users reset their passwords when they first sign in, and you might require certain standards for these passwords, such as length and complexity rules. Afterward, every time a user signs in with these credentials, the user is connected with the permissions granted to the user account.

These identities and their credentials are stored in a repository called an identity vault, which might draw information from other authoritative sources (called sources of truth) and becomes the centralized source of truth for authentication and authorization functions in this cloud platform.

Figure 7-1 AWS IAM dashboard



Source: Amazon Web Services, Inc.

Account Types

Before digging deeper into the discussion of account management in the cloud, it's important to clarify the different types of accounts that need to be managed and the various ways users will access resources in the cloud. When you create a free account with a CSP such as AWS, the account you create is generally a standard administrative account; that is, it's an account intended to be used by admins, developers, and cloud operators to create, manage, and work directly with cloud services (such as IAM) and cloud-hosted resources (such as VMs). Because it is a root user, it can be used to perform nearly any action in the account, such as creating resources, firewall rules, or other users, and it can limit the permissions granted to these other users. However, creating other users in a standard account such as this typically would only be needed for managing workload access to resources or possibly for very small teams. It would not be used for managing large numbers of users.

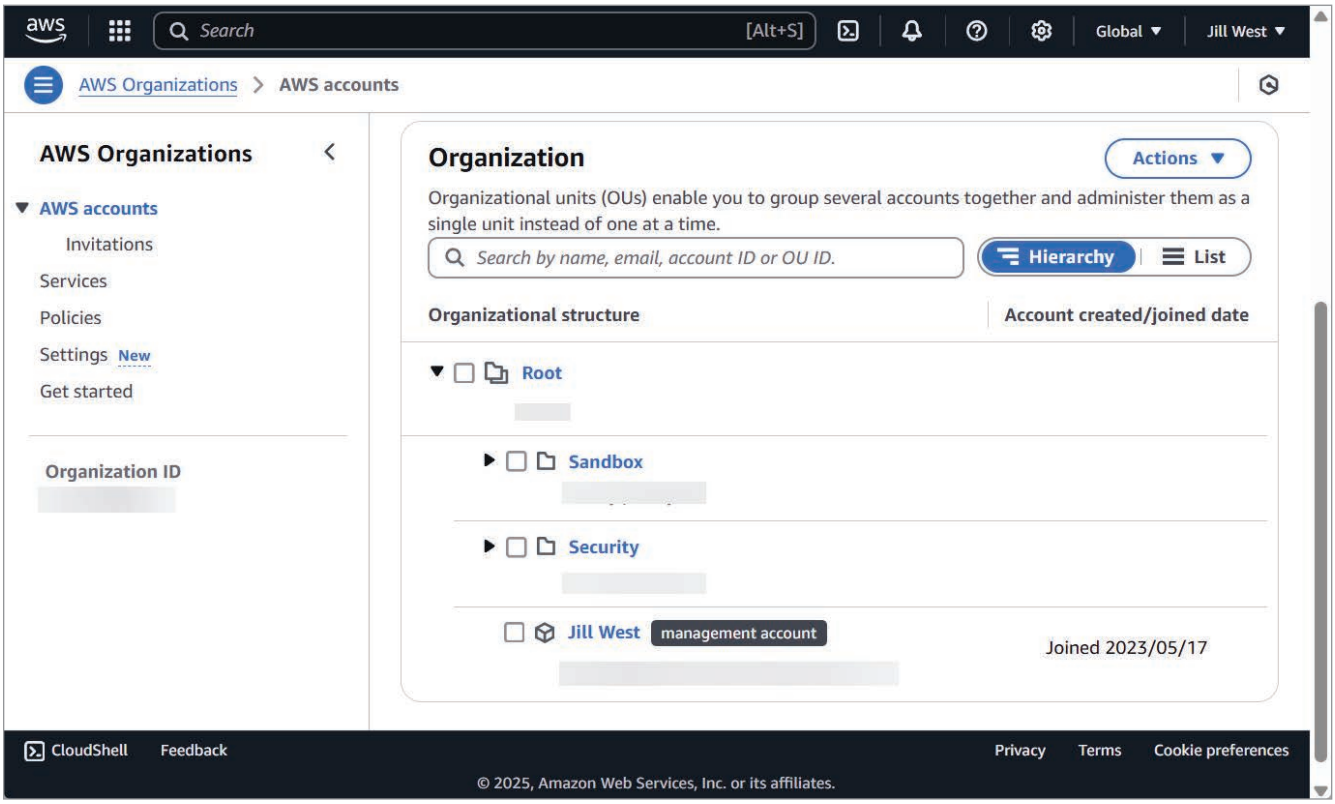
Organizations

User management from within a single, standard account is not scalable in the sense that it would become very difficult to handle many users with only these tools. Instead, other user management features and services are needed to organize and support users in larger groups or organizations. For this purpose, most

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

public cloud platforms now offer newer identity and access control services that are designed to organize many accounts across a large team or company. Typically, this service is called something like Organizations (Figure 7-2). An **organization** resource is an identity management object at the top of a hierarchy of accounts that is used to control access to all of a company's resources in that cloud. The organization, then, can be used to manage many accounts and users as well as apply policies across accounts, enforce security requirements throughout the company, and simplify billing for the entire organization.

Figure 7-2 An organization in AWS



Source: Amazon Web Services, Inc.

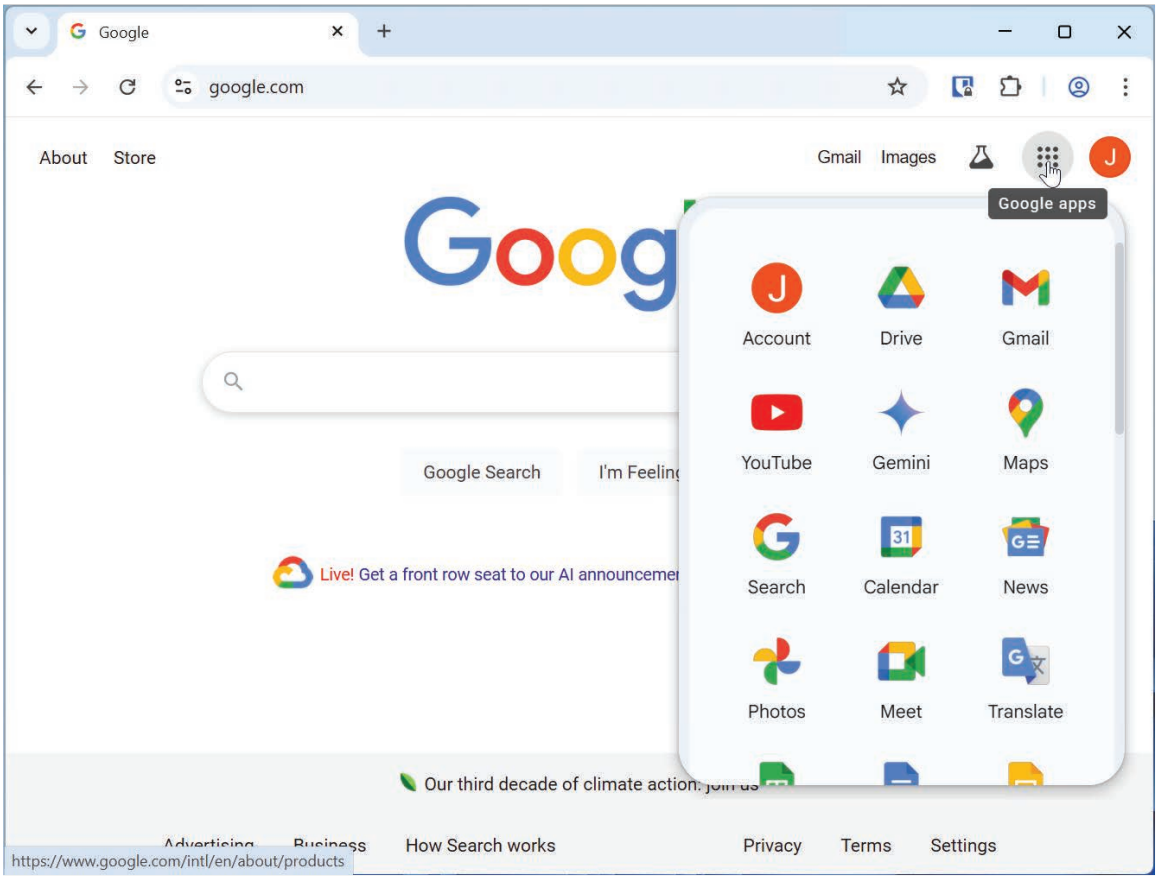
Consumer Accounts

In addition to the oversight provided by an organization resource, some types of cloud accounts additionally integrate with other types of systems. Consider the functionality provided by a Google account. If you've created a Gmail account, you might have explored the tools and web apps made available through that account. For example, with a free Gmail account, you also have free access to other Google services such as Google Maps, Google Drive, Google Docs, and Google Meet (Figure 7-3). Similarly, a Microsoft account gives you free access to OneDrive web storage and to web apps such as Word Online and Excel Online. You can access these same web apps on your mobile device and even install a mobile version of the apps for more ease of use on a small touchscreen. You might have used your Microsoft account or your Google account to sign up for a free trial on the Microsoft Azure or Google Cloud platforms. Licensed versions of these accounts can also be managed at an organizational level, such as a school or employer, for access to additional features, services, and software that are paid for by that organization.

Unlike Microsoft and Google, AWS does not have an extensive catalog of its own web or mobile apps made available for free to nontechnical users. When you create an AWS account, it's designed primarily for use in the AWS cloud to build your own cloud deployments. However, AWS does offer a service where developers can manage large numbers of nontechnical end users of the developers' web or mobile apps deployed on AWS-hosted resources. Amazon Cognito is a CIAM (customer IAM) service where developers can give each external user a **consumer account** with access to web or mobile apps without creating AWS accounts for all these users.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-3 Google apps



Source: Google LLC

Consider a mobile app with a few hundred thousand users, only some of which are paid subscription users. You wouldn't want to create and manage AWS accounts for all these people. Not only would it be difficult to create all these accounts, it would also be painful to manage, secure, and track individual expenses for this many users, especially considering the fact that many people sign up for an account with a free app and then never really use it. In Azure or GCP, you could tap into the Microsoft or Google account management tools and have customers use the relevant account to access your app. But with AWS, you need a dedicated service for this purpose. To fill this need, Cognito offers many relevant features, such as MFA (multifactor authentication), passwordless authentication, self-registration, fine-grained authorization, and specialized security features.

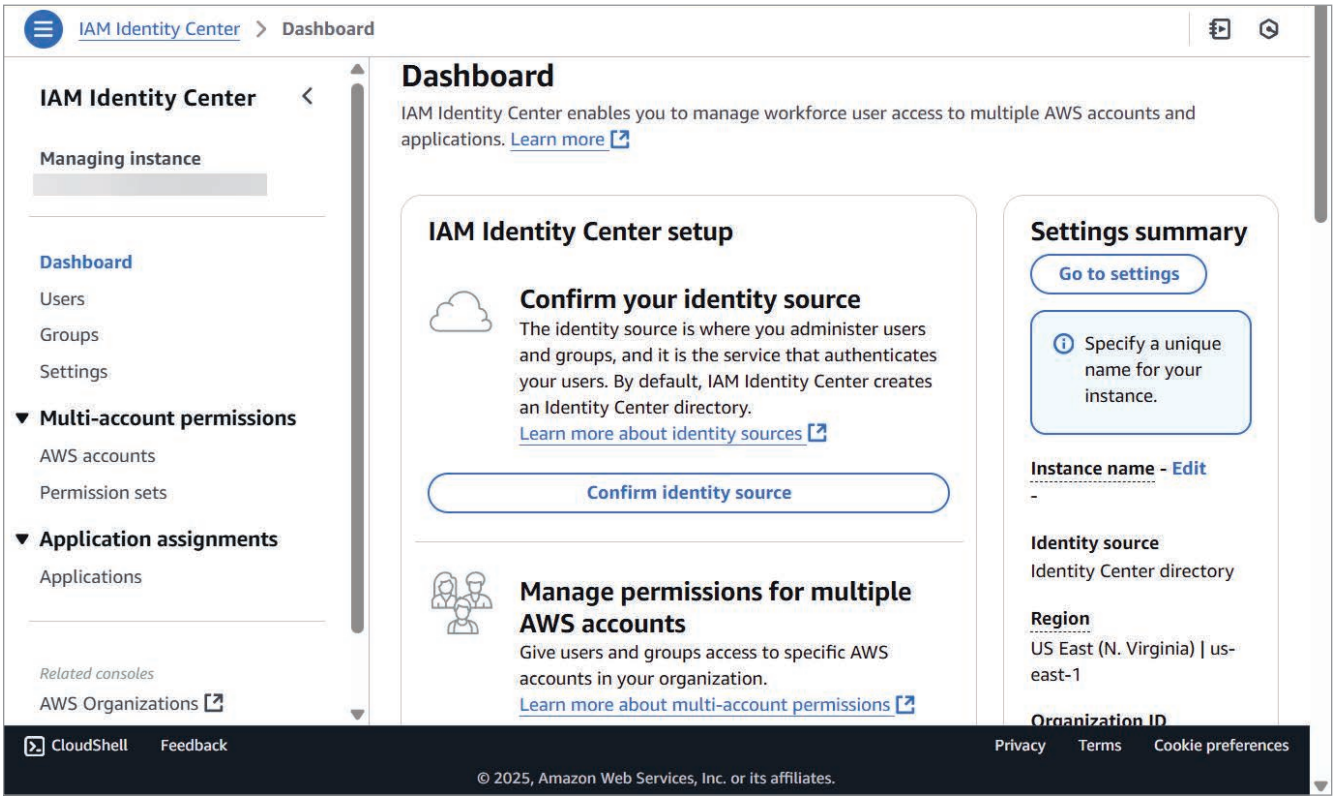
Workforce Identities

Similar to consumer users, an organization likely will not want to create administrative, technical-oriented accounts for all (or even most) of the employees in the company. These users do not need access to the cloud management environment. Instead, most employees will only need to interact with company applications as nontechnical end users. Although internal to the organization, these accounts require a different approach than providing administrative accounts to these users. Instead, you can create accounts that are called **workforce identities**, workforce users, or internal users, and they are designed for employees or contractors.

In Azure or GCP, you might still use Microsoft accounts or Google accounts for these users, as you can also administrate other organizational resources to your managed users, such as Microsoft 365 licenses or Google Workspace features. In AWS, you might use Cognito for access to individual apps, or you can use a more encompassing service called IAM Identity Center (Figure 7-4). IAM Identity Center also provides the ability to manage internal users' access to other AWS-managed applications. Examples include Amazon Q Developer—an AI-powered assistant (Figure 7-5) specially designed to aid in development processes—and Amazon QuickSight—a BI (business intelligence) service to access dashboards and visual analytics displaying company data using ML-driven (machine learning) insights.

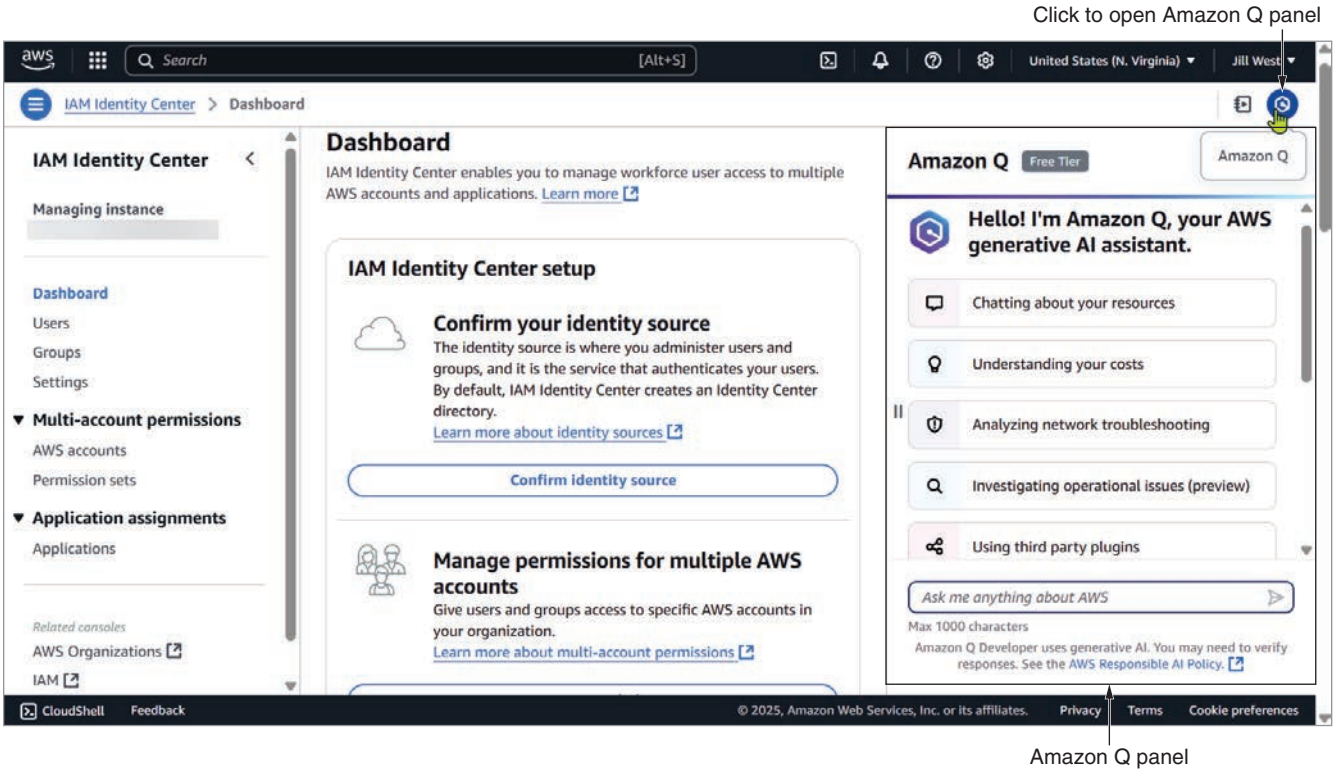
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-4 IAM Identity Center dashboard in AWS



Source: Amazon Web Services, Inc.

Figure 7-5 Amazon Q AI-powered assistant



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

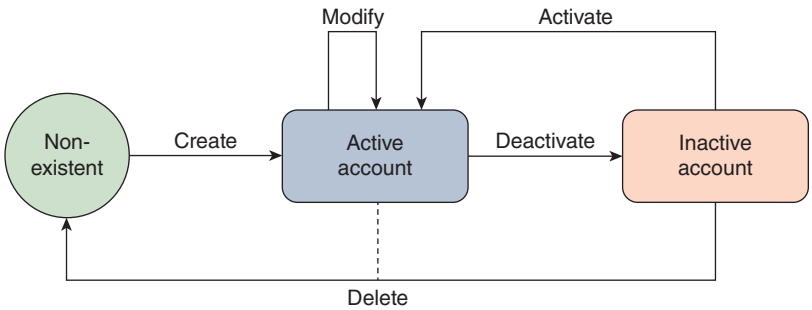
Account Management

Account management plays a critical role in a successful IAM deployment. People come and go in an organization, and the existing user accounts at any given time should accurately reflect the people who can legitimately access digital resources. For example, when an employee leaves an organization for another job, that person's user account(s) should be disabled as a matter of course. Effective account management policies can help ensure that the following tasks occur reliably and consistently:

- Accounts are set up, or provisioned, in a timely manner for new users.
- Compromised accounts are locked out for protection.
- Privilege creep, the gradual increase of disorganized and unmonitored privileges, is limited.
- Unused accounts are closed to further activity, a process called deprovisioning.

Figure 7-6 illustrates the life cycle of user accounts from creation through deletion. An active account might be modified or deactivated and sometimes deleted. Typically, it's better to deactivate an account when an employee leaves the company in case there are resources that only this account can access. This way, an admin could still access the protected resources, if needed.

Figure 7-6 Account life cycle



Another consideration in managing accounts is whether and when users should be given multiple accounts. For example, a user who needs administrative privileges to create new accounts might not need access to those privileges for day-to-day use. Giving that user a different account for each level or type of privilege they need can help minimize inherent vulnerabilities from those privileges. Other reasons to generate multiple accounts per user include the need to limit visibility of workloads between groups or teams, to limit the so-called blast radius of a critical event such as a security breach, and to comply with strict regulations on data isolation and auditing activities.

PAM (Privileged Access Management)

PAM (privileged access management) is a subset of IAM that applies stricter rules and safety precautions specifically to users who are given elevated permissions to do their jobs. These users might have backend access to databases or other critical systems containing highly sensitive information. They can perform more sensitive tasks, such as viewing or changing financial information, making configuration changes, or adjusting access privileges for other users. These users are also more likely to be targeted by cyber criminals.

Special security precautions for these accounts include the following:

- **Limited use**—These accounts should only be used when higher privileges are necessary to accomplish a task. Employees who are given a privileged user account should also be given a lower-level account for normal activities. In fact, anyone who has a user account of any kind should be given only the least privilege, or least amount of access, needed to do their specific job. This is called the **principle of least privilege**.
- **Limited location**—Many companies require that the privileged account is only accessed on premises so that no one, not even a legitimate network administrator, can access the system remotely and make high-level changes from outside the protected network.
- **Limited duration**—Privileged accounts should be carefully accounted for and disabled as soon as they're not needed, such as when an employee is terminated or transferred.
- **Limited access**—The passwords for these accounts should be kept especially secure and designed to be difficult to crack. The most important factor for password hacking resistance is length; privileged accounts should require much longer passwords than regular user accounts do. Passwords should be stored securely and, when possible, multifactor authentication should be required.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Limited privacy**—A privileged account can be used for destructive activity, whether malicious or not. For that reason, every user action in these accounts should be logged and monitored by someone other than the owner of that account.

PAM software is available from companies such as CyberArk, ManageEngine, BeyondTrust, and Delinea.

Self-Check

1. A user account is a(n) _____.
   a. root
   b. privileged account
   c. vault
   d. identity

2. What is a goal of successful account management?
   a. Limit users to one account each.
   b. Expand a blast radius.
   c. Limit privilege creep.
   d. Increase time for account provisioning.

Check your answers at the end of this chapter.

Section 7-2: Authentication

Certification

4.3 Given a scenario, implement identity and access management.

4.4 Given a scenario, apply security best practices.

Once a user or service has an identity, there must be a way to validate that account before allowing access to resources. This is where authentication comes in. Authentication is part of a three-tiered approach to NAC (network access control) called AAA (authentication, authorization, and accounting) and pronounced "triple-A." Put simply:

- Authentication gets you into a system.
- Authorization lets you do things while you're there.
- Accounting tracks what you're doing for later review.

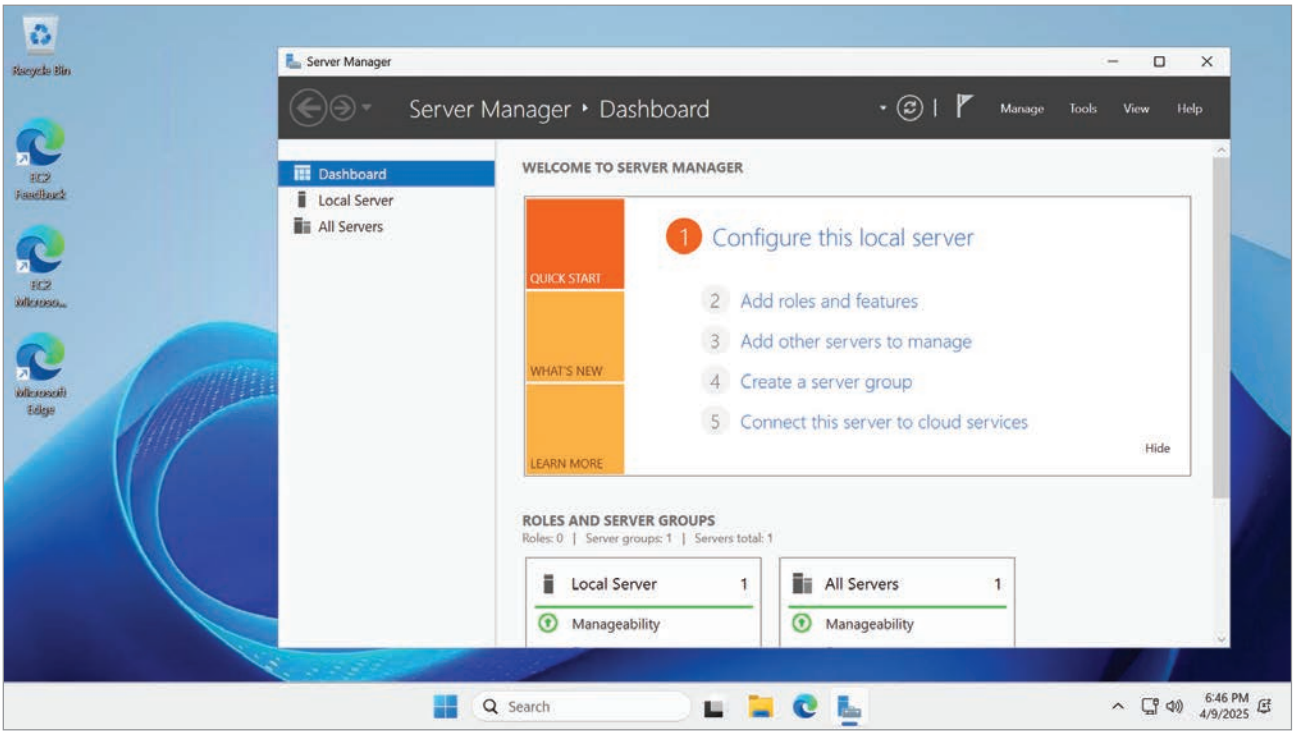
This section discusses authentication. You learn about authorization in the next section, and you explore accounting techniques (logging and monitoring) in a later chapter.

Authentication Processes

Authentication is a service that proves the identity of a client (which could be a user, service, application, or device) and determines whether that client is allowed to access a secured system. Essentially, it answers the question, "Who are you?" To confirm a client's identity, some sort of directory service must maintain a database of account information such as usernames, passwords, and any other authentication credentials. Often this is accomplished in **AD (Active Directory)**, a Microsoft product that runs in Windows Server (Figure 7-7). AD relies on the older NTLM (NT LAN Manager) authentication protocol, the newer and open-source Kerberos authentication protocol, or something more Linux focused such as OpenLDAP or 389 Directory Server.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-7 Active Directory in Windows Server



Microsoft Corporation

All these options are built to be LDAP compliant. LDAP (Lightweight Directory Access Protocol) is a standard protocol for accessing an existing directory. The mechanisms of LDAP dictate some basic requirements for any directory it accesses. Therefore, directory servers are configured and function in similar ways, regardless of the software used. LDAP can query the database, which means to draw information out of the database. It can also be used to add new information or edit existing data.

API-Based Communication

To support communication between the various systems involved in authentication processes in the cloud, these systems also often rely on REST APIs. Let's break this down. An API (application programming interface), as you've already learned, is a request made in a specific format to an application or platform. It's a standardized means of communicating data between applications and services. For example, you might have an app that needs 360-degree panoramas of specific locations on a map. You can program the app to collect map data from Google's Street View Static API, knowing the data will be available as pure data that your app can easily incorporate into its algorithms.

REST (representational state transfer), as you've also learned, is a kind of architecture standard that requires certain characteristics for HTTP (Hypertext Transfer Protocol) or HTTPS (HTTP Secure) communications over the web. A REST-compliant system will conform to the following characteristics:

- Client and server systems run independently of each other. You can replace one or change the code on one without being concerned about how it will affect the other. This creates a modular separation of duties that allows for each to evolve separately.
- The server saves no client data. Session information is stored only on the client system. This means that either side can respond to a single message without having read any earlier messages.

Now you can put API and REST concepts together in the context of authentication: REST APIs can be used to perform basic, standardized functions across the web to interact with an authentication directory. This approach allows organizations to set up relationships between various services or apps and a centralized authentication service, knowing the communication standards will allow these different systems to reliably communicate with each other.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

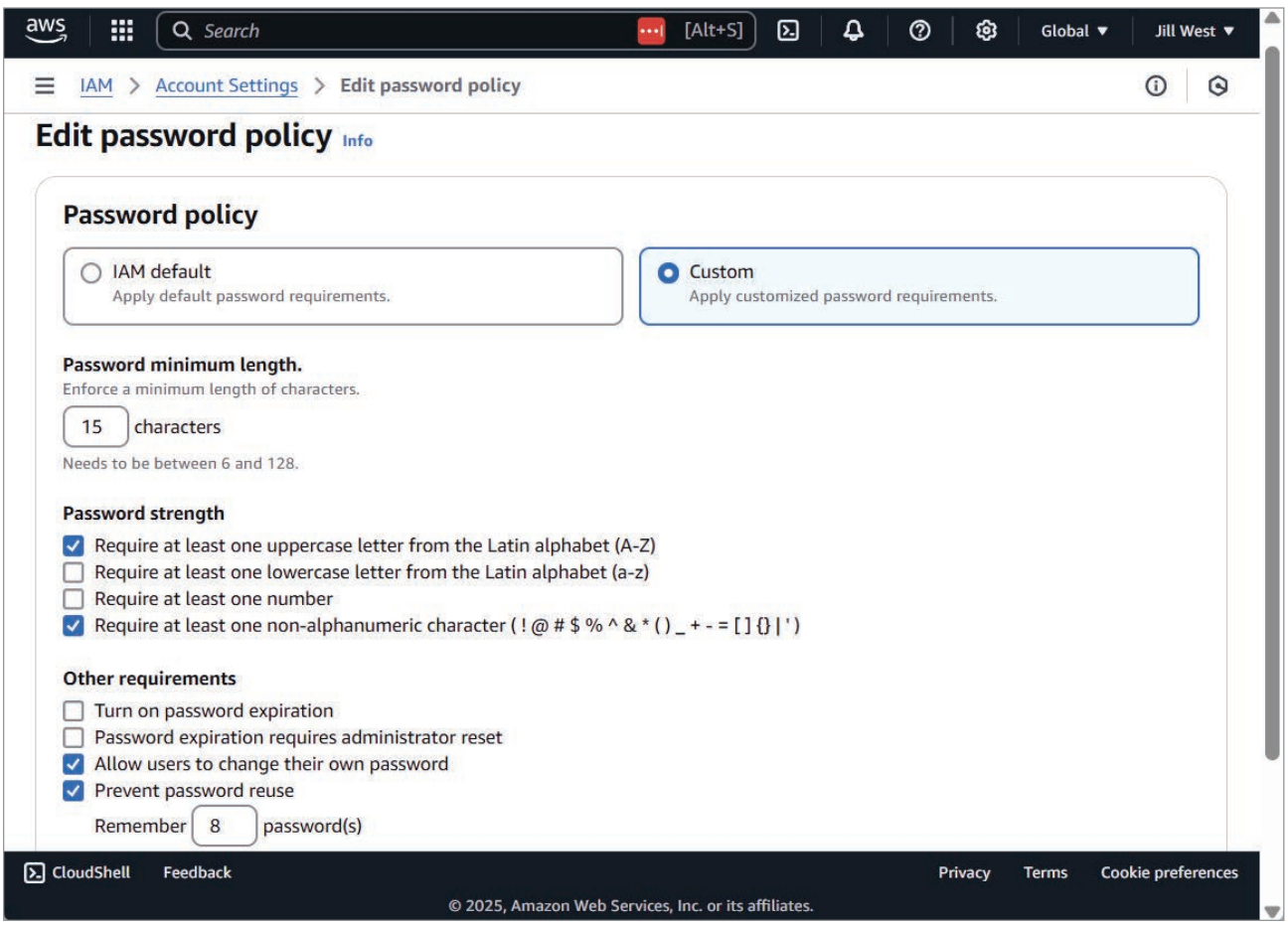
Password Policies

One way to increase security of user authentication is through a well-designed password policy. Password policies can require that passwords meet certain requirements, such as the following:

- **Complexity**—Password complexity can be increased by requiring a certain combination of character types, such as uppercase and lowercase letters, numbers, or non-alphanumeric characters.
- **Length**—According to many cybersecurity experts, password length is the factor most predictive of a password's resistance to brute force attacks. Statistically, a long but simple password trumps a short but complex password almost every time.
- **Expiration**—Requiring users to change their passwords somewhat frequently can, in some cases, increase password resistance to hacking. However, users tend to resort to insecure password practices when required to change their password often. Therefore, many cybersecurity experts and organizations—including NIST (National Institute of Standards and Technology)—only recommend requiring password replacement if there is some indication a password has been compromised. Often, you can also configure the policy to remember previous passwords and prevent users from reusing an earlier password.
- **Lockout**—An account lockout policy locks down an account after a specified number of failed attempts to sign in. A low lockout threshold (such as five attempts) reduces the risk of an attacker succeeding in guessing the right password within the allowed trials. Some systems will also generate an alert for IT staff even after one or two failed login attempts.

Figure 7-8 shows the password policy options for AWS member accounts.

Figure 7-8 Set a password policy in AWS



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 7-2 | Another way to increase password security and also increase convenience is to use a password manager. A password manager such as BitWarden, 1Password, Dashlane, NordPass, LastPass, or KeePass securely stores passwords and other credentials in a **password vault**. The user enters a long, complex master password to sign into the password vault, which then auto-fills passwords in websites and can even generate highly complex and very long passwords. Password information for other resources can also be stored, such as application passwords or server credentials. Relying on a password manager can help users create longer, more complex passwords and avoid reusing passwords, both of which increase password security. Although no password manager can guarantee complete protection, using additional security methods such as multifactor authentication (described later in this section) can add a layer of assurance.

Applying Concepts 7-1: Start Using a Password Manager

Estimated time: 20 minutes
Objective 4.4: Given a scenario, apply security best practices.
Resource:

- Internet access

Description: Security experts recommend that you use long, complex passwords and use a different password for every account. A password manager is a free and secure way to store passwords so it's easier to observe these security recommendations. All you need to remember is a single, complex master password.

In this activity, you create an account for a password manager. Always keep your master password secure and never share it with anyone. Complete the following steps:

1. Research online to compare some of the top-rated password managers. For example, you might compare features, reputation, and security elements for BitWarden, 1Password, Dashlane, NordPass, LastPass, or KeePass. Select one of these password managers for this activity—make sure it has a free tier subscription option (not just a free trial). What password manager did you choose?
2. Create an account with your chosen password manager. You'll need to create a master password—the longer, the better. Make sure to document your master password in a safe place.
3. If there is a site tour, use that resource to help familiarize yourself with your password manager. Otherwise, poke around on the site to discover the various tools, features, and settings. What feature do you think will be most useful for you? Why do you think this is?
4. Search for any account recovery options or emergency access options and configure those. Also add MFA requirements, if available.
5. If the password manager offers the option to install an app on your phone, consider using this app as a more convenient way to access your passwords when you're on the go. If the password manager offers the option to install a browser extension, consider doing this as well for convenient access to your vault when you're on your laptop or desktop computer. However, make sure to always sign out of the mobile app or browser extension when you're not using it.
6. Practice adding a password to your vault. For which website did you store a password? (Note: Do not share your password, even with your instructor!)
7. When you're finished, sign out of your password manager account. Always remember to sign out of your account before moving away from your computer or other device. Store a copy of your master password in a very secure place, such as a lockbox in your home, a safe deposit box at a bank, or a securely encrypted file on your computer or in the cloud.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Caution | No password manager is 100 percent reliably secure. Hackers target these services, and occasionally they're successful. No password manager is secure if you leave your account open on a computer or device you're not using or if you write your master password where someone else can find it. Follow these guidelines consistently:

- Never save passwords in your browser's own vault, which does not incorporate the same security standards offered by dedicated password managers.
- Always sign out of your password manager account or any other secure account when you're not using it.
- Always close browser windows where you have been signed into a secure account of any kind, even though you have signed out of that account.
- Always lock or sign out of the OS before leaving your computer or other device.

MFA (Multifactor Authentication)

A simple form of authentication is providing a username and password. Most people are familiar with this process, and most people (especially IT professionals) are aware that this kind of authentication is readily cracked if the password is compromised. A major weakness of passwords is their susceptibility to social engineering attacks. Social engineering doesn't require a great deal of technical skill at all. In fact, acting clueless is often the best way to convince people to give out their information. Search on YouTube for examples of social engineering attacks and notice the techniques attackers use to successfully gain the trust of the person they're talking with.

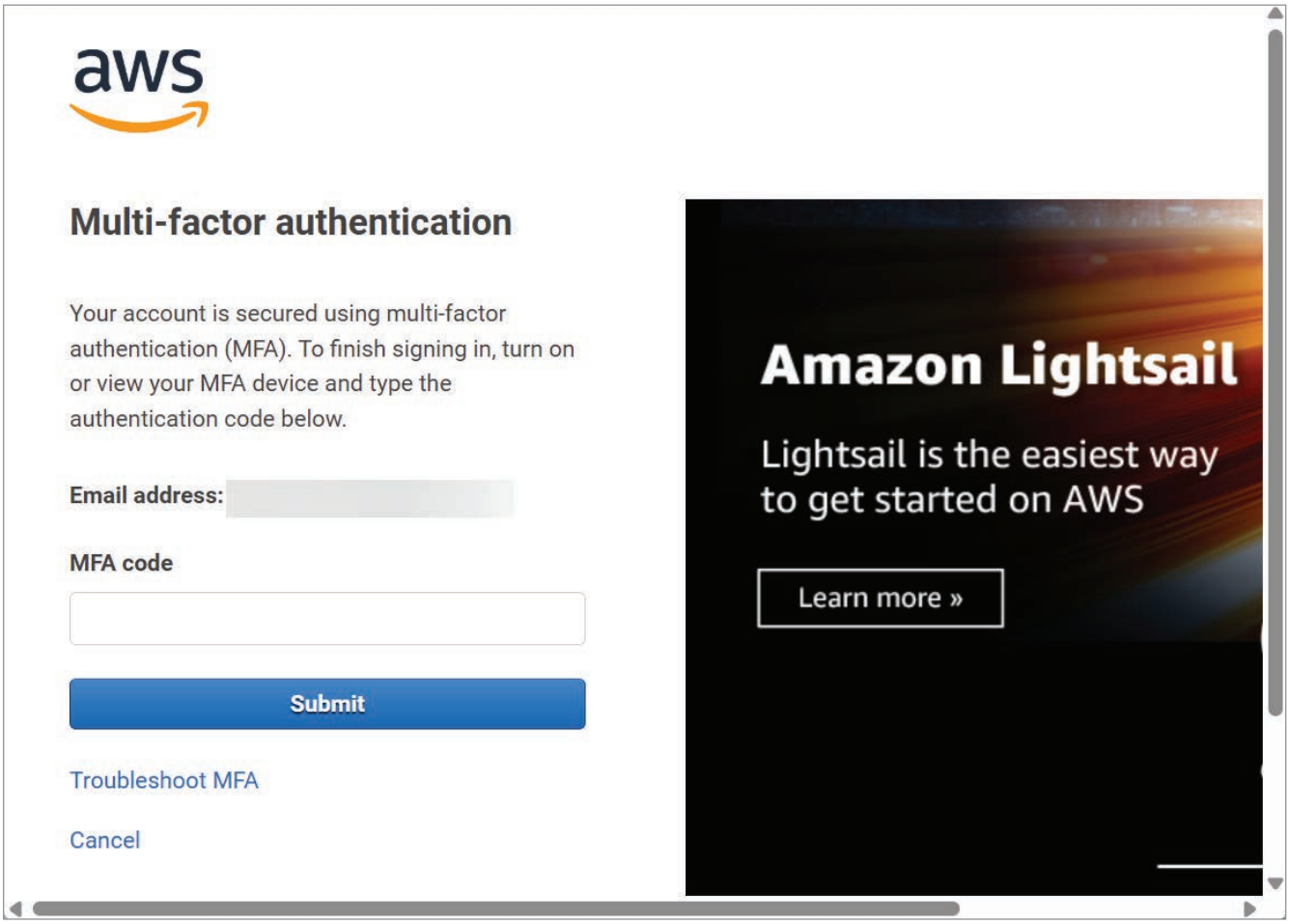
One way to increase authentication security is by requiring **MFA (multifactor authentication)**, which requires two or more pieces of information—called factors—from two or more categories of authentication factors. The following is a list of the most common MFA factor categories with a few examples of each:

- **Something you know**—A password, PIN (personal identification number), or biographical data (such as first elementary school)
- **Something you have**—An ATM card, ID badge, key, or smartphone with an authentication app
- **Something you are**—Your fingerprint, facial pattern, or iris pattern
- **Somewhere you are**—Your location in a specific geopolitical area (such as a country), a company's building, or a secured data closet
- **Something you do**—The specific way you type, speak, or walk

MFA requires at least one factor from two or more *different* categories. For example, signing into a network might require a password (something you know), a fingerprint scan (something you are), plus a code generated by an app on your smartphone (something you must have with you to find the code). Figure 7-9 shows a request for an MFA code to sign into AWS. The MFA app generates a random six-digit code every 30 seconds that must be entered on the secured site in addition to the user's password to access the account.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-9 MFA on AWS



Source: Amazon Web Services, Inc.

AI in IT | You've probably noticed tools on websites that attempt to distinguish between human visitors and bots, especially during an authentication process. For example, a variety of CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) tools prompt the user to identify photos containing a type of item, such as traffic lights.

However, with AI, it's becoming increasingly difficult to distinguish between humans and bots. In September 2024, researchers at ETH Zürich in Switzerland successfully trained an AI to pass Google's image-based reCAPTCHAv2 tests every time. Additionally, in March 2025, researchers at UC San Diego found that the GPT-4.5 LLM (large language model) was determined by users to seem more human in its interactions than actual humans do.

References:

Jones, C. & Bergen, B. (2025). Large Language Models Pass the Turing Test. Preprint. https://doi.org/10.48550/arXiv.2503.23674.

Plesner, A., Vontobel, T., & Wattenhofer, R. (2024). Breaking recaptchav2. *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1047-1056. https://doi.org/10.1109/compsac61105.2024.00142.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Applying Concepts 7-2: Require MFA on Your Cloud Account and Other Sensitive Accounts

Estimated time: 15 minutes
Objective 4.3: Given a scenario, implement identity and access management.
Resources:

- A standard public cloud account (such as AWS, Azure, or GCP)
  Note that an AWS Academy or other educational cloud account likely will not be sufficient for this activity. You can instead choose a different type of account, such as an email, social media, or financial account.
- Internet access

Description: If you've not yet done so, this is a good time to set up MFA in your cloud accounts and any other sensitive accounts (such as email). Complete the following steps:

1. Sign in to your account.
2. Locate your account settings, specifically the option to change your password.
3. If available, enable MFA and set up the secondary authentication factor. For example, you might need to sync your account with an authentication app on your phone.
4. In your password manager, make a note on the affected account's entry that identifies the MFA method you chose, such as the authentication app.
5. While you're here, check your password to ensure it is long enough. Security experts recommend at least 12–15 characters for standard accounts and longer passwords for more sensitive accounts.
6. Repeat these steps for any other sensitive accounts you want to protect, such as email, banking, or social media accounts. Which accounts did you secure with MFA?

Certificate-Based Authentication

One example of a "something you have" authentication factor can be a digital certificate. You've already learned how a key pair can be used to support secure encryption services. Public key encryption can also be used to prove the identity of a person, business, or server. This is accomplished through digital certificates.

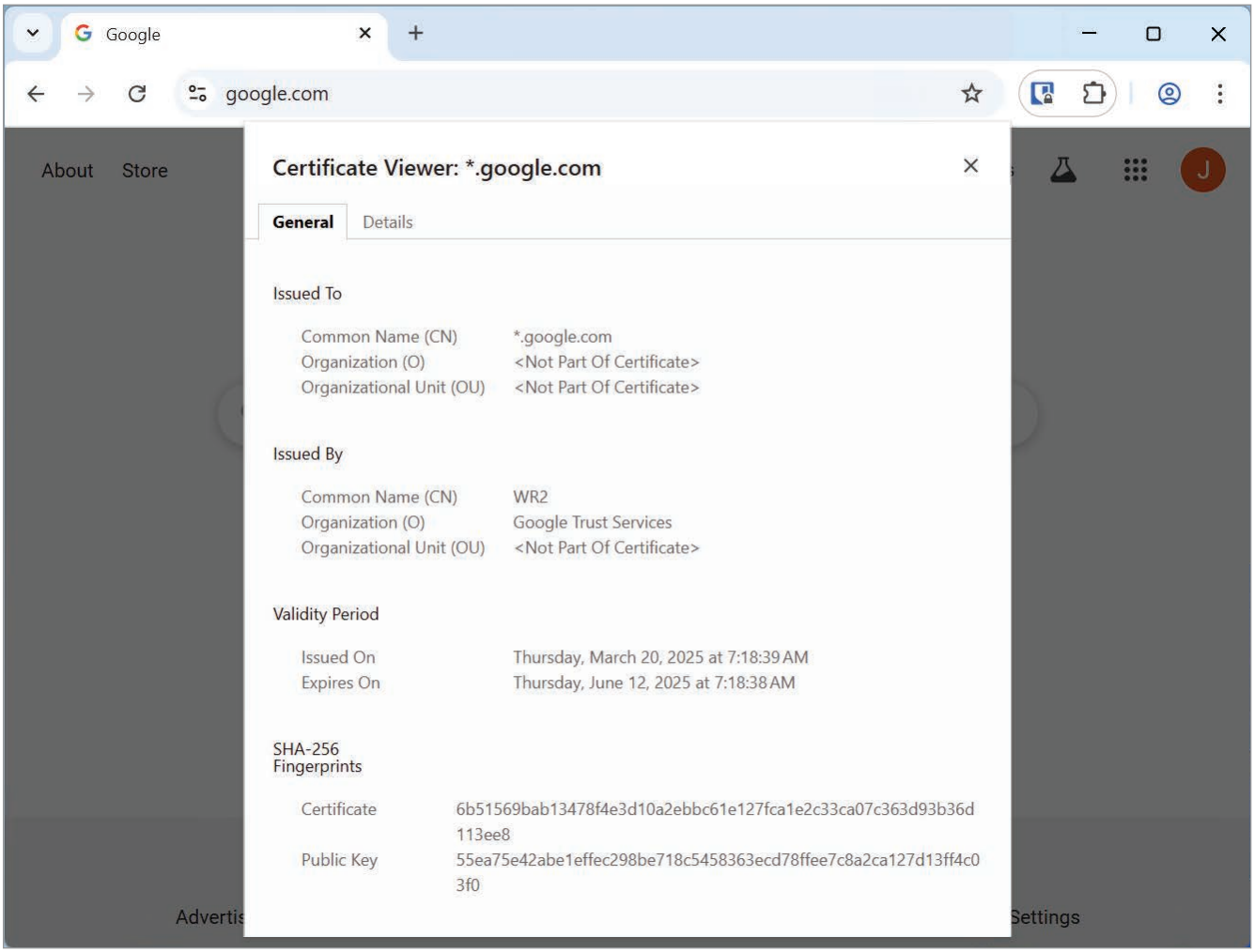
A **digital certificate** is a small file containing verified identification information and the public key of the entity whose identity is being authenticated. If the public key can successfully decrypt data contained within the certificate, the entity has proven possession of the related public and private keys. Other information contained within the digital certificate provides additional identity information, which is validated by a third party called a CA (certificate authority). CAs issue, maintain, and validate digital certificates. CAs also maintain a CRL (certificate revocation list) to identify which certificates have been withdrawn prior to their expiration date and should not be trusted. The use of CAs to associate public keys with certain users is known as **PKI (public key infrastructure)**.

Digital certificates are primarily used to certify and secure websites where financial and other sensitive information is exchanged. They're also used for other types of websites (Figure 7-10) and to secure email communications, to authenticate client devices in a domain, or to authenticate users to a network. When surfing the web, at some point you might have gotten an error message that said the website's certificate was untrusted. This means the website's digital certificate used by the encryption protocol SSL (Secure Sockets Layer) or TLS (Transport Layer Security) was not signed by a trusted CA or associated with a trusted root certificate.

Establishing and maintaining an organization's PKI can be expensive and complicated. Doing this on-prem requires significant infrastructure investment, and these PKI services are limited to authentication of on-prem resources. By moving to a cloud-based PKI solution, companies can reduce overall expenditure while simplifying the work required to support PKI services and extending the reach of those services to cover both on-prem and cloud-based resources.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-10 Certificate for Google.com



Source: Google LLC

Certificate Management

Effective certificate management includes monitoring certificates for compromise, revocation, or expiration and renewing certificates in a timely manner. The following list describes various certificate statuses:

- **Good**—A certificate is good, or trusted, during the timeframe when it was approved (that is, it has not yet reached its expiration date), and there is no indication it has been compromised.
- **Retired**—Once a certificate reaches its expiration date, it is retired and can no longer be used.
- **Revoked**—If there is any indication the certificate has been compromised or destroyed, or if the certificate is no longer needed (such as when an employee leaves a company), the certificate must be revoked.

To revoke a certificate, the CA that issued the certificate requires submission of a revocation request, and sometimes supporting evidence will need to be included. Once revoked, the certificate is added to a CRL, which other entities can check to confirm whether a certificate has become invalidated. However, these lists are not consistently checked because they are long, flat files that take a while to download and search. Therefore, a revoked certificate might still sometimes be used successfully to perform unauthorized access or to impersonate someone else. Alternatively, a revoked certificate might be identified with OCSP (Online Certificate Status Protocol), which offers an updated and scalable method of tracking certificate revocation. The OCSP server collects CRLs from multiple CAs and can respond to requests for a certificate's status.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A compromised certificate is not as useful to an attacker without the corresponding private key. For this reason, private key security is especially critical. Even if there is no evidence that a private key has been compromised, simply discovering the key has been temporarily exposed presents sufficient risk that the key should be considered compromised and should be replaced.

Note 7-3 | The simplest authentication model is **local user authentication**, which performs the authentication process on a single machine to access that machine's own resources. Local user accounts can be a good backup mechanism in case your IAM system goes down or otherwise isn't accessible. However, local authentication typically is not as secure as other authentication options, and it presents an administrative nightmare when handling a large number of users and resources.

Programmatic Access

Throughout the course, you've used the GUI (graphical user interface) console, or web portal, to perform many tasks in the cloud. However, there are other ways to interact with your cloud resources, which you learn about in this section. Later in this course, you'll learn how to use a CLI (command line interface) tool to work in the cloud.

Exam Tip | In the context of securing access to the cloud management environment, the CompTIA Cloud+ objectives list CLI (Common Language Infrastructure), which is a different term than the CLI (command line interface) typically associated with cloud IAM. CLI (Common Language Infrastructure) was originally developed by Microsoft to provide an intermediate language that allows various coding languages to work on different hardware. Use context, such as information provided in an exam question, to determine which acronym is intended.

Generating Access Keys

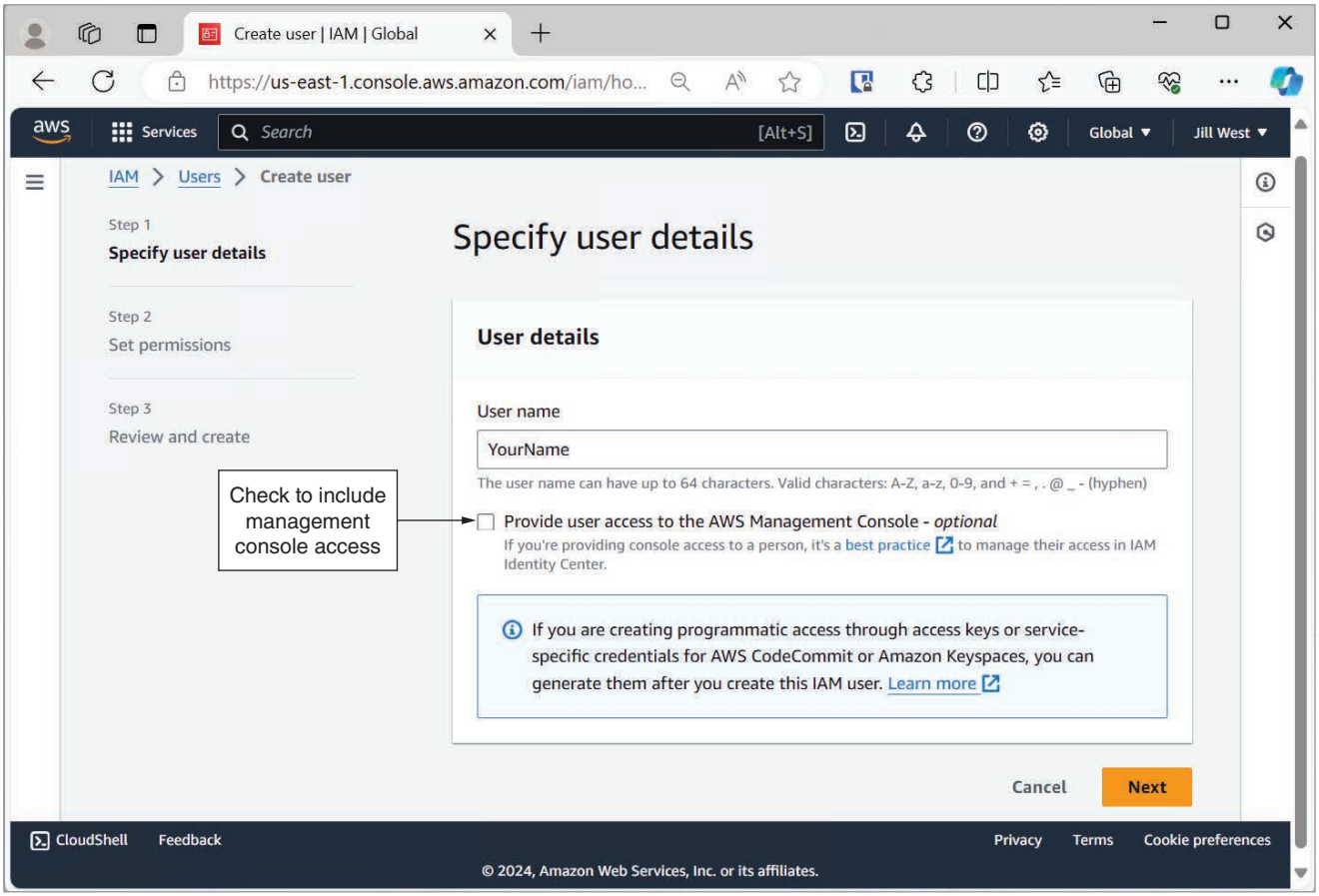
Programmatic access refers to the ability to interact with cloud resources through a command line, SDK, API, or a third-party tool. An **SDK (software development kit)** is a platform-specific collection of tools and resources developers need to create, deploy, and troubleshoot their applications. For authentication, programmatic access requires access keys instead of a username and password. For example, when you create a user account in AWS's IAM service, you can choose programmatic access and/or management console access for that user (Figure 7-11).

Management console access relies on a username and password to authenticate. However, if you choose programmatic access, the user is assigned an access key ID and a secret access key. Both values are required for the user to interact with their AWS account through a CLI or another tool. You only get one opportunity to download a new account's secret access key, which is immediately after creating it (Figure 7-12).

The user could be a human user signing in through a CLI, such as Windows PowerShell or Linux Terminal, or accessing resources through an SDK. The "user" could also be a computer system of some type, such as an API from another system. For example, a monitoring tool running in your on-prem data center will need programmatic access to your AWS infrastructure.

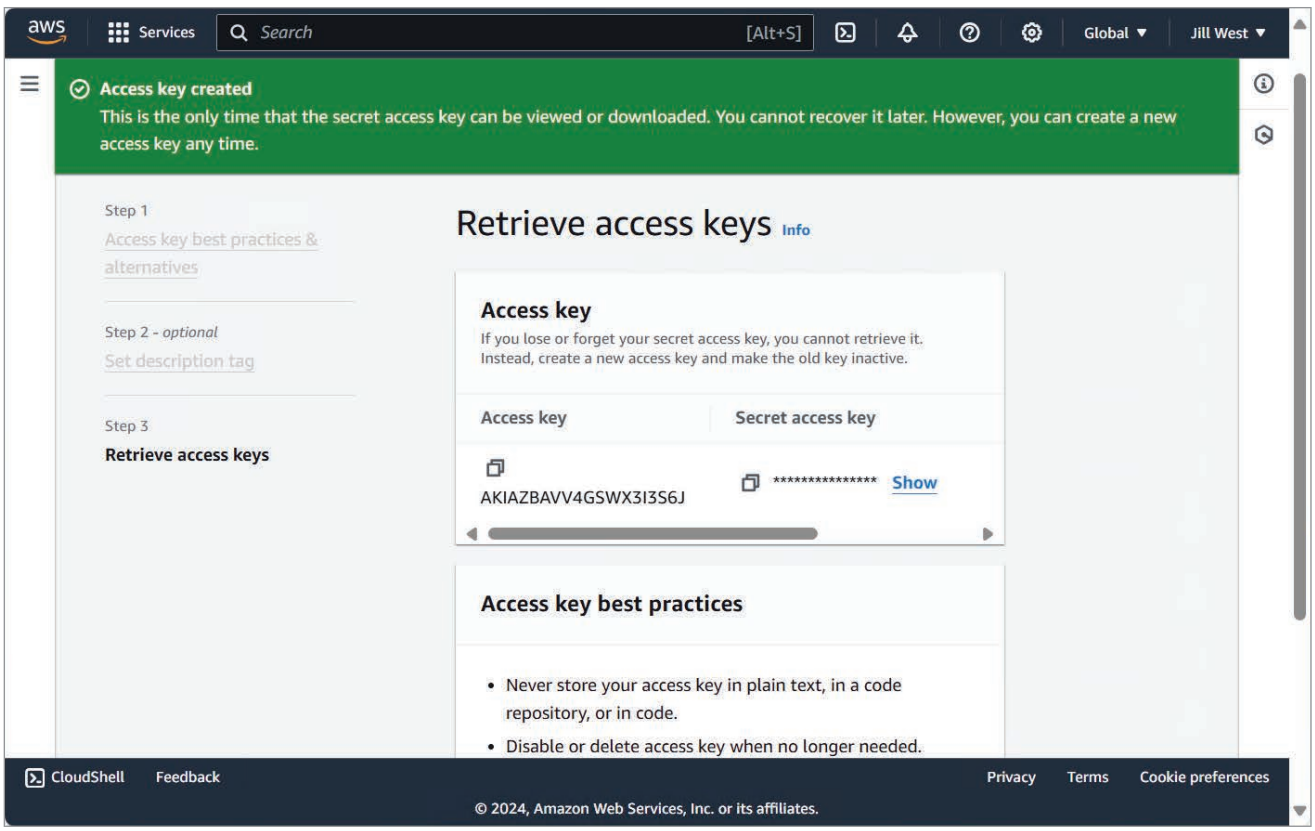
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-11 Grant an IAM user programmatic access or management console access



Source: Amazon Web Services, Inc.

Figure 7-12 The access key can only be downloaded immediately after it is created



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 7-4 | Recall that systems can communicate with each other via APIs. This process does not always require access keys; however, authentication should be applied to any public-facing APIs whenever possible such as when allowing a user programmatic access to an AWS account. In AWS, API authentication to your own APIs can be managed through API Gateway. Among other things, API Gateway provides an endpoint target where API requests can be sent and then routed to the appropriate backend API in your cloud deployment. It also incorporates several security features to protect your APIs, including authentication and authorization for APIs throughout your architecture. AWS offers many options to implement API Gateway authentication, including IAM roles, resource-based policies (that is, a policy created on the API resource's configuration page that defines "allow" or "deny" access rules), or Cognito user pools. Cognito is an AWS service that offers customizable authentication and authorization features to secure REST API access.

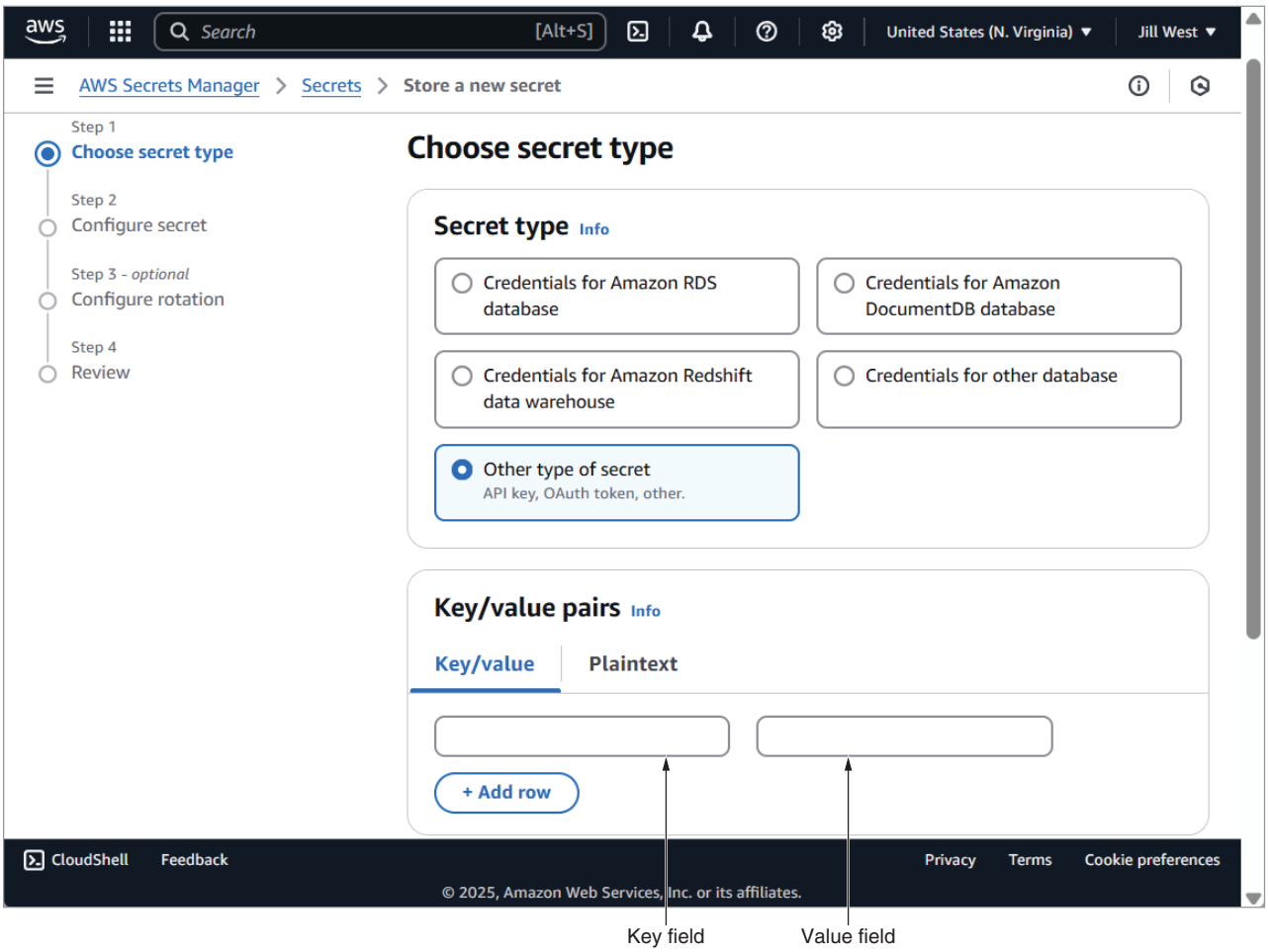
Securing Secrets

Information such as passwords, private keys, and certificates must be carefully managed to ensure they are available when legitimately needed but not vulnerable to discovery by attackers or unauthorized users. Let's explore some of the technologies used in safely storing and using these sensitive files.

Secrets Management

Secrets management systems handle the security and distribution of keys, passwords, authorization tokens, and other files used to secure access to resources. Many CSPs include platform-native tools for handling these kinds of secrets. For example, AWS offers Secrets Manager, which supports the secure storage, retrieval, rotation, and retirement of secrets such as database credentials and API keys. Figure 7-13 shows some of the options for storing a new secret in Secrets Manager.

Figure 7-13 Store a secret as a key-value pair



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Securely storing secrets in password vaults and similar services avoids the need to hard-code this sensitive information into applications, configuration management scripts, and other cloud resources. For example, an application might expose an API endpoint that other applications can use to allow both applications to interact with each other. This connection requires some kind of authentication, such as providing an authorization token or a private key. Instead of storing these secrets in the code of the application requesting authentication, they are encrypted in storage or generated only when needed. An application, script, or other resource can request access to the secret at the time of use, which is then securely transmitted to the authorized service or process. Access to these secrets is also logged for live monitoring and retroactive analysis, as needed.

Self-Check

3. Which protocol defines how most authentication directories work?
   - **a.** LDAP
   - **b.** OCSP
   - **c.** HTTPS
   - **d.** MFA
4. If a certificate is included in a CRL, what is its status?
   - **a.** Retired
   - **b.** Revoked
   - **c.** Trusted
   - **d.** Renewed
5. Which access type is most appropriate for application-to-application communication?
   - **a.** SDK
   - **b.** API
   - **c.** CLI
   - **d.** GUI

Check your answers at the end of this chapter.

Section 7-3: Authorization

Certification

4.2 Compare and contrast aspects of compliance and regulation.

4.3 Given a scenario, implement identity and access management.

4.5 Given a scenario, apply security controls in the cloud.

While authentication gets you into a system, authorization determines what you can do while you're there. For example, if you sign into the AWS Management Console as the root user, you're then authorized to do pretty much anything with the resources, services, and processes contained in that account. However, if you're signed into a different user account, such as a web developer user or a network admin user, you're only authorized for limited actions. **Authorization**—a critical element of IAM—enforces these access control limitations.

Authorization Processes

Each cloud platform manages IAM permissions in different ways. For example, all three of the major public cloud platforms that you're learning about in this course (AWS, Azure, and GCP) control access to system resources—such as compute services, storage objects, or virtual networks—by managing requests to access

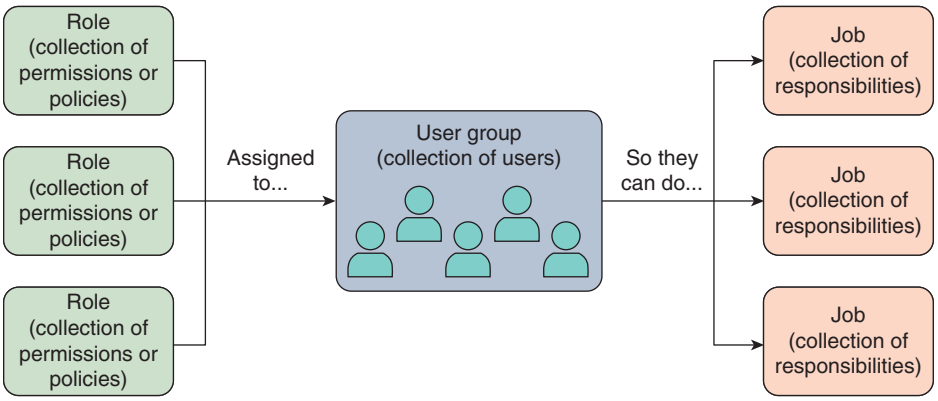
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

those resources. However, each platform uses different terminology and access control architecture. This section evaluates these variations. But first, let's begin with a quick description of how roles are used for access control.

Access Control Methods

Many cloud platforms rely on **RBAC (role-based access control)** to manage users' access to resources. With RBAC (also called nondiscretionary access control), a network administrator receives from a user's supervisor a detailed description of the roles or jobs the user performs for the organization. The admin is responsible for assigning exactly the privileges and permissions necessary for the user to perform these roles. To manage this, the admin creates a **role** object with the needed privileges and permissions for each type of job or other collection of tasks in the organization. The admin then associates the role directly with a user or with a user group. Users are assigned to groups that match the roles they need to do their jobs (Figure 7-14). RBAC often is additive, so each additional role given to a user increases that user's access to resources. In some cases, such as with AWS, a user must switch between roles in order to access the permissions associated with each role.

Figure 7-14 Roles allow users to do their jobs

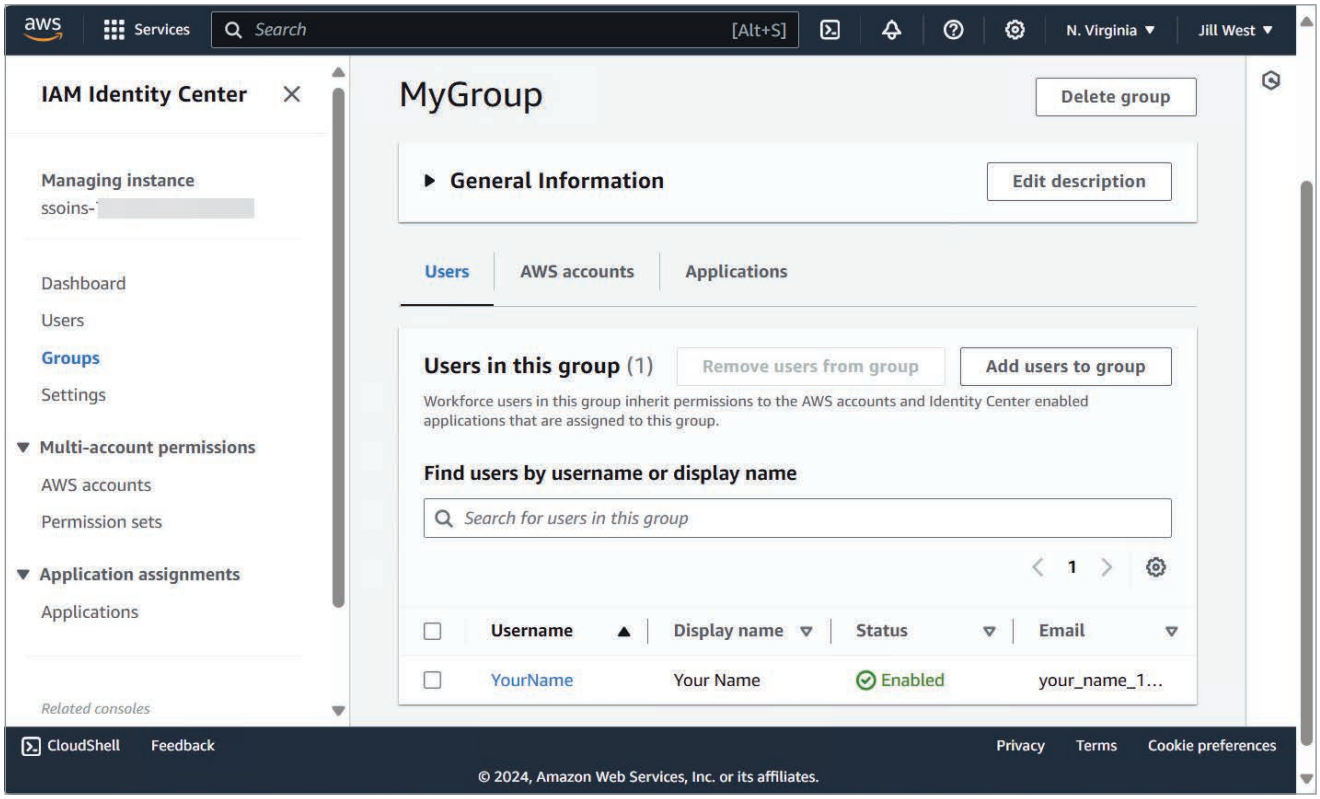


In contrast to RBAC, other popular access control methods you need to know about include the following:

- **DAC (discretionary access control)** is the least secure of these methods. With DAC, users decide for themselves who has access to their resources. You might have used DAC when setting share permissions on a Windows computer to determine who on your home network has access to your files.
- **MAC (mandatory access control)** is the most restrictive of these methods. In this case, resources are organized into hierarchical data classifications, such as "confidential" or "top secret." Resources are also grouped into categories—for example, by department. Users, then, are also classified and categorized. If a user's classification and category match those of a resource, then the user is given access to that resource.
- **GBAC (group-based access control)** is a method of access control where users are organized into groups. GBAC can be used separately or alongside role-based access control—where RBAC is used to organize permissions, GBAC is used to organize users. Figure 7-15 shows a group in AWS's IAM Identity Center service; the group contains one user but could also include AWS accounts or applications. Once assigned to a group, the user receives certain permissions that all members of the group share in common. In Azure, dynamic groups can automatically join users to groups based on rules, such as the user's location or job title.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-15 IAM Identity Center group with one user in the group



Source: Amazon Web Services, Inc.

AWS IAM

As you've already learned, AWS creates a user object to represent a human or another resource. The user itself is also an AWS resource and is limited to the account in which it was created. A user has a name; an ARN (Amazon Resource Number); and credentials such as a password, access keys (including an access key ID and a secret access key), or a server certificate.

A policy is a collection of permissions, such as the permissions listed in the AmazonEC2ReadOnlyAccess policy shown in Figure 7-16. It can be assigned directly to a user, but this is not recommended. Instead, best practice is to assign policies to users indirectly via roles. Using roles, you can add or take away permissions without having to manage each permission or user account separately.

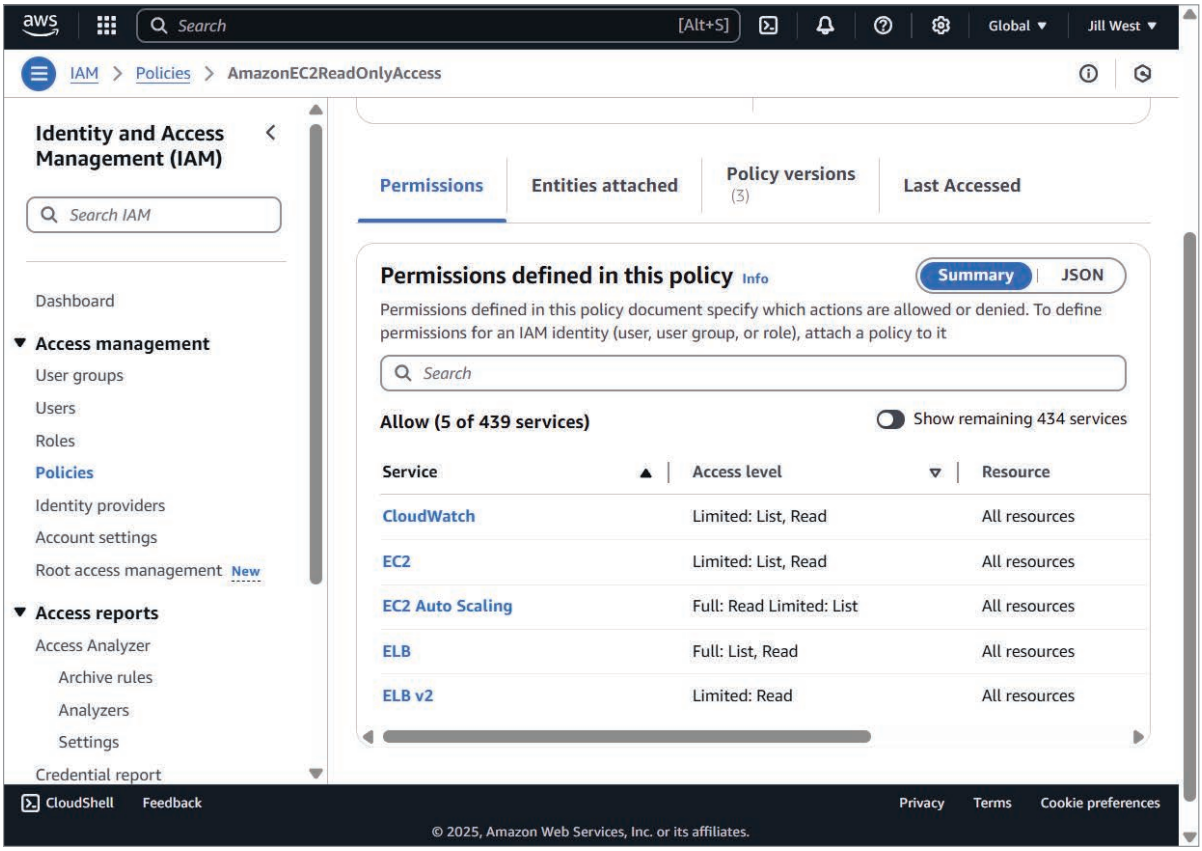
AWS offers many kinds of policies. An identity-based policy is a policy assigned to identities (usually via roles) to define the permissions granted to those identities, such as access to resources. In contrast, a resource-based policy is a list of rules applied to a resource, such as a storage bucket, that defines who can do what to the resource. This is often used to manage access to resources by external entities. For example, a storage bucket could contain the files for a static website, and a resource-based policy could allow public access to those files so visitors can explore the website. Only a few services offer this option to define a resource-based policy.

Another categorization of policies in AWS refers to who controls the policies. An AWS managed policy is a policy created by AWS that lists permissions often needed together for certain roles. If you need a customized policy, you can create a customer-managed policy instead. You can even start with an existing AWS managed policy, create a copy, and then edit it as desired.

IAM policies in AWS are written in JSON (JavaScript Object Notation) or YAML (YAML Ain't Markup Language). While you do not need to know how to write in JSON for the CompTIA Cloud+ exam, you should be able to interpret a basic JSON policy to determine what permissions it grants and to whom. Figure 7-17 shows an example of a policy being created in JSON. This policy will allow two actions in EC2: describe EC2 instances and describe VPCs.

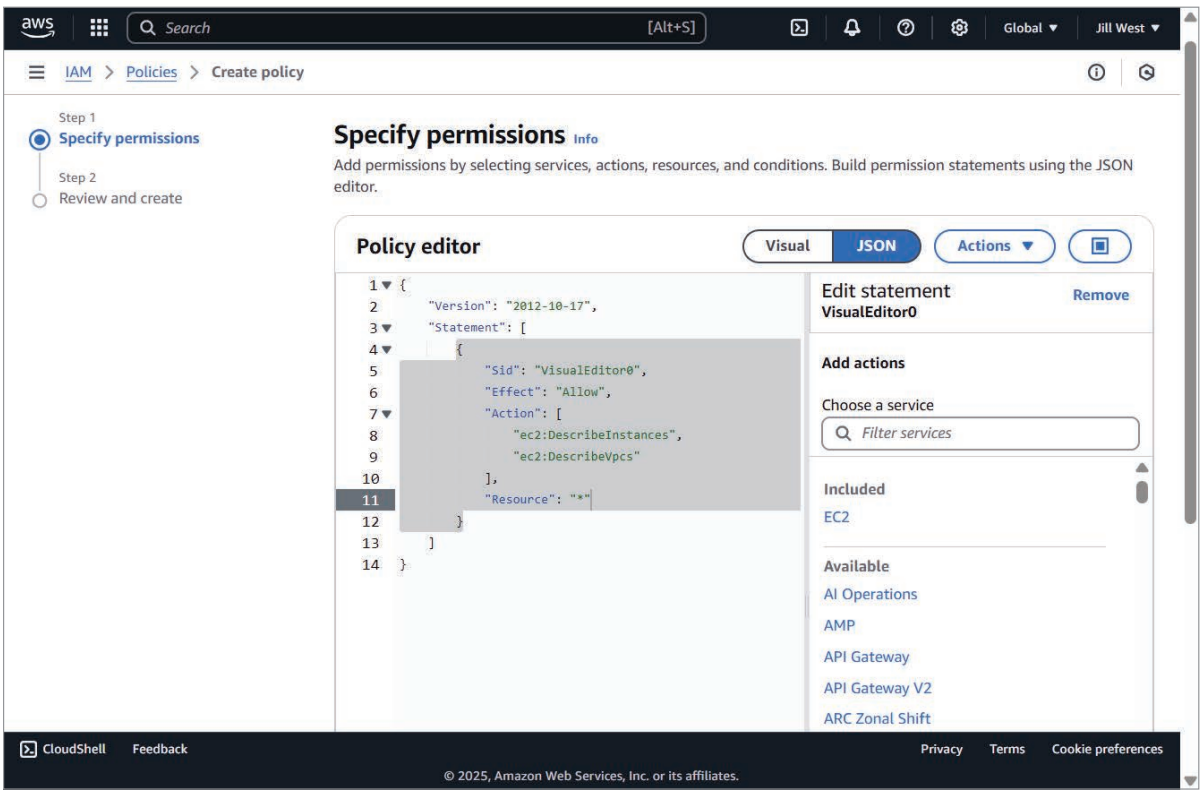
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-16 An AWS policy is a collection of permissions



Source: Amazon Web Services, Inc.

Figure 7-17 An IAM policy written in JSON



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A policy's effect defines either allow or deny permissions. An action defines what can be done to specified resources, such as the following examples:

- **List action**—Allows the user to view a list of the specified resources, such as EC2 (Elastic Compute Cloud) instances or S3 (Simple Storage Service) buckets.
- **Describe action**—Allows the user to view information about the resources, such as EC2 instances or VPCs (virtual private clouds).
- **Get action**—Allows the specified resource to be retrieved, such as opening a storage bucket to view its objects.
- **Put action**—Allows the user to create or edit the specified resource.

An asterisk (*) is a wildcard that refers to all actions or all resources. For example, ec2:* refers to all actions that can be performed on EC2 instances, and s3:* refers to all actions that can be performed on S3 buckets.

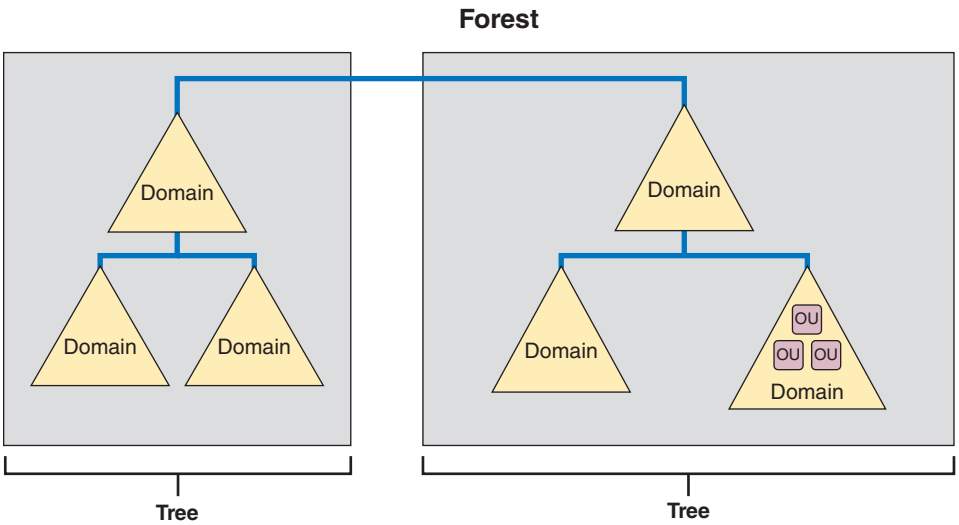
Azure IAM

Azure manages IAM services with Entra ID, which is the cloud-native cousin of Active Directory found in Windows Server for on-prem networks. Active Directory offers many services for a traditional network, including the following:

- **Domain Services (AD DS)**—Provides authentication and authorization support for users and devices on an organization's network.
- **Certificate Services (AD CS)**—Creates, distributes, and manages secure certificates.
- **Lightweight Directory Services (AD LDS)**—Supports applications that rely on LDAP.
- **Federation Services (AD FS)**—Supports SSO (single sign-on).
- **Rights Management Services (AD RMS)**—Provides IRM (information rights management) protection of data by enforcing data access policies.

One of the defining aspects of AD is its hierarchical structure (Figure 7-18). The all-encompassing container is the forest. Within a forest, you can have many domains branching out to multiple layers. A root domain along with its branch domains collectively is called a tree. Within each domain, you might have multiple OUs (organizational units). Inside the OUs is where you'll find leaf objects, such as users, computers, or printers.

Figure 7-18 A forest contains trees, which contain domains, which contain OUs

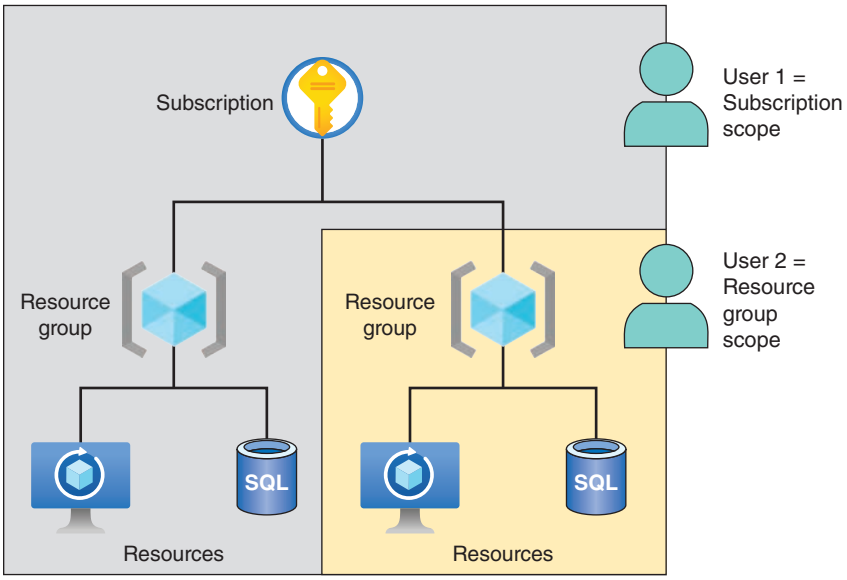


Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In contrast, Entra ID uses a flat-file structure. This means there is no container hierarchy like OUs or forests. Entra ID does use a domain, but this is more of a naming system and not designed to hold OUs or leaf objects. Entra ID relies instead on groups, such as resource groups or user groups.

Entra ID also applies permissions in a way that is naturally limited by a concept called a scope, which is the set of resources that a user *might* have access to (Figure 7-19). You can think of a scope as being a boundary defined at a certain level in an access hierarchy. For example, if a user has access at the subscription scope, that user can't access any resources in a different subscription but could be given permission to access resources in multiple resource groups within that subscription. If a user has access at the resource group scope, that user's access is limited to resources within the one resource group. By itself, a scope does not give a user access to resources. The scope merely defines the extent of access a user *can* receive from a role assignment.

Figure 7-19 A scope defines which resources a user might have access to



GCP IAM

GCP's IAM service revolves around principals, which are entities that represent human users, groups of users, domains, or workloads. The IAM service grants principals access to resources via roles, which are collections of permissions. Note that permissions cannot be granted directly to principals. Permissions can only be granted by assigning a role to the principal. And when a role is assigned to a principal, all permissions contained in that role are granted to the principal. For example, Figure 7-20 shows only a small portion of the 187 permissions contained in the pre-defined Compute Network User role.

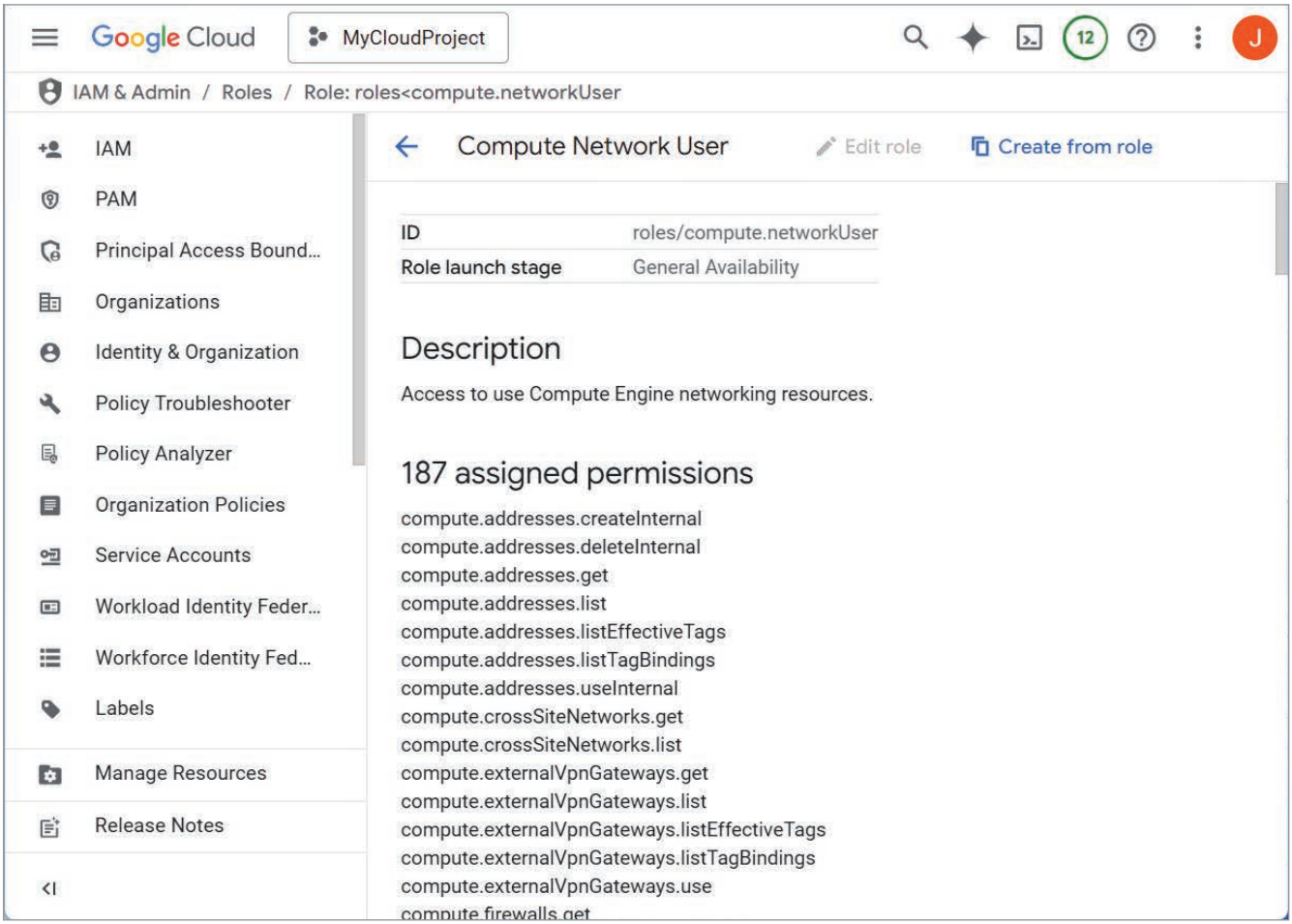
A permission defines an operation allowed on a resource, such as a VM instance or a storage bucket. You can grant permissions at the project level or, in some cases, on a specific resource. Permissions are defined in the form <service>.<resource>.<verb>, such as compute.instances.delete.

Similar to AWS, a GCP account can create an organization resource to expand functionality in ways that are often needed for a corporate GCP deployment. For example, an organization resource provides identity management across a company's users and workloads, lifecycle management to better monitor resources in disparate deployments, and recovery mechanisms for greater durability.

Consider what could happen to a company if an employee creates a system's resources using their own account and then they leave the company. The resources would be lost to the company. Instead, an organization ensures that resources deployed within it remain part of the organization even as employees come and go. The organization contains projects, and each project must be part of an organization, reducing the possibility for shadow projects that are not tracked or secured by the company. Additionally, permissions can be granted across projects through the organization resource.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-20 Assign a role to grant permissions



Source: Google LLC

Self-Check

6. The following policy grants what kind of permission?

```
"Sid": "FirstStatement",
"Effect": "Allow",
"Action": "s3:ListAllMyBuckets",
"Resource": "*"
```

a. Open
b. Read
c. Edit
d. Delete

7. What access control method does Azure use?

a. AD
b. RBAC
c. DAC
d. MAC

8. Groups are used to organize _____ .

a. roles
b. resources
c. permissions
d. users

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 7-4: IAM for Hybrid Clouds

Certification

4.3 Given a scenario, implement identity and access management.

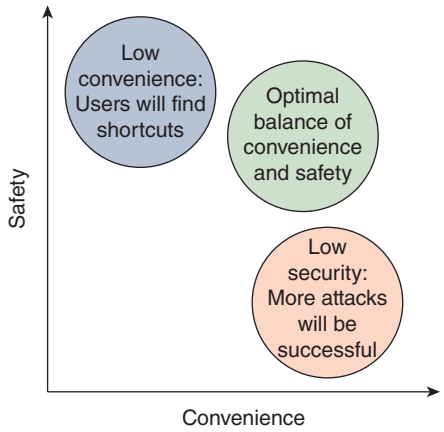
With so many organizations camping out somewhere in the middle of the spectrum between all-cloud and everything on-prem, managing identity and access across a hybrid cloud becomes essential. One way to do this is to adapt on-prem AAA methods to the cloud environment. For example, you might install Active Directory on a VM server in the cloud. As you've discovered throughout this course, however, much of the advantage of moving to the cloud is only realized when deploying cloud-native solutions. IAM is no different in this regard.

Consider each of the elements of identity management that you've been reading about in this chapter: identification, authentication, and authorization. Blending these services across cloud and on-prem deployments requires techniques that work with multiple systems, outside of the more limited IAM service of a single cloud. Let's explore some of these technologies, and then you'll learn about some of the available services that support hybrid identification, authentication, and authorization in AWS, Azure, and GCP.

SSO (Single Sign-On)

Security is always a tug-of-war between convenience and safety (Figure 7-21). As you increase security functions to lock down a resource, convenience decreases. Conversely, as you increase the convenience of authentication or other security processes, you sacrifice some of the assurance that your resources are secure.

Figure 7-21 Convenience versus safety



This conflict isn't just theoretical. If security measures are too inconvenient, people won't use them and will find ways to get around security protocols. On the flip side, if you focus more on making these processes convenient for users, your resources won't truly be secure. The secret to good security is to find and use tools that increase both convenience (so people will use them correctly) and safety (so they keep your resources secure). This way, you can avoid sacrificing safety when increasing convenience.

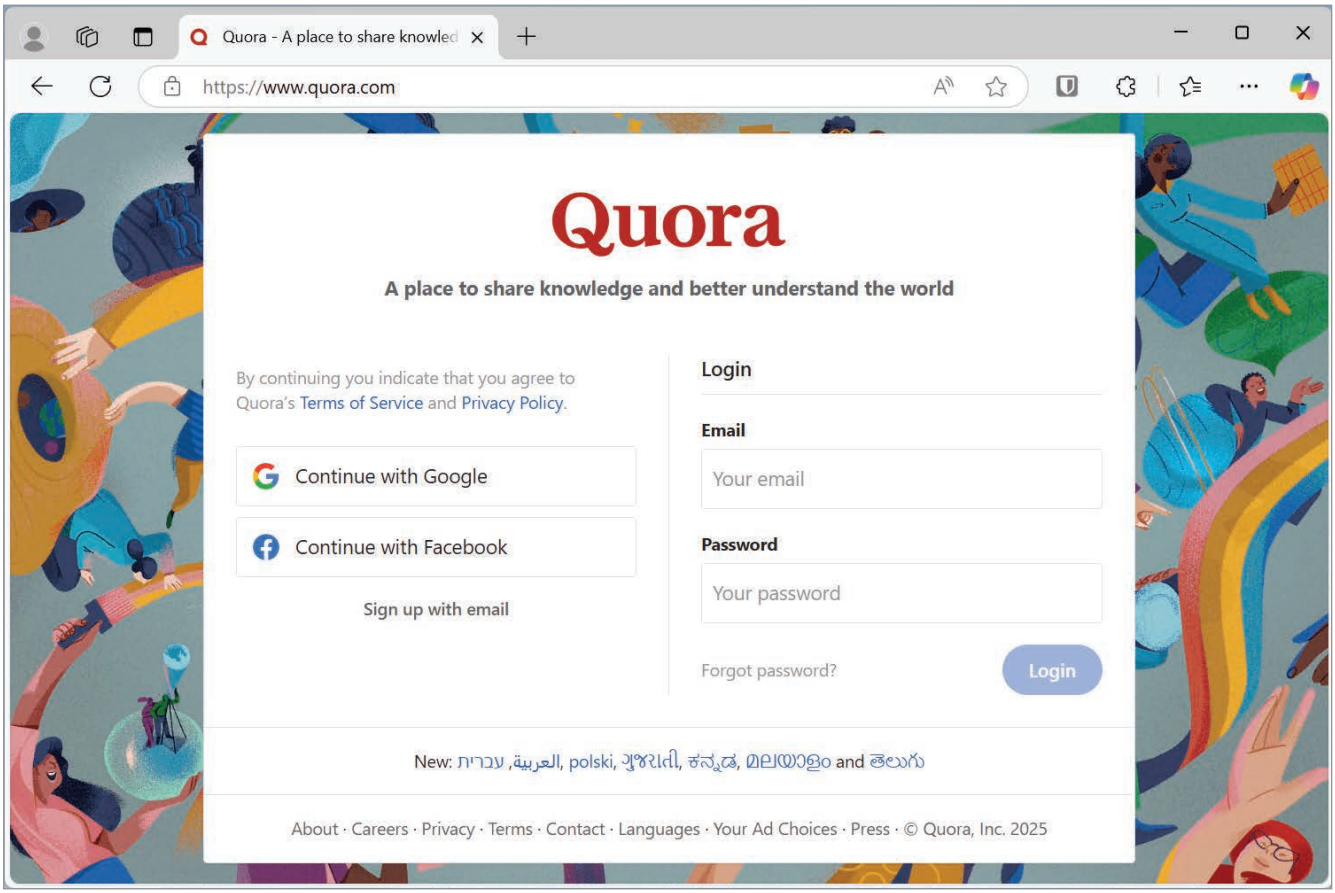
Single sign-on is one part of the solution to this conundrum. **SSO (single sign-on)** is a form of authentication in which a client signs in one time to access multiple systems or resources. This way, users don't have to remember several passwords and, depending on the system configuration, might not need to complete the authentication process multiple times in a single work session for every new resource they need to access.

Many applications are designed to support SSO functionality through the use of standards such as SAML (Security Assertion Markup Language) or OIDC (OpenID Connect), which relies on authorization by OAuth 2.0 (Open Authorization, version 2.0). These standards bridge the connection between IdPs (identity providers) and SPs (service providers). SP websites can subscribe to managed directory and identity services offered by an IdP. This way, the service provider does not have to manage identities and authentication processes themselves.

For example, you've likely encountered the option on some websites to "Sign in with Facebook" or "Continue with Google" (Figure 7-22). In these cases, Facebook and Google are acting as IdPs, offering a validated authentication of the user so the person can access whatever service the SP is offering (such as information, media, or retail access). To make applications readily compatible with these authentication mechanisms, developers often use pre-coded PAMs (pluggable authentication modules). A PAM is an authentication API designed to work with standardized input and output.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7-22 Google and Facebook both function as identity providers



Source: Quora, Inc

SSO offers the simplicity of using a single source of truth so user data isn't duplicated through multiple systems. Additionally, users themselves don't have to keep track of so many different sets of credentials. SSO can be implemented within a single organization that allows their users to access multiple services, or it might bridge services across multiple organizations. In this case, federation services create a trust relationship between organizations. **Federation**, also called federated identity, is the process of managing user identities across organizations on the foundation of this trust relationship.

Token-Based Authentication and Authorization

The CompTIA Cloud+ objectives list "token-based" as an authentication model, so let's chat about what this is. In reality, **token-based authentication** blurs the lines between authentication and authorization—it relies on generated tokens to validate a user's identity to services without requiring repeated authentication. Tokens are generated as part of the authentication process, and token-based systems are often used to enable SSO. Tokens both confirm the user's identity to other services, which is authentication, and allow evaluation of the user's permissions, which is authorization.

Different tools work with various parts of this process of authenticating the user, generating tokens, and managing the tokens to evaluate access permissions. For example, **OAuth 2.0 (Open Authorization, version 2.0)** is the industry-standard protocol for the authorization side of these processes. In general, here's how the process works:

- **Step 1**—Initially, the user submits their credentials to an IdP, which validates the user's identity. This process can be performed by any of several tools or services, and it can also include features such as MFA.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

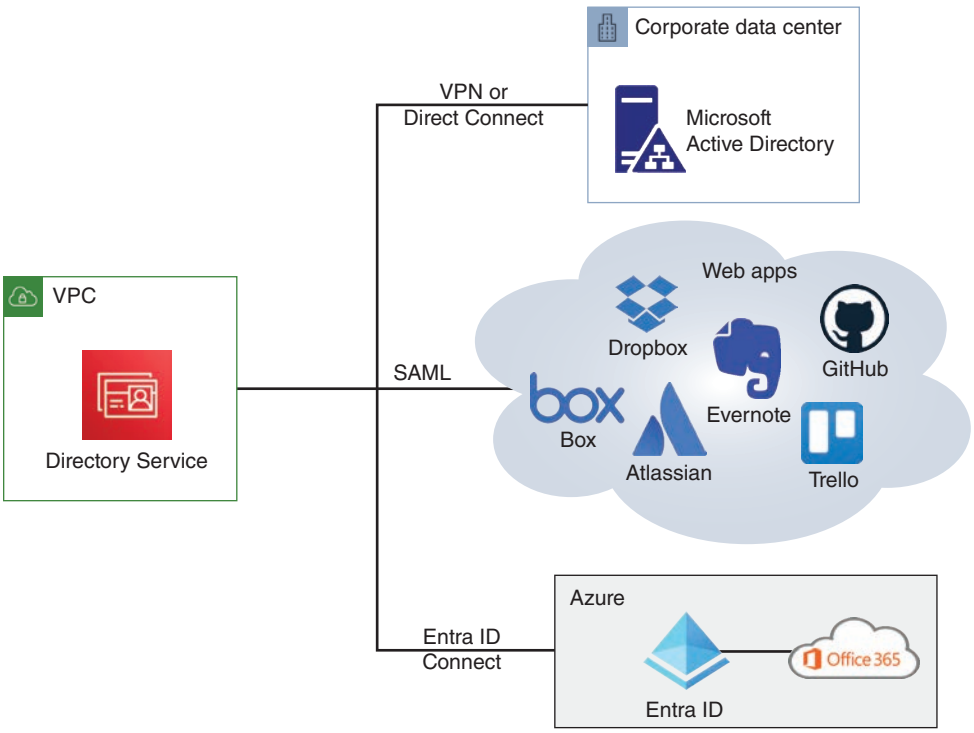
- **Step 2**—Once validated, the IdP generates a token that is provided to the user. This token includes information other systems will need to determine the user's access permissions. For example, the token might include the user's ID, applicable roles, and token expiration information.
- **Step 3**—The token is provided to the user's browser or other application, such as a mobile app.
- **Step 4**—As the user requests access to specific services, the user provides the token to the SP rather than needing to reauthenticate.
- **Step 5**—Each time an SP receives a request with a token, the SP validates the authenticity and integrity of the token and compares the request against the user's assigned roles and permissions. If the user's request is allowed, then the request is granted. Otherwise, the request is denied.

AWS IAM in a Hybrid Cloud

One form of hybrid or multi-cloud IAM relies on managing user access to individual applications or other kinds of systems. However, a more comprehensive approach to hybrid or multi-cloud IAM requires a blending of directory services, often through the configuration of trust relationships in Active Directory. Sometimes this is accomplished through an AD deployment or AD-native service in each cloud platform and on-prem.

AWS offers AWS Managed Microsoft AD, also called AWS Directory Service for Microsoft Active Directory (or just Directory Service for short). This service is essentially Active Directory in the AWS cloud. Directory Service is a managed directory service in AWS and can apply AD features, such as Group Policy and SSO, to other AWS services, such as EC2 or RDS (Relational Database Service). Directory Service can also support traditional AD applications such as Remote Desktop, SharePoint, and SQL Server. The AWS Directory Service coverage area can extend beyond AWS-hosted resources (Figure 7-23).

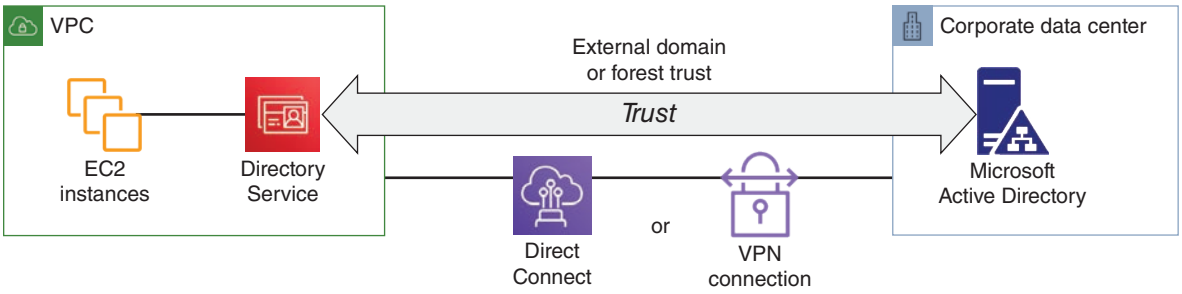
Figure 7-23 Use AWS Directory Service with many kinds of services



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Specifically relevant to extending IAM from an on-prem AD to AWS, an AD trust relationship can be established between AWS Directory Service and the on-prem AD (Figure 7-24).

Figure 7-24 Extend AAA from on-prem Active Directory to AWS

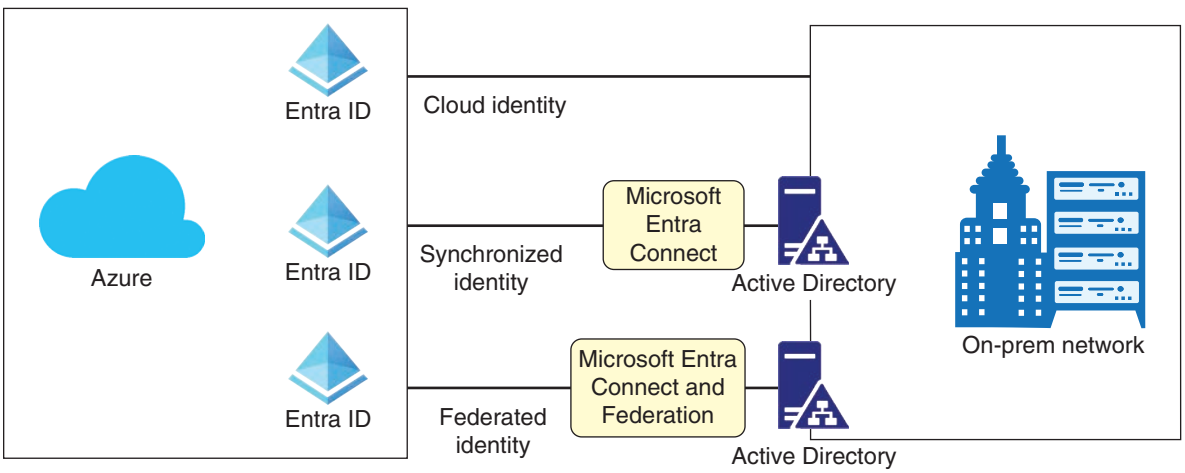


Azure IAM in a Hybrid Cloud

Azure offers three options for integrating IAM services in the cloud with on-prem workloads (Figure 7-25), as described next.

- **Cloud identity**—The pass-through authentication option consists of all IAM services residing in the cloud and includes limited synchronization of passwords or other credentials.
- **Synchronized identity**—With the synchronized authentication option, identities exist both on-prem and in the cloud, and password hashes are shared between the two locations. As changes are made in one place, that information is synced to the other location. However, there can be a delay of up to three hours when disabling a user on-prem, leaving cloud resources vulnerable to that user's credentials during the delay.
- **Federated identity**—Federated authentication is the more complicated and expensive option, offering more complete SSO functionality and immediate synchronization between locations so that approvals or deletions propagate throughout both the on-prem and cloud portions of the network simultaneously. This arrangement maintains a more consistent access policy.

Figure 7-25 Azure hybrid cloud options



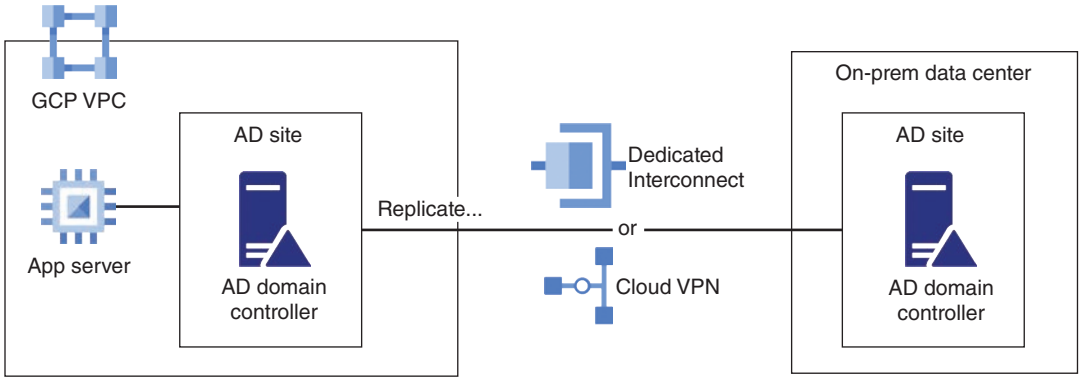
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

GCP IAM in a Hybrid Cloud

Google suggests using one of two approaches to extend identity services across a hybrid cloud, as follows:

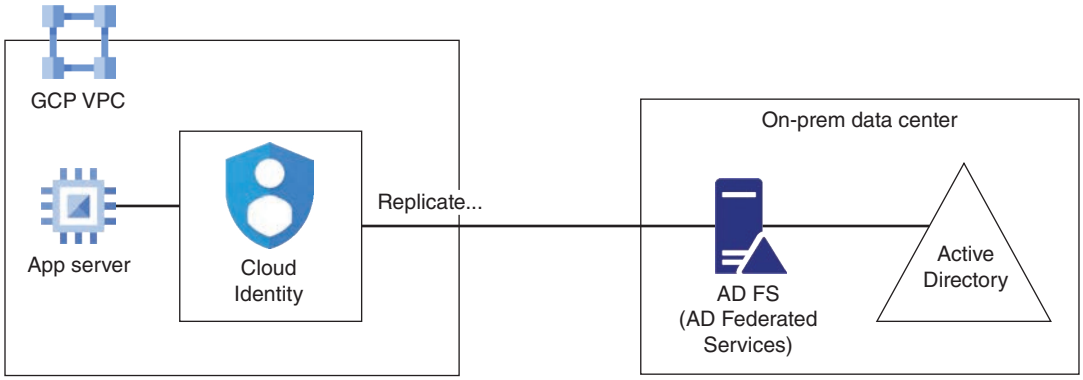
- Extend the existing IdP (such as Active Directory) to GCP. Figure 7-26 shows an example of an on-prem AD that has been extended across a Cloud VPN or an Interconnect connection to an AD domain controller running in a GCP VM.

Figure 7-26 The GCP-hosted AD domain controller services GCP resources



- Federate the existing IdP with GCP's Cloud Identity service. Cloud Identity is a robust, paid service within Cloud IAM that includes app and device management in addition to identity and access management. In Figure 7-27, GCP's Cloud Identity serves as an IdP for GCP resources, and a trust relationship is established between Cloud Identity and the on-prem Active Directory.

Figure 7-27 GCP's Cloud Identity service provides IAM functions for GCP resources



Self-Check

9. True or False: Active Directory can only be run on-prem.
   a. True, because AD must be run directly in a customer-managed server.
   b. False, because many cloud providers offer cloud-native AD services.
   c. True, because AD cannot be used to manage cloud users.
   d. False, because AD can manage any kind of user account in any kind of environment.

10. Which technique bridges the gap between IdPs and SPs?
   a. Source of truth
   b. IAM
   c. Federation
   d. AD

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 7-5: Troubleshooting Cloud IAM

Certification

4.4 Given a scenario, apply security best practices.

6.1 Given a scenario, troubleshoot deployment issues.

6.3 Given a scenario, troubleshoot security issues.

Most IAM problems you face when you begin working in the cloud are the responsibility of more qualified cloud administrators. However, there are some symptoms that, if you know what to look for, can indicate where the problem is located and what kind of action will be needed to fix the problem. Having these kinds of insights can help you ask better questions as you're initially investigating a problem, even if it will likely be escalated to a higher support tier.

Common IAM Issues

Consider the following common authentication and authorization issues:

- **Authentication issues**—Authentication issues can be caused by expired credentials, mistyped passwords, misconfigured certificates, or misconfigured trust relationships. Repeatedly attempting to access an account with wrong or expired credentials can cause it to lock out, which protects the account from brute force attacks. If a user is locked out, an admin with sufficient privileges will need to reset the account's credentials. Specific examples of issues and their solutions include the following:
  - **Expired certificate**—Most certificates renew automatically, or the system prompts an admin to approve the renewal process. An expired certificate error could be caused by an incorrect time on your local computer. If the certificate truly is expired, you might need to replace it with a new one.
  - **Misconfigured certificate**—First, confirm you're attempting to use the correct certificate. If that doesn't work, replace it with a new one with the correct configurations.
  - **Federation and single sign-on issues**—SSO issues are often caused by mistyping the username (such as entering an email address instead of the username, or vice versa). Double-check the credentials being used. Also check the trust configuration between the SSO service (such as AWS SSO) and the service provider's application. Keep an open mind that other items or processes could also be an unexpected cause of the problem. For example, authentication failures caused by a keyboard with a broken key won't be obvious if the password being typed is obscured.
- **Authorization issues**—Authorization issues are often caused by misconfigured permissions, groups, or roles. Read the error message carefully for clues, and check permissions, policies, and resource levels for referenced policies or permissions. Also consider that switching to a different internal role within your organization's account or to an external role, such as in a business partner's account, changes your permissions. You can only use one set of permissions at a time, so if you're not able to access the resources you think you should be finding, check that you're using the correct role.
- **Unauthorized access**—How do you monitor for and detect unauthorized access when it does occur, and then what can you do about it? Some access breaches can be detected fairly quickly. Something like a user logging into a system from an unusual location might be flagged right away, triggering alerts, notifications, or even a temporary shutdown of the user's account. But other types of breaches can be more long-standing. For example, in an APT (advanced persistent threat), the intruder gains unauthorized access to a system and camps there for a while, gathering intelligence to inform their next steps in the attack.

A lot of user behavior can be studied to detect patterns, which then can be used to establish baselines. This is often performed with a UEBA (user entity and behavior analytics) solution. When behavior exceeds

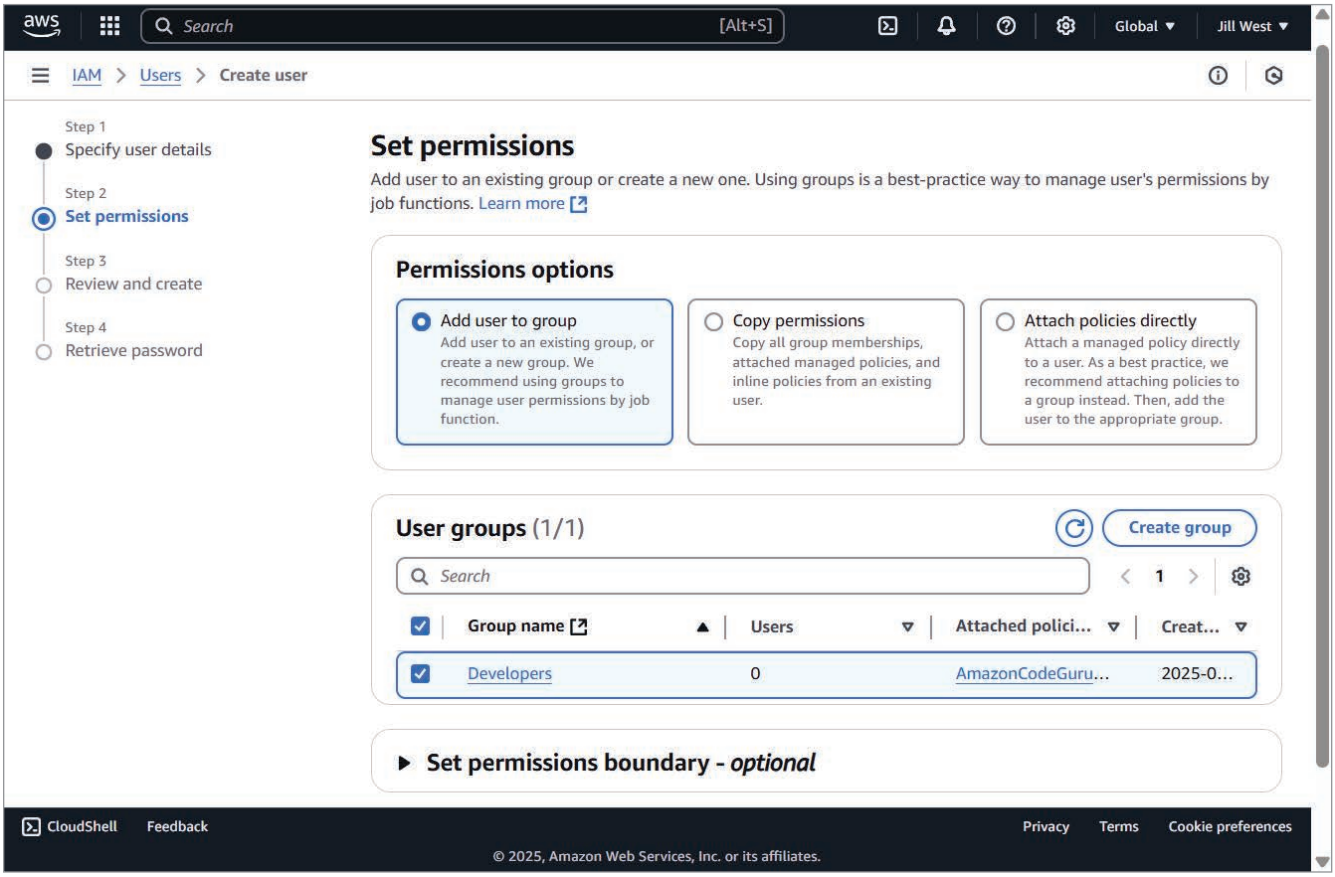
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

these baselines in some way, anomalous activities can be flagged for further investigation. Similarly, network traffic patterns can be established and monitored so that when certain types of traffic occur, such as unusual uploads that might indicate the transfer of sensitive data from your organization's systems to an attacker, that traffic will more likely be noticed and flagged. DLP (data loss prevention) solutions can detect when certain sensitive files have been tampered with or accessed in ways they shouldn't be. Tools like DLP can also prevent unauthorized use of data and other resources that might not be intentionally malicious but could still cause damage if allowed to continue.

Troubleshooting Privilege Issues

When assigning privileges for activities in the cloud, it's important to limit privileges to the minimum required for someone to do their job. As you learned earlier, this is called the principle of least privilege. In the interest of minimizing each person's permissions, however, you might find that someone doesn't have permissions they need for their job. Missing or incomplete privileges can be resolved by confirming the person does, in fact, need the identified privileges, and then assigning those privileges to that person's account. In AWS, for example, these privileges might be assigned by giving the user access to a new role or adding the user to a group that already has access to the needed role (Figure 7-28).

Figure 7-28 Add a user to a group to grant permissions



Source: Amazon Web Services, Inc.

In contrast to missing or incomplete privileges, attackers can also use existing privileges—or even just basic, unprivileged access—to perform a **privilege escalation attack**. There are two types of privilege escalation attacks:

- **Horizontal**—In a horizontal attack, the attacker uses typical user privileges to access information available to the compromised account, such as information about other users. This attack type is typically used for reconnaissance.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Vertical**—In a vertical attack, the attacker escalates privileges by manipulating privileges granted to the account being used or using the current account to gain access to more privileged accounts. This attack type requires more sophisticated skills to manipulate privilege assignments.

A system is vulnerable to these attacks through misconfigurations or system loopholes. For example, a user with permission to attach a role to a user group could give themselves a role with full administrative access. The following best practices can help reduce a cloud system's vulnerability to privilege escalation attacks:

- Use wildcards sparingly when creating permission policies because these wildcards can inadvertently grant permissions beyond the minimum required for a user to perform their job duties.
- Design permission policies as specifically as possible to avoid giving more privileges than intended.
- Design an overall IAM approach that clearly identifies the purpose of each policy and role.
- Use directory security groups to logically organize users according to the permissions they should have and to provide better oversight of each user's permissions.
- Watch for inherited permissions where membership in one group provides access that is denied in another group or in a subgroup.
- Require MFA for every user account.
- Enable logging as well as tools that monitor these logs, and frequently check the logs yourself to identify signs of unapproved changes to users, groups, roles, and policies.
- Use platform-specific tools designed to monitor IAM access and configurations.

Self-Check

11. An expired certificate is a(n) _____ issue.

a. authorization
b. accounting
c. authentication
d. privilege

12. What kind of attack causes changed account privileges?

a. External
b. Internal
c. Horizontal
d. Vertical

Check your answers at the end of this chapter.

Summary

Section 7-1: Cloud Accounts

- Once other user accounts are established, the root account should not be used again except when absolutely necessary.
- Users and resources (such as an application) both can have an identity, which is a digital entity to which you can grant permissions.
- An organization resource is an identity management object at the top of a hierarchy of accounts that is used to control access to all of a company's resources in that cloud.
- Effective account management policies can help ensure that accounts are provisioned and deprovisioned in a timely manner, compromised accounts are locked out, privilege creep is limited, and unused accounts are closed to further activity.

Section 7-2: Authentication

- Authentication is a service that proves the identity of a client (which could be a user, service, application, or device) and determines whether that client is allowed to access a secured system.
- Password policies can require that passwords meet certain requirements, such as complexity, length, expiration, and lockout.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- MFA requires two or more pieces of information—called factors—from two or more categories of authentication factors.
- A digital certificate is a small file containing verified identification information and the public key of the entity whose identity is being authenticated.
- Programmatic access refers to the ability to interact with cloud resources through a command line, SDK, API, or a third-party tool.
- A secrets management system handles the security and distribution of keys, passwords, authorization tokens, and other files used to secure access to resources.

Section 7-3: Authorization

- RBAC manages users' access to resources by assigning privileges and permissions to users via roles.
- In AWS, an identity-based policy is a policy assigned to identities (usually via roles) to define the permissions granted to those identities, such as access to resources.
- Azure manages IAM services with Entra ID, which is the cloud-native cousin of Active Directory found in Windows Server for on-prem networks.
- GCP's IAM service revolves around principals, which are entities that represent human users, groups of users, domains, or workloads.

Section 7-4: IAM for Hybrid Clouds

- SSO is a form of authentication in which a client signs in one time to access multiple systems or resources.
- Directory Service is a managed directory service in AWS and can apply AD features, such as Group Policy and SSO, to other AWS services.
- Azure offers three options for integrating IAM services in the cloud with on-prem workloads: cloud identity, synchronized identity, and federated identity.
- Google suggests using one of two approaches to extend identity services across a hybrid cloud: extend the existing IdP to GCP or federate the existing IdP with GCP's Cloud Identity service.

Section 7-5: Troubleshooting Cloud IAM

- Authentication issues can be caused by expired credentials, mistyped passwords, misconfigured certificates, or misconfigured trust relationships, whereas authorization issues are often caused by misconfigured permissions, groups, or roles.
- Missing or incomplete privileges can be resolved by confirming the person does, in fact, need the identified privileges, and then assigning those privileges to that person's account.

Key Terms

For definitions of key terms, check the Glossary.

AD (Active Directory)
authentication
authorization
consumer account
digital certificate
federation
IAM (identity and access management)
identity
local user authentication
MFA (multifactor authentication)
OAuth 2.0 (Open Authorization, version 2.0)
organization
PAM (privileged access management)
password vault
PKI (public key infrastructure)
principle of least privilege
privilege escalation attack
programmatic access
RBAC (role-based access control)
role
root
SDK (software development kit)
secrets management
service account
SSO (single sign-on)
token-based authentication
user account
workforce identity

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

CLI command line interface
CLI Common Language Infrastructure
IAM identity and access management
LDAP Lightweight Directory Access Protocol
MFA multifactor authentication
OAuth Open Authorization
OIDC OpenID Connect
SAML Security Assertion Markup Language
SDK software development kit
SSO single sign-on

Review Questions

1. Which of the following employees will most likely need a cloud privileged user account?
   a. A sales manager who trains sales reps on how to use a new application
   b. A custodian who needs access to every office for cleaning
   c. A new hire who must submit employment documents to HR
   d. A technician who helps users reset their passwords
2. Which part of AAA lets you perform the creation of an EC2 instance?
   a. Accounting
   b. Authentication
   c. Authorization
   d. Auditing
3. When signing an electronic document, what secret must you keep safe in your possession?
   a. The other party's public key
   b. Your identity
   c. Your private key
   d. The other party's private key
4. Which of the following combinations is MFA?
   a. ATM card and smartphone
   b. Fingerprint and password
   c. Password and PIN
   d. Fingerprint and facial pattern
5. Which key is used to decrypt information contained within a server's digital certificate?
   a. The server's private key
   b. The server's public key
   c. The client's private key
   d. The client's public key
6. Which access control method is controlled by individual users?
   a. DAC
   b. MAC
   c. RBAC
   d. GBAC
7. Haphazardly adding permissions every time someone needs to do a new task causes increased _____.
   a. inheritance drift
   b. privilege creep
   c. privilege escalation attacks
   d. blast radius
8. Which account type would be most appropriate for a software developer who tests code on a variety of servers?
   a. Workforce account
   b. Consumer account
   c. Root account
   d. Administrative account
9. When would a technician most likely need to use a privileged account?
   a. To store a report
   b. To restore a backup
   c. To update a database record
   d. To check email
10. What process does a token replace so it doesn't have to be repeated with each access request?
    a. Authentication
    b. Accounting
    c. Auditing
    d. Authorization

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Scenario-Based Questions

Question 7-1

Kason works for a real estate company that owns and manages apartment complexes. He recently deployed a new application designed to track maintenance calls for apartments across three of his company's larger properties. The information is stored in an AWS RDS database. Kason designed the database with the intent of expanding this app to track much more information than just maintenance calls. However, he's currently experiencing problems with the connection between the app and the database. It seems the database is not yet aware that the app is allowed to work with the information stored there. What kind of IAM account should Kason check to confirm its permissions are configured correctly?

a. Service account
b. Root account
c. Consumer account
d. Workforce account

Question 7-2

Paulo has configured multifactor authentication on his AWS account using a virtual MFA application on his phone. When he signs into his account on a computer in his school's computer lab, he is prompted to enter his password, and then he's asked for a code from his phone's app. Which category of MFA does the code represent?

a. Something you are
b. Something you know
c. Something you have
d. Something you do

Question 7-3

Alicia and her two business partners have been steadily building their homegrown company that sells custom-embroidered bags and other accessories. They've recently received a few large contracts that required them to invest in new machinery, and they want to expand their customer base to help keep that machinery working at full capacity. Alicia, the IT specialist in her company, decided to invest in a Salesforce subscription so that she, her business partners, and their sales reps can better track their sales efforts. She already has a website running in AWS and wants to set up SSO from her AWS account to her Salesforce account. Which standard makes this possible?

a. AAA
b. SSH
c. PKI
d. SAML

Question 7-4

When Lisa needs access to make edits to the tables in a database, her boss gives her user account the same role he has. What tenet of IAM security has been compromised?

a. Privilege escalation
b. Password complexity
c. Mandatory access control
d. Least privilege

Hands-On Projects

Note 7-5 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Project 7-1: Store Secrets in AWS KMS and Secrets Manager

Estimated time: 30 minutes
Objective 4.4: Given a scenario, apply security best practices.
Resources:

- AWS account
- Internet access

Description: AWS offers KMS for securely storing and rotating encryption keys and Secrets Manager for securely storing other kinds of secrets such as passwords. In this project, you practice using both services. Complete the following steps:

1. In your AWS Management Console, go to the **KMS** dashboard. KMS is listed under the **Security, Identity, & Compliance** service category.
2. In the left pane, click **AWS managed keys**. On this screen, **take a screenshot** of the AWS managed keys listed; submit this visual with your responses to this project's questions.
3. Notice you do not have the option to create keys here, as all these keys are managed solely by AWS for your use with various services. Click one of the listed keys and read the information block at the top of the page. You cannot change AWS managed keys, but what *can* you do with them?
4. Click **Customer managed keys**, and then click **Create key**. Recall that a symmetric key is used to encrypt and decrypt text, while asymmetric keys work in pairs where one encrypts and the other decrypts. Choose **Symmetric**. Then click **Advanced options**.
5. You're given four options for key material origin. Although you can import your own key material, the service provides the most automation when you use KMS-generated keys. Make sure **KMS** is selected.
6. A single-region key is never replicated for backup, but a multi-region key is copied to other regions for backups. Additionally, a multi-region key can be used to decrypt data that was encrypted by one of its copies in another region, such as when data is duplicated across regions. This feature supports disaster recovery, global data management, and other data encryption scenarios. However, multi-region keys are also more expensive. Make sure **Single-Region key** is selected. Click **Next**.
7. Give the key an alias, such as my-key. Click **Next**.
8. Without changing any permissions, read the information about key administrators on this page and key usage permissions on the next page. What is the difference between key administrators and users?
9. Finish creating the key without changing any other settings. Click through the information about your key. According to the documentation, why might you want to perform an on-demand rotation? How many times can you perform an on-demand rotation of each key?
10. Go to the **Secrets Manager** dashboard. Secrets Manager is listed under the **Security, Identity, & Compliance** service category. Click **Store a new secret**. If you're using an AWS Academy account, you might get a notification that you are not authorized to list DocumentDB clusters. This is not a problem for this project—you can clear the notification.
11. Secrets are typically credentials for other services. What types of secrets can you store in Secrets Manager?
12. Select **Other type of secret**. Each key/value pair defines a field of information for the secret you're creating. In the first row, type **Username** in the first field and type your name in the second field. Click **Add row**. In the new row, type **Password** in the first field and make up a password to type in the second field. Notice the password is displayed in plaintext.
13. You have the option of creating your own encryption key to encrypt your secret, or you can have KMS create a key for you. Leave this field at its default value of **aws/secretsmanager** so KMS will handle the key for you. Click **Next**.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14. Give the secret a name and click **Next**.
15. Secrets Manager can automatically rotate the secret for you. However, you have to set up the code in Lambda to perform the rotation or use one of the services that offer managed rotation, such as Amazon ECS (Elastic Container Service) or Amazon RDS (Relational Database Service). If you require a Lambda function, AWS offers a template you can alter so you don't have to write the code from scratch. For this project, you will not enable automatic rotation. Click **Next**, and then click **Store**.
16. Click your secret to display its details page. On the Overview tab, click **Retrieve secret value**. Notice the information you stored is now displayed in plaintext because your account has permission to decrypt this data.
17. Return to the **KMS** dashboard. What new key is listed on the **AWS managed keys** page?
18. Delete all the resources you created in this project, including the KMS customer managed key and the secret in Secrets Manager. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
19. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 7-6 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 7-2: Store Secrets in Azure Key Vault

Estimated time: 30 minutes
Objective 4.4: Given a scenario, apply security best practices.
Resources:

- Azure account and subscription
- Internet access

Description: Azure offers the Key Vault service for secrets management, key management, and certificate management. In this project, you practice using Key Vault for secrets and keys. Complete the following steps:

1. In your Azure portal, in the search box, type **key vault**, and then click **Key vaults** in the search results.
2. Click **Create key vault**. Use the following details as you begin to configure your key vault:
   - Create a new resource group.
   - You must use a unique key vault name, which might require some creativity in naming.
   - Keep the Standard pricing tier selection.
   - Click **Next**.
3. On the Access configuration tab, what is the default permission model for key vault data plane access?
4. Select **Vault access policy**. What changes on the Access configuration page when you make this selection?
5. In the Access policies section, click **Create**. What are the three categories of available permissions?
6. In the Secret permissions column, select **List**. Click **Next**.
7. For Principal, search for and select your account from the list. Click **Next** twice, and then click **Create**.
8. On the Networking tab, disable public access. Click **Review + create**, and after the validation completes, click **Create**.
9. When the resource is created, navigate to its resource page. In the resource's navigation pane, click **Objects**, and then click **Secrets**. What message is displayed?
10. To enable public access, click **Settings** and then click **Networking**. Select **Allow public access from all networks** and click **Apply**. After the settings are successfully updated, return to the list of secrets. What message is displayed now?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11. Click **Generate/Import**. To avoid fees, you will ***not*** create the secret. Use the following details as you configure your secret:
    - Type text in the Secret value field. Before you click from that field, click the eye icon to view the plaintext. When you click in a different field, the view in plaintext option will disappear.
    - Creating the secret will incur fees. Instead, click **Cancel**.
12. In the resource's navigation pane, click **Access control (IAM)**. On the Check access tab, click **View my access**. What is your role assignment for this resource? What scope is applied to your role assignment?
13. Close the assignments blade, and then click the **Roles** tab. Which role do you think would be the best fit for an employee who needs to audit objects stored in this key vault without editing the objects?
14. In the resource's navigation pane, click **Overview**. Scroll down and click the **Tools + SDKs** tab. **Take a screenshot** showing the tools and SDKs that can have programmatic access to your key vault; submit this visual with your responses to this project's questions.
15. Delete all the resources you created in this project, including the key vault and resource group. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
16. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 7-7 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 7-3: Store Secrets in GCP Secret Manager

Estimated time: 20 minutes
Objective 4.4: Given a scenario, apply security best practices.
Resources:

- GCP account
- Internet access

Description: GCP offers KMS for securely storing and rotating encryption keys and Secret Manager for securely storing other kinds of secrets such as passwords. Working with KMS requires a CLI. Therefore, in this project, you only practice using Secret Manager. Complete the following steps:

1. In your GCP console, create a new project.
2. In the navigation menu, point to **Security** and click **Secret Manager**, which is listed in the Data Protection section. Enable the Secret Manager API.
3. When the API is enabled, click **CREATE SECRET**. Give the secret a name. What's the maximum size file you could upload to store as a secret?
4. Enter text for the Secret value. **Take a screenshot** showing the CRC-32C checksum; submit this visual with your responses to this project's questions.
5. What are the two encryption key options to encrypt the secret?
6. Click **CREATE SECRET**. After the secret is created, click the dots icon under Actions and click **View secret value**. How can you know if the displayed secret value is the correct value?
7. Click **DONE**, and then click the **LOGS** tab. **Take a screenshot** showing the log entries for your interactions with this secret; submit this visual with your responses to this project's questions.
8. Click the **VERSIONS** tab, and then click **NEW VERSION**. Enter new text for the Secret value, and then click **ADD NEW VERSION**. Return to the **LOGS** tab. What has changed?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9. Click the **PERMISSIONS** tab. What is your role in relation to this secret?
10. Go to the list of projects and shut down the project you created in Step 1. Confirm the project is shut down and scheduled for deletion.
11. Make sure to properly sign out of your account.

Note 7-8 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Challenge Projects

Note 7-9 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 7-1: Practice Interpreting IAM Policies

Estimated time: 20 minutes (+10 minutes for group work, if assigned)
Objective 4.5: Given a scenario, apply security controls in the cloud.
Group work: This project includes enhancements when assigned as a group project.
Resource:

- Internet access

Description: In the chapter, you read about IAM policies written in JSON or YAML. Other cloud platforms might use other languages, such as CEL (Common Expression Language) in GCP. You should be able to examine an IAM policy in a common language and determine the basic effect of the policy. In this project, you'll practice creating JSON policies with an AWS policy generator or using a different policy generator of your choice. Complete the following steps:

1. AWS offers a free policy generator tool and other cloud platforms might offer similar tools. For this project, you can use AWS's policy generator at awspolicygen.s3.amazonaws.com/policygen.html, or search for an updated link. The AWS policy generator does not require an AWS account, and these instructions are written for this tool. Alternatively, you can search for a different IAM policy generator and adjust the steps as needed.
2. Select **IAM Policy** as the type of policy.
3. Experiment with adding statements containing all the required elements. For example, you might add a statement that allows EC2 instance creation, which requires the action ec2:RunInstances.
4. Make up an ARN with generic information in the required fields. For example, for this EC2 statement, you could use the following data: arn:aws:ec2:$Region:$MyAccount:$NewResource:$AnyPath. Notice the generator simply needs text as a placeholder, not valid data.
5. Generate the policy.
6. In the policy, identify each of the data points that you initially provided, noting where this information appears in the policy. **Take a screenshot** of your policy and label each component of the policy, including actions, effect (that is allow or deny), and resources; submit this visual with your responses to this project's questions.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7. Try again with two or three other services, each with two to four actions included in the policy. You might use a GenAI tool, such as ChatGPT, to help you choose relevant actions for your selected services. Additionally, you might use a search engine to figure out the name of an action for something specific that you want to allow or deny. Make sure to pause after generating each policy. Examine the policy statements closely to understand how your information was translated into the required code. This will make it easier later for you to interpret policies. What patterns do you notice about how these policies are structured?
8. **For group projects:** Generate two new policies that incorporate simple actions you've practiced performing in projects earlier in this course. Share your policies with another group member. When you receive your group member's policies, write a detailed, plain-language description of the policies' effects. Then check your group member's descriptions of your policies to determine if your classmate accurately interpreted your policies. Give each other feedback on your policies and your interpretations. What are some actions in your group member's policies that you had not practiced incorporating into your own policies? What parts of your group member's policies were challenging to interpret?

Solutions to Self-Check Questions

Section 7-1: Cloud Accounts

1. A user account is a(n) _____.
   Answer: d. identity

   Explanation: Users and resources (such as an application) both can have an identity, which is a digital **entity** (such as a user account) to which you can attach roles and permissions. Some privileged accounts have elevated permissions, and root is a special privileged account that has full administrative access throughout all services in the cloud account. Identities and their credentials are stored in a repository called a vault.

2. What is a goal of successful account management?
   Answer: c. Limit privilege creep

   Explanation: Account management policies can help ensure that **privilege creep is limited** and monitored. Giving some users multiple accounts with different privileges can limit the blast radius of a critical event, such as a security breach. Automation can help decrease time required for account provisioning.

Section 7-2: Authentication

3. Which protocol defines how most authentication directories work?
   Answer: a. LDAP

   Explanation: The mechanisms of **LDAP (Lightweight Directory Access Protocol)** dictate some basic requirements for any directory it accesses. Therefore, directory servers are configured and function in similar ways, regardless of the software used. OCSP (Online Certificate Status Protocol) offers an updated and scalable method of tracking certificate revocation. SSL (Secure Sockets Layer) and TLS (Transport Layer Security) are used to encrypt HTTP (Hypertext Transfer Protocol), resulting in HTTPS (HTTP Secure). MFA (multifactor authentication) increases authentication security by requiring two or more authentication factors from two or more factor categories.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. If a certificate is included in a CRL, what is its status?

Answer: b. Revoked

Explanation: Once **revoked**, a certificate is added to a CRL (certificate revocation list), which other entities can check to confirm whether a certificate has become invalidated. A certificate is retired when it reaches its expiration date. A certificate is good, or trusted, during the timeframe when it was approved if there is no indication of compromise. A renewed certificate's validity period has been extended so that it continues to maintain a trusted status for a longer period.

5. Which access type is most appropriate for application-to-application communication?

Answer: b. API

Explanation: Systems, such as applications, communicate with each other via **API (application programmable interface)**. A human user can access the cloud via a GUI (graphical user interface) or a CLI (command line interface). Developers can access cloud resources indirectly using an SDK (software development kit).

Section 7-3: Authorization

6. The following policy grants what kind of permission?

```
"Sid": "FirstStatement",
"Effect": "Allow",
"Action": "s3:ListAllMyBuckets",
"Resource": "*"
```

Answer: b. Read

Explanation: The List action allows the user to display and **read** a list of all the buckets in the account. To open the bucket and view objects in it would require the Get action. Edit permissions would require the Put action. Delete permissions would require the Delete action.

7. What access control method does Azure use?

Answer: b. RBAC

Explanation: Azure relies on **RBAC (role-based access control)** to manage users' access to resources. AD (Active Directory) is not an access control method. DAC (discretionary access control) is typically used in less formal situations, such as on a home network. MAC (mandatory access control) requires that users and resources be organized into a hierarchical security classification system.

8. Groups are used to organize _____.

Answer: d. users

Explanation: GBAC (group-based access control) is used to organize **users**. RBAC (role-based access control) is used to organize permissions according to roles. Permissions are used to grant access to resources.

Section 7-4: IAM for Hybrid Clouds

9. True or False: Active Directory can only be run on-prem.

Answer: b. False, because many cloud providers offer cloud-native AD services.

Explanation: This statement is **false because many cloud providers offer cloud-native AD services**; additionally, Active Directory can be run in a VM instance on the cloud. AD is not

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

limited to customer-managed servers, and it can be used to manage cloud users. However, there are limitations on scenarios where AD is an appropriate fit. For example, the chapter discussed using Cognito to manage mobile app users rather than setting up more formal accounts through IAM or AD.

10. Which technique bridges the gap between IdPs and SPs?

Answer: b. Federation

Explanation: Federation is the process of managing user identities across organizations on the foundation of an established trust relationship. A repository containing identities and their credentials is sometimes called a source of truth—this repository must be shared for federation to succeed, but its existence is not, by itself, what bridges this gap. A typical IAM (identity and access management) service is limited to a single cloud platform—other services are required to bridge the gap to other systems. Establishing trust relationships between AD (Active Directory) deployments can blend directory services, but a single AD deployment provides AAA only for a single location, such as on-prem.

Section 7-5: Troubleshooting Cloud IAM

11. An expired certificate is a(n) _____ issue.

Answer: c. authentication

Explanation: An expired certificate can result in an **authentication** failure. Once authentication is complete, the certificate is not relevant for authorization or accounting processes. Privileges are defined by authorization processes.

12. What kind of attack causes changed account privileges?

Answer: d. Vertical

Explanation: In a **vertical privilege escalation attack**, the attacker escalates privileges by manipulating privileges granted to the account being used or using the current account to gain access to more privileged accounts. In a horizontal attack, the attacker uses existing privileges to gather information available to the compromised account. Internal and external accounts can both be used for either privilege escalation attack type.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Chapter 8

Cloud Storage and Databases

Chapter 8 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Compare storage technologies.
2. Explain options for storage configurations.
3. Configure common cloud storage services.
4. Describe options for creating and storing backups.
5. Identify storage security techniques.

Chapter 8 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 8-1: Storage Types
- Data Types
- On-Prem Storage Technologies
- Cloud Storage Technologies
- Types of Databases

Section 8-2: Storage Configuration
- Tiering
- Storage Performance
- Storage Migration
- Data Destruction

Section 8-3: Cloud Storage and Database Services
- AWS Storage Services
- Regulations for Storage Security
- Azure Storage Services
- GCP Storage Services

Section 8-4: Creating and Storing Backups
- Backup Strategies
- Backup Retention
- Backup Types
- Backup Considerations

Section 8-5: Storage Security
- Storage Security Techniques
- Data Classification
- Data Obfuscation
- Regulations for Storage Security

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 8 Scenario "Someone stole my snack again!" you exclaim, feeling frustrated as you dig through your desk drawer searching for a stray granola bar or something else edible.

"What happened?" asks Kendra as she leans against the doorway to her office.

"I brought some leftover quesadillas to hold me over until our late lunch meeting this afternoon. I put them in the refrigerator in the snack room *with my name on the box*," you emphasize. "But now they're gone! Some people have no manners."

"I have some extra yogurt and a brownie you can have," offers Malik. "I made the brownies myself!"

"I guess it's important to have backups," laughs Kendra.

As you rummage through your desk drawer a little more, you mumble, "I thought I had an old granola bar in here . . . where *is* that thing?"

"A stale granola bar in your desk drawer hardly qualifies as a good backup," she points out.

Kendra turns to go back into her office and then stops and comes back. "Hey, can you two take these file boxes to the storage unit for me? That'll get you out of the office for 20 minutes or so—it's a beautiful day, and you can hit the drive-through for a snack while you're out."

"Yeah, sure," you say as you stand up and grab your jacket. Then you add, "Thanks."

"No problem," says Kendra. "It's good timing anyway—when you get back, we need to talk about how to design some storage space in our cloud."

"What kind of storage space do we need?" asks Malik. "Files and folders?"

"Not exactly," replies Kendra. "We'll need a file system to support a couple of applications, but mostly we'll use block storage and object storage."

"What are those?" you ask, genuinely intrigued.

"Well, block storage is kind of like our storage unit where you're taking these boxes. Each box on the shelf is like a block of storage space. You can put a certain amount of stuff in there, and you track the location of the stuff based on the name of the box."

"What about object storage?" Malik asks.

"That's a little different," says Kendra. "That's more like . . . well, that's kind of like the stuff stored on your desk," she laughs.

"Hey, I like my desk! I can find everything I need," Malik says, as you all three glance over his desk piled with papers, boxes, bottles, electronics, books, and all kinds of other things. You're pretty sure you saw a rubber chicken staring out at you from one of the piles last week.

"This is true," Kendra says. "*You* can always find exactly what you need because, in your mind, you know where everything is. But from the outside, it just seems to be piles of *stuff*. Object storage is the same idea: You pile lots of different kinds of things together in whatever amount of space it takes. It's a very flexible system that can handle massive amounts of data."

"Okay, that kind of makes sense," you nod. "I guess we won't need to back up anything we move to the cloud, right? I mean, the cloud provider keeps everything backed up already, don't they? It's kind of like going to a restaurant—there's always food there."

"Ah, well, that's a common misconception," explains Kendra. "Not everything in the cloud is backed up automatically. Just like your favorite restaurant can be out of your favorite food, or they can be closed, or maybe you can't eat out every time you're hungry, you need to have backup plans for food when something goes wrong. And with the cloud, you still need to back up your data."

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

"Tell you what," Kendra adds. "Let's do a little research . . .," she says, as she pulls a spare notepad off Malik's desk.

- What are different ways data is stored in the cloud?
- How is data organized in cloud storage?
- What cloud services store data?
- How is data backed up in the cloud?
- What can be done to keep cloud data safe?

Section 8-1: Storage Types

Certification

1.4: Compare and contrast storage resources and technologies.

1.7: Compare and contrast virtualization concepts.

1.9: Explain the importance of database concepts.

Whether you're working on-prem or in the cloud, managing an enterprise-scale network, or typing a research paper on a home computer, you're creating, saving, and accessing data. If you think about it, all of IT exists for the sake of data. The companies that do data well have a competitive edge in their respective industries. Nearly every business decision hinges on the quality of data made available to decision makers.

Knowing the key role data plays in business highlights the need for good data storage methodologies. Compromised data can destroy an organization. Therefore, as companies move their data centers to the cloud, it's the responsibility of IT professionals to ensure that data is kept both safe and accessible.

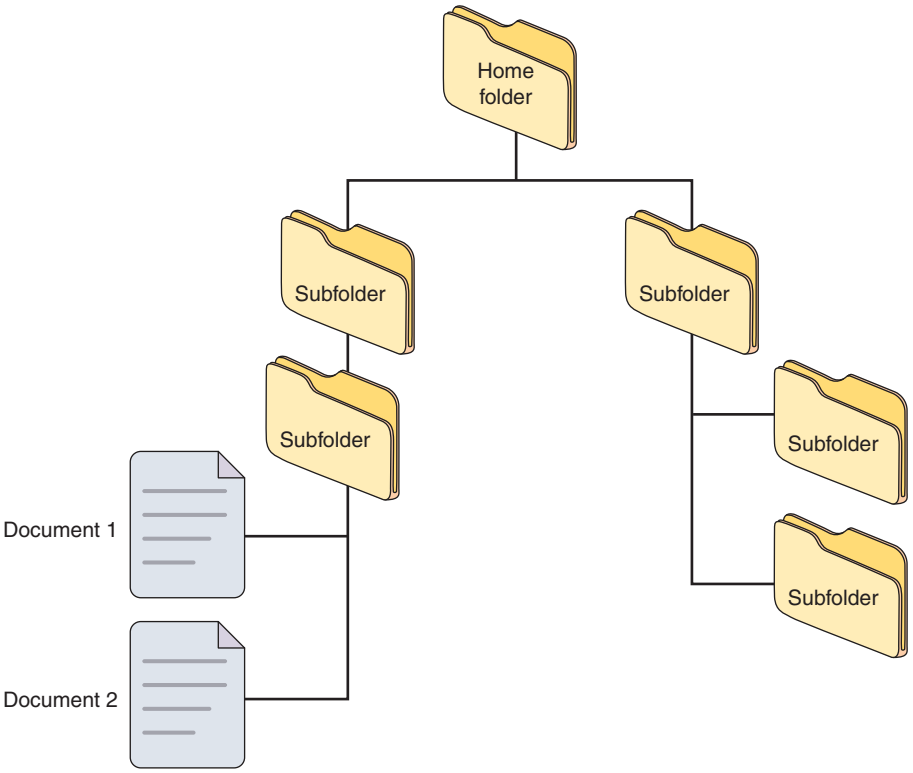
Data Types

Consider the many kinds of workloads that rely on data. A CRM (customer relationship management) application draws on data stored in databases. Media services access what are often very large media files, such as videos or high-resolution images. Office staff create documents, spreadsheets, and presentations that are stored on their workstations or in file shares accessed across their company's network. All these examples of data might be stored on-prem or in the cloud. Other examples of data types stored in the cloud include the following:

- **Office and media files**—Probably the most recognizable type of data, office files include file types such as documents, spreadsheets, and presentations. These files might be created in a desktop app (such as Word, Excel, or PowerPoint) or created in a web app (such as Google Docs, Google Sheets, or Google Slides). Media files include file types such as photos or videos. Files are then typically stored in a hierarchical file system (Figure 8-1). This file system might live on a local computer or in the cloud, such as in Microsoft OneDrive or Google Drive. Files are grouped within folders (also called directories), and a folder can also hold other folders (also called subfolders or subdirectories). Common file systems you might be familiar with include the old FAT32 (File Allocation Table), the newer NTFS (New Technology File System), the even newer ReFS (Resilient File System), the long-time default Linux file system ext4 (fourth extended file system), and the newer Linux file systems BTRFS (B-Tree File System, pronounced "butter f-s") and ZFS (Z File System, pronounced "zed f-s").

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-1 A file system consists of folders within folders



- **Applications**—Applications can be made up of many parts stored in various places on a network. These parts interact as needed to generate requested output. An application might run on an operating system, in a container, inside another application, or across the web. Applications often rely on data supplied by other sources, such as a database, an operating system, or another application.
- **Websites**—Websites usually consist of many files stored on a web server. Each webpage lives in a different file and often collects data from many files. A webpage might also pull data from other sources, such as APIs (application programming interfaces) from another website, a database, or an application.
- **Logs**—A log file, or log for short, is a special kind of file that collects information on events associated with a system or application. For example, a log might record information about traffic crossing a network interface. The information could include data about the time a particular network message was received, where it came from, where it was going, and whether it was allowed to pass. Keeping logs generates a tremendous amount of data in a short period and can take up a lot of storage space. But logs are important for troubleshooting problems or investigating security issues.
- **Databases**—A database is a large collection of data that is organized for quick retrieval of information. A phone directory is a simple kind of database, where information on thousands of people is organized by name, address, and phone number. The information for each person listed in the phone directory is contained within a record. These records can be filtered or sorted to investigate patterns, guide business endeavors, or interact with other databases. However, most databases are more complex. Databases often hold large amounts of data in ways that make it faster and easier to pull needed information from the data. The types of interactions performed on databases typically require special storage types. Therefore, in most cloud platforms, databases are considered a separate service category.

Clearly, data is integral to running a business, whether that's a small, one-person shop or a global organization. As an IT professional managing data in the cloud, you'll need to know how to keep data safe while also ensuring that it's accessible by those authorized to use it.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

On-Prem Storage Technologies

While cloud-based technologies are the focus of CompTIA Cloud+, on-prem storage technologies that use similar principles are also covered in the exam objectives. It's important that you understand how the principles of cloud computing can apply to on-prem technologies, and vice versa. Let's begin with a discussion of on-prem storage types.

On-Prem Storage Types

There are three basic networking approaches to storing data on-prem, as described next:

- **DAS (direct attached storage)**, the most familiar storage type to the typical user, is directly connected to the computer or server that uses it; this storage type is also called **local storage**. For example, the hard drive in a laptop is typically accessed only by that laptop. Even if a storage volume on the local drive is shared with other computers on the network, these files are only accessible if the storage drive's host computer is turned on. Attaching several disks (called "just a bunch of disks," or JBOD) to a server is also an example of DAS. These disks might be shared among several servers, but each server has its own connection directly to the disks and does not rely on a switched network to access data stored on the disks. These disks must initially be formatted using a partitioning system, such as the older MBR (Master Boot Record) or the newer GPT (GUID Partition Table), before installing a file system on it, such as NTFS.
- **NAS (network attached storage)** relies on network infrastructure to support clients' access to files. NAS is most often provided by a NAS device containing multiple disk drives and accessed by client devices over the network. The NAS device is attached to a network switch, which allows access from network servers and other networked devices. It relies on access protocols such as CIFS (Common Internet File System) or SMB (Server Message Block) for Windows systems, or NFS (Network File System) for Linux or UNIX systems. NAS often incorporates RAID to provide data redundancy, which would protect data in the event one or more disks fail. **RAID (Redundant Array of Independent Disks** or **Redundant Array of Inexpensive Disks)** is a collection of storage drives (called an array) interfaced in a way to ensure one or more layers of redundant copies of data across the drive array. NAS works well for SMBs (small or medium-sized businesses) that need to make on-prem data available to several users on the network but can't afford an expensive SAN, described next.

Note 8-1 | Although NFS (Network File System) has "file system" in its name, it's really a DFS (distributed file system) protocol used to give file server clients access to files across a network. A related technology is HDFS (Hadoop Distributed File System), where Hadoop is a framework for distributing big data across a network and efficiently processing that data in parallel functions.

Big data is large amounts of data generated at high speeds and collected from many sources, resulting in data of such high volume, variety, and velocity that special tools are required to capture, analyze, and store the data. Notice the "three Vs" of big data: volume, variety, and velocity. These massive data sets might be collected by companies, governments, or entire industries, and the data might be structured or unstructured. One example is global climate data made available by the National Oceanic and Atmospheric Administration at ncdc.noaa.gov/cdo-web.

- A **SAN (storage area network)** is a specialized portion of an organization's network dedicated entirely to supporting storage hardware (collectively called a storage array). Data in a SAN array is stored in blocks rather than in a file system hierarchy. Each block is a specific size, much like storage units in a self-storage facility. Think about consistently sized storage units offered by Public Storage,

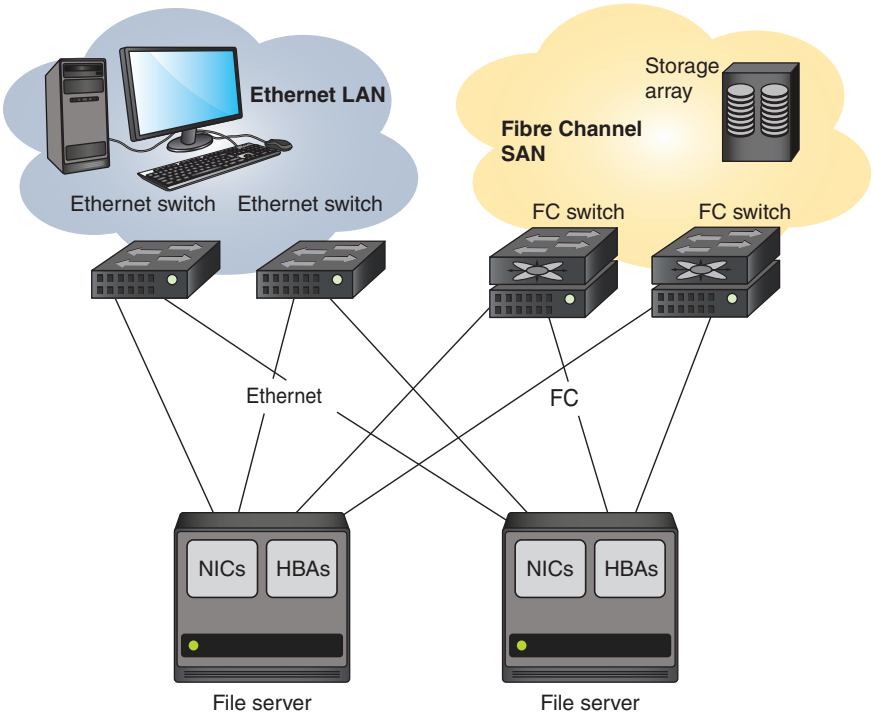
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Extra Space Storage, or CubeSmart, where you can rent a locker based on its size and store furniture or other items you don't want in your home. Each block in a SAN contains data that is referenced by the block's LUN (logical unit number). A server regards each block as a separate volume and can treat that volume as an attached drive, even installing a file system on it if the block size is large enough. The server then adds data to the block over time. Here are a few helpful points to consider when learning about SANs:

- SAN devices don't usually communicate by Ethernet like the rest of the network does—at least, not directly. Internally, the SAN typically relies on different kinds of access protocols, such as FC (Fibre Channel), SCSI (Small Computer Systems Interface, pronounced "scuzzy"), SAS (Serial Attached SCSI), SATA (Serial Advanced Technology Attachment or Serial ATA), or NVMe (non-volatile memory express).
- External to the SAN, variations of these protocols and other supporting protocols can make SAN devices more accessible to other networked devices, such as FCoE (Fibre Channel over Ethernet), FCIP (Fibre Channel over IP), IFCP (Internet Fibre Channel Protocol), iSCSI (Internet Small Computer System Interface, pronounced "i-scuzzy"), and ISNS (Internet Storage Name Service).
- Nodes on an FC network are identified by WWNN (World Wide Node Name)—which is similar to a MAC address on an Ethernet network—and by WWPN (World Wide Port Name) for individual ports. On an iSCSI network, nodes are identified by an IQN (iSCSI Qualified Name).

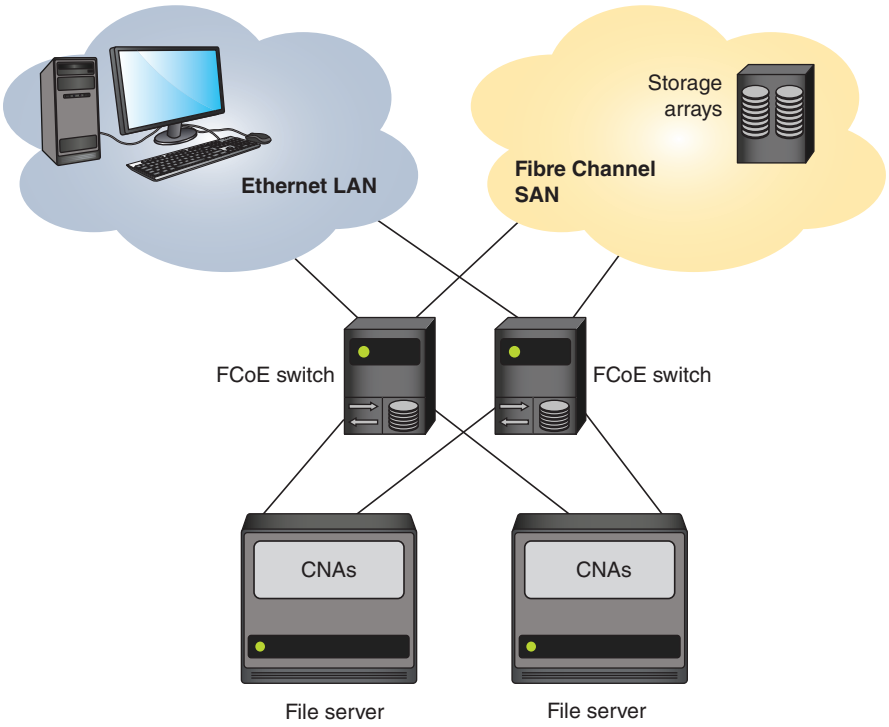
SANs are more complex and expensive to set up and maintain than other storage options. As illustrated in Figure 8-2, a SAN requires specialized switches that connect to servers using HBAs (host bus adapters) instead of NICs. However, more recent SAN technologies use CNAs (converged network adapters) or UTAs (unified target adapters), which can use typical, multipurpose LAN Ethernet devices in service of the SAN (Figure 8-3).

Figure 8-2 A Fibre Channel SAN connected to an Ethernet LAN



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-3 A SAN using FCoE to connect to a LAN



Within a single computer, you can choose from several standards to connect a storage device with the CPU: SCSI, SAS, SATA, or NVMe. Adjustments to these protocols are needed for them to work in a SAN. For example, iSCSI applies the basic principles of SCSI to SAN connections. Similarly, NVMe-oF (nonvolatile memory express over fabrics) applies NVMe technology to connections between servers and the SAN. Improvements in storage technology led to the transition from hard disks to SSDs (solid-state drives), thus improving performance of the disks themselves. However, the connection technologies at the time imposed a bottleneck that reduced experienced performance of storage devices. NVMe-oF offers connectivity performance capable of matching the data read and write performance of SSDs in the SAN.

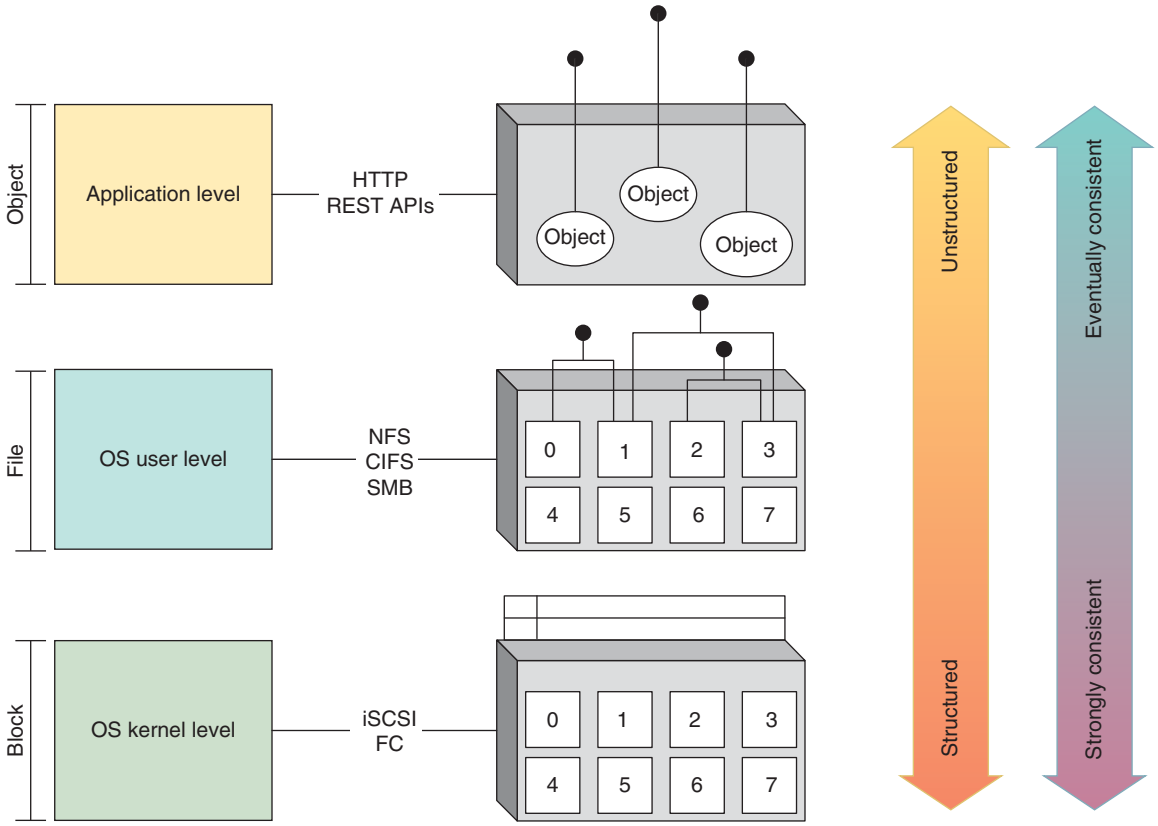
Exam Tip | In preparation for the CompTIA Cloud+ exam, make sure you understand the basic infrastructure for how a SAN connects to the rest of the network, what devices generally go where, and which protocols are implemented in each segment (that is, internal to the SAN, in connection with the SAN, and external to the SAN).

Cloud Storage Technologies

You've learned that one of the primary differences between NAS and a SAN is that NAS relies on a file system hierarchy while a SAN stores data in consistently sized blocks. Although you won't set up NAS or SAN storage in the cloud per se, these storage structures continue to provide helpful categories of storage services available in the cloud. Cloud storage services, also called STaaS (Storage as a Service), additionally offer one more type of storage: object storage. Figure 8-4 illustrates how these storage categories compare to each other. Let's explore a more detailed comparison of these three storage types.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-4 Each storage type serves different purposes



File System Storage

File system storage is easy to conceptualize if you think about putting papers in folders inside a file cabinet. You might have folders within folders, which are then organized by hanging folders. In a digital environment, you can theoretically create many more layers of virtual subfolders than would be reasonably accessible with physical folders.

File system storage lends itself well to organizing a single person's or a small department's files. However, it does not scale as well for large amounts of different kinds of data (such as a video repository) held in very large files (such as medical imaging) used by sizable teams distributed across various work locations. To address this issue, file system architects design a logical structure for the folder hierarchy with consistent naming schemes that avoid excessive levels of nesting folders within folders. Additionally, file system services in the cloud can support concurrent requests from large numbers of users (a functionality that is often limited with on-prem file systems), and load balancers can distribute requests across multiple file systems. Further, cloud file system services incorporate file- and directory-level permissions for fine-grained access control, consistency techniques to ensure changes synchronize across connected devices, and tiered storage classes for cost optimization.

Like many cloud storage options, cloud file system services integrate redundancy features to ensure data durability and availability. **Durability** refers to the resistance of data to errors due to corruption or loss. And these services offer many options for connecting configured file systems with many other kinds of workloads, applications, or resources—such as VM instances, containers, and serverless functions. A single file system deployment can support a large number of high-performance connections while implementing robust security measures.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For some situations, file system storage offers an interesting advantage over other storage types: the metadata that can be stored with each file. Metadata is data about the file, such as creation date, author, and permissions. This can be helpful when you're searching for a particular file or need to sort files by creation date, last edited date, or file size. In large, distributed file system deployments, this metadata is decentralized—meaning it is spread across multiple host nodes—in a process called sharding.

Block Storage

Block storage scales up easily because each block is addressed by a unique number rather than being organized into a hierarchy. This is similar to the difference between your street address versus your email address. Your home address might be identified by information at the following layers:

a number → on a street → in a zip code → in a city → within a state

The street name represents a group of dwellings; that group is part of a larger group represented by a zip code; and that zip code is part of a larger group of zip codes in a city, which is joined by many cities in your state. These are groups within groups within groups. With your email, however, you're identified only by your account name. Every other email account by the same provider (such as Gmail) also has a unique name, and they're not grouped within groups in a hierarchical structure. The namespace is said to be flat.

Block storage works much the same way as an email provider's namespace. Every block in the storage network is addressed by a unique LUN. This avoids the complexity from multiple layers of nested folders. It also allows for low-latency access to databases, virtual machines, boot drives, and other resources stored within the blocks.

In the cloud, block storage often acts more like local storage with a physical computer. For example, in AWS, the EBS (Elastic Block Store) service provides virtual storage drives that can be attached directly to an EC2 instance. Each storage volume supports a block of storage space. Within these storage volumes, you can install a file system, such as how Windows uses folders or how Linux uses directories to organize files stored on a hard drive.

Object Storage

Object storage is a more cloud-native storage infrastructure—it takes some of the best characteristics of file and block storage and reimagines storage in a way that capitalizes on the cloud's auto-scaling functionality. Similar to file systems, an object packages stored data along with metadata about the object. This metadata can give extensive information about the object and about the data stored within the object. This metadata is also optimized for quick search and retrieval functions. Where file systems keep basic metadata about a file, such as creation date and author, object metadata can contain extensive and customized information about the data itself. This is especially useful when the data can't be easily searched, such as with photos, videos, or complex CAD (computer-aided design) 3-D images.

Similar to block storage, object storage also relies on a flat namespace, that is, the directory structure is managed through a naming scheme where prefixes are used to identify related files and virtual folders, rather than positioning files and folders within multiple layers of folders. For example, Figure 8-5 shows the ARN (Amazon Resource Number) for an object in an AWS S3 (Simple Storage Service) bucket.

Notice the bucket and folder both appear in the ARN. However, the "folder" is not a separate entity in a true file hierarchy; it's simply another prefix added to the simulated file path. Unlike a folder in a file system, a folder in S3 has no metadata of its own and it can't be renamed (Figure 8-6).

Additionally, this space is not organized into consistently sized blocks. Objects are held in a single repository, or pool of storage. This space can be expanded indefinitely, even across data centers and regions. Unfortunately, the expansiveness and flexibility of the space can result in higher latency when retrieving data from large buckets. Still, object storage lends itself well to unstructured data storage (such as massive collections of images or video), archival storage, or big data storage (such as IoT raw data stored in a data lake).

Object storage offers one more benefit specific to the cloud environment: simpler protocols. Where block storage relies on complex protocols such as FC, object storage uses REST APIs and HTTP.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-5 Prefixes in a flat namespace

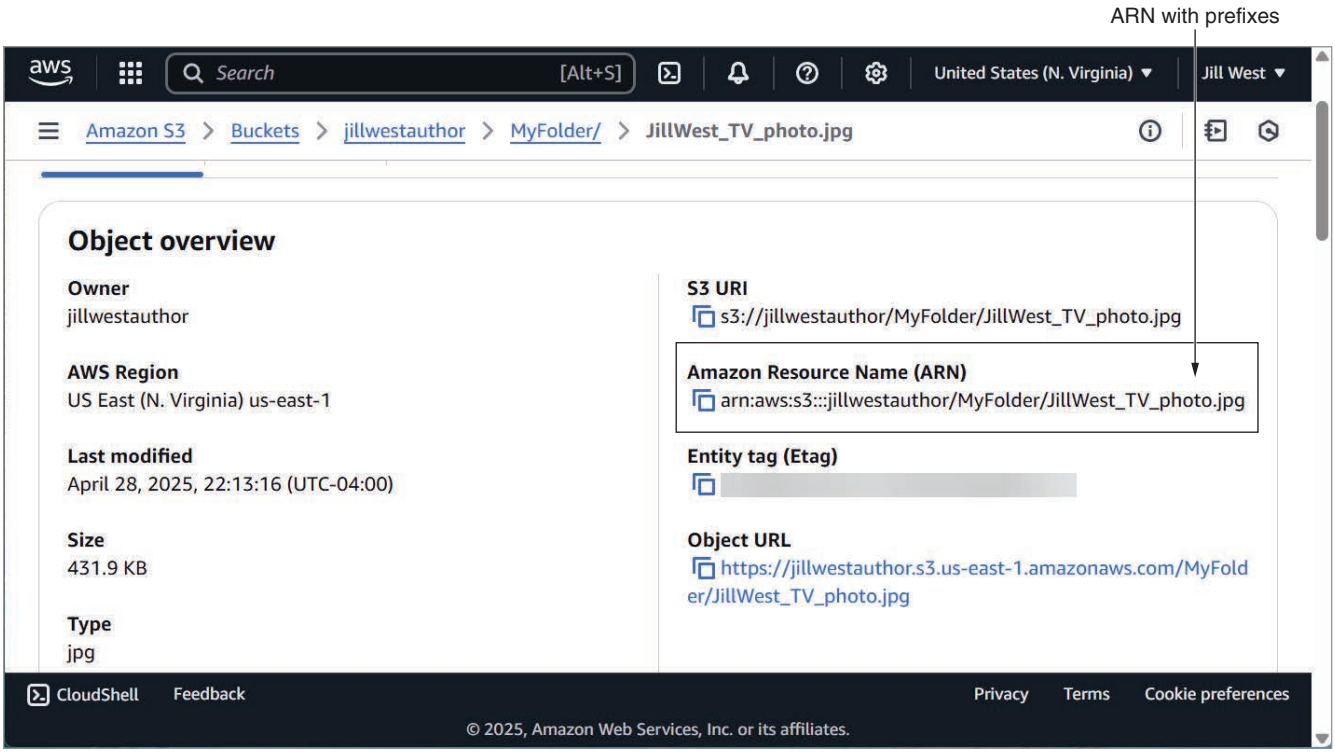
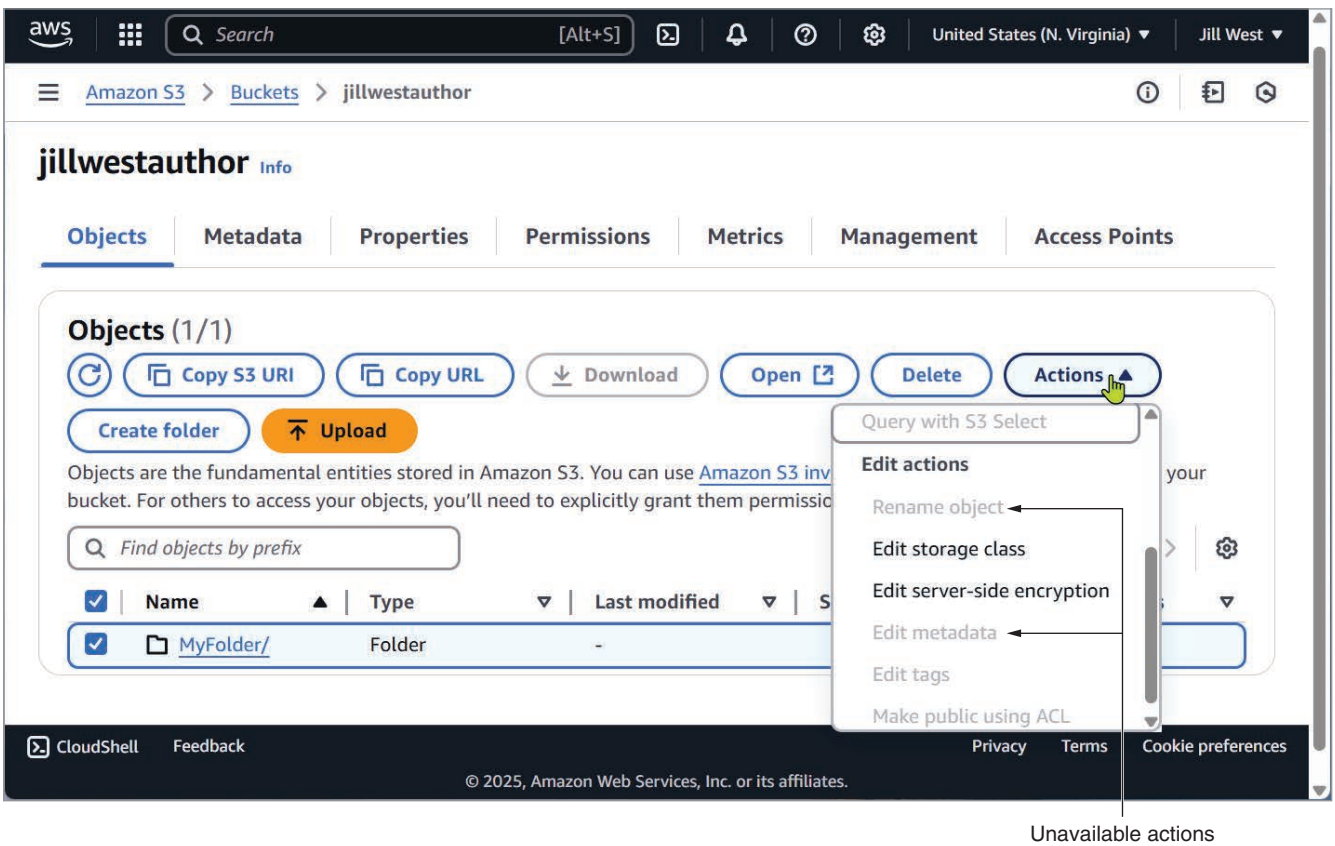


Figure 8-6 Simulated folders in S3 can't be renamed



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Choosing Between File, Block, and Object Storage

Whereas block storage in a SAN is the Cadillac of on-prem storage options, object storage maximizes the potential of storage capabilities in the cloud. To be fair, each storage type best supports different use cases, no matter where the storage is located. For example, object storage is generally less expensive and more flexible, while block storage provides faster performance at a higher price point.

Another difference between storage types is consistency of data. Data consistency refers to how well copies of data stored in a particular environment are consistent with each other. In block storage, the most current version of data is immediately available across all copies of the data after a change is made. This is called "strongly consistent." In object storage, however, more time is required for data modifications to propagate through the system and its various copies. This is said to be "eventually consistent." Data that is infrequently updated is better suited to object storage than is data that is changed often.

For this reason, object storage lends itself well to data that is written once and then left alone or that is only read, such as log files. This data type is referred to as WORM (write once, read many) and is similar in concept to the on-prem data center's CAS (content addressed storage), where content is written once and then referenced by its content rather than by its location on a disk. The data goes in and then it stays put and doesn't have to be changed. This kind of storage is highly scalable but not easy to plan for how to organize it (such as log files that will increase in size almost indefinitely). And it's not as well suited to more organized data, such as a database, unless it's only hosting backup or archived data.

Block storage works well for hosting databases, operating systems, and VM-attached storage where unlimited expansion of space is not needed. In these situations, data changes frequently and must provide low-latency performance. However, in the cloud, block storage is generally considered an older, less sophisticated method of configuring storage.

File system storage provides managed file system organization of data and can often be mounted to multiple VM instances, unlike block storage. Again, the focus with file system storage is on data organization rather than storage capacity.

Types of Databases

A database is a type of storage in the sense that massive amounts of data can be held in a database. Most of the larger cloud platforms position database services in a separate category, however, because databases are much more than simply data storage. In a database, data is structured in a way that makes it easier to pull specific pieces of information from the database, perform operations on that data, search for patterns in the data, and edit detailed portions of the data.

At a basic level, databases can be categorized as relational or nonrelational, and these are two very different types of databases. Relational databases incorporate a specific structure to the data, while nonrelational databases are more flexible in their design. Let's compare these database categories.

Relational Databases

A **relational database** is a type of database that must conform to a rigid structure consisting of tables that each holds different kinds of text-based information, which yields insight into relationships between those tables. To better explain the purpose of database tables, imagine you own a cupcake bakery. You might have one table for customer information, one table to list all orders placed, another table listing all the food products you create along with a list of ingredients, and another table listing your inventory of supplies you must keep in stock for your business to operate smoothly. Each entry in each table is called a record. For example, each customer listed in the Customers table is a different record. Each record must contain essentially the same kinds of information about each customer, such as name, address, phone number, and email address. A relational database defines meaningful connections between these tables, such as the relationship between which customers in the Customers table have placed which orders in the Orders table.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Using a relational database, you could determine which customers are most likely to visit the store on National Chocolate Day in October because they've ordered many chocolate cupcakes in the past. From that list, you can determine how much more chocolate you should order to be prepared after running an email campaign advertising to those customers. In other words, the relational database shows the relationship between customers who like chocolate and your chocolate supplies to help you determine how much extra chocolate you'll need in preparation for the holiday special.

A relational database requires sophisticated DBMS (database management system) software such as SQL Server or MySQL. SQL (Structured Query Language and pronounced "*see*-quel") is a programming language designed specifically for communication with databases. SQL standardizes the format of queries, which are requests to the database. Queries might be used to create new records, edit existing data, or list all data that matches specific parameters. Popular, cloud-based, relational database services include AWS's RDS (Relational Database Service) and Azure's SQL Database.

Nonrelational Databases

In many situations, the enforced design consistency of a relational database (with the same kinds of information required for every record in a table) is an advantage. However, this structure comes with the limitation that data must generally be represented by text or numbers rather than images, videos, or other file types. As the Internet—and particularly web applications—became more popular, this restriction on database design led to the emergence of more powerful database technologies better suited to managing less consistent data types. Nonrelational databases, also called NoSQL (not-only SQL) databases, resolve many of the weaknesses of relational databases. A **nonrelational database** is a type of database that can store a wide variety of data types using an unstructured, semi-structured, or structured design, allowing for flexibility, scalability, and high performance to adapt to changing needs. Popular, cloud-based, nonrelational database services include AWS's DynamoDB and Azure's Cosmos DB.

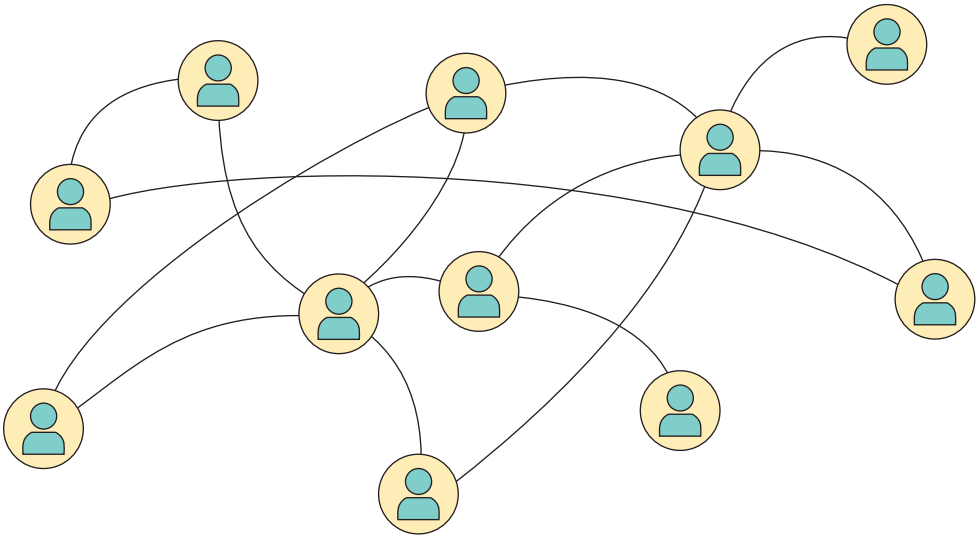
> **Note 8-2** | NoSQL originally stood for non-SQL, but more recently has been called not-only SQL because some of these systems do support SQL-based languages.

Nonrelational databases take a variety of approaches to store many kinds of data. Consider the following examples:

- **Key-value databases** (also called key-value stores) list any number of key-value pairs for each record. For example, you use key-value pairs to create tags for cloud resources, such as the key "Name" and the value "Web server 1." You can also create many other tags to track your resources, such as identifying resources by department, project, geographical location, or billing account. Similarly, in a key-value database of students, you might add key-value pairs for some students to list their hobbies, for some students to list their interests, and for some students to list their career plans. But you don't have to include all pieces of information for every student, only for those where it makes sense. And if you come across a situation where you need to store a unique piece of data about only one student (such as one student's unusual funding source for school expenses), you could add this information without needing to fill in similar information for all other students.
- **Graph databases** store data according to each entity's relationship with other entities. For example, consider the relationships between accounts on a social media platform where each account, or entity, is connected with several other accounts (Figure 8-7). A graph database displays the connections between these accounts, revealing patterns in the relationships. The graph database can also record details, called properties, about each account and each relationship. Graph databases can provide penetrating insights into evolving data. For example, credit card companies often use graph databases to detect signs of fraudulent charges. Similarly, e-commerce websites can use graph databases to make product recommendations to customers.

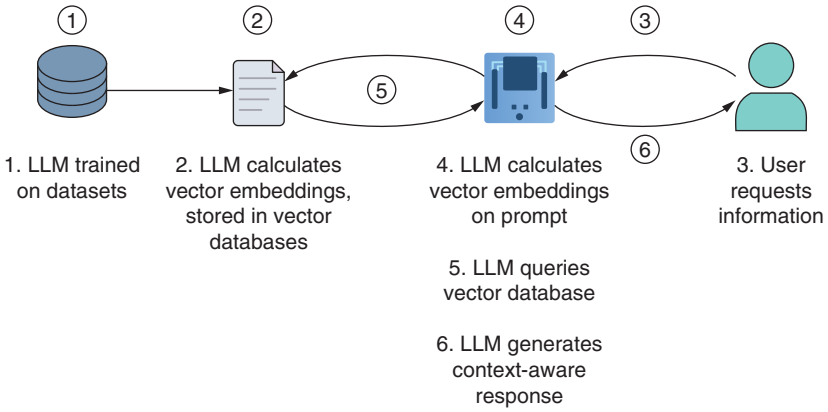
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-7 Connections between social media accounts in a graph database



- **Vector databases** store collections of values, called vector embeddings, in a way that can be searched quickly for values that are similar. Consider how AI chat tools, such as ChatGPT, use LLMs (large language models) to generate text related to your prompt when you've requested information from specific sources. LLMs can use RAG (retrieval-augmented generation) technology to ensure the LLM outputs reliable, documented information rather than just making up text. For example, consider the following scenario (Figure 8-8).

Figure 8-8 Role of vector database in RAG



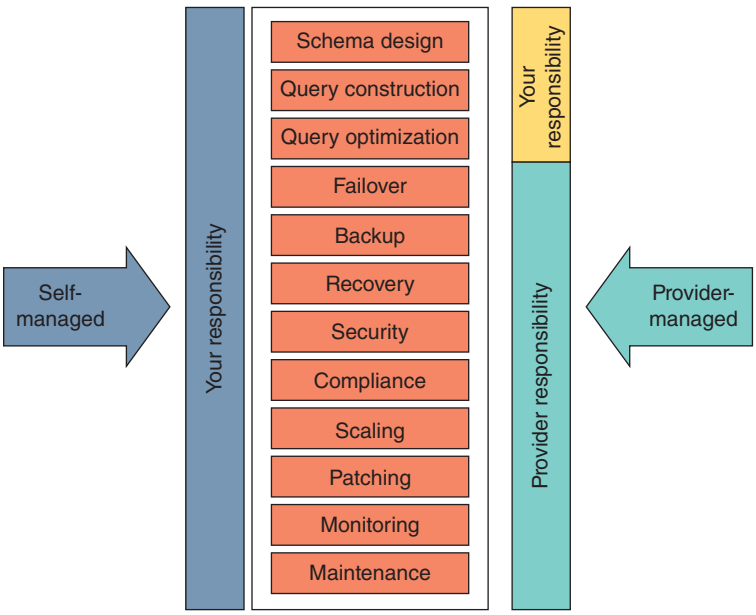
1. Suppose your employer has obtained a pretrained, general-purpose LLM to support internal users.
2. The company equipped the AI chatbot with access to company policy information. The LLM calculates vector embeddings for the data contained in this documentation and stores these vectors in a vector database.
3. Perhaps you ask the chatbot how many paid vacation days you can accrue during the next year, given your job role and the number of years you've worked for the company.
4. The chatbot must reference the connected documentation to answer your question accurately. How does it find this information among the possibly hundreds or thousands of pages of policy documents? When you submit a prompt requesting this specific detail, the LLM calculates vector embeddings for the data in your prompt (such as the words in the prompt).
5. It then searches its vector database for embeddings from the RAG-sourced documentation that have similar vector embeddings.
6. The LLM uses these similarities to then generate text in its output to you that is, theoretically, directly related to what you requested.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Database Management Types

Some cloud database services are specialized for relational databases while others are optimized for nonrelational databases. However, there's another method of categorizing database services. Ask the question, Who is in charge of managing the underlying infrastructure that hosts the database components? The answer to this question is not a clearly defined "this" or "that" situation. Instead, the division of responsibility for database management exists on a spectrum with a variety of available options (Figure 8-9), as described next.

Figure 8-9 Shifting responsibility to the cloud provider



- **Self-managed**—Self-managed databases are exactly what they sound like: a database where the customer manages the database and the server and other infrastructure hosting it. Typically, the host is a virtual server configured and managed by the cloud customer. For example, you can create an EC2 instance and install a DBMS such as MySQL on the instance. In terms of cloud service models, this is an IaaS (Infrastructure as a Service) design.
- **Fully managed or provider-managed**—With a managed database service, someone else manages the servers hosting the database. As the customer, you can still make some choices as to what types of servers host the database, how much capacity and bandwidth these servers have, and any limits you might want to place on the number of servers in the host cluster to control costs or ensure performance. In some cases, such as with AWS's RDS, you can even choose serverless hosting. However, the provider manages failover configurations, backup and recovery, security, patching, monitoring, and maintenance of any host servers. In terms of cloud service models, this is still an IaaS design.
- **Serverless**—With a serverless database, the customer does not manage the servers at all. In some cases, the customer can choose the type or size of servers in relation to performance and cost, but the customer is not responsible for updating, patching, scaling, or otherwise maintaining the servers hosting the customer's database. The host cluster scales automatically, and for very large deployments, this option is often more responsive to large fluctuations in demand. In terms of cloud service models, this is a PaaS (Platform as a Service) design. In AWS, two popular serverless database services are DynamoDB and Amazon Aurora Serverless.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

1. What type of storage is the hard drive in a laptop?
   a. SAN
   b. LUN
   c. DAS
   d. NAS
2. Which type of cloud storage would be the best fit to host a relational database?
   a. Object
   b. Block
   c. Cloud-native
   d. File system
3. Which database type supports RAG?
   a. Key-value databases
   b. Graph databases
   c. Relational databases
   d. Vector databases

Check your answers at the end of this chapter.

Section 8-2: Storage Configuration

Certification

1.4: Compare and contrast storage resources and technologies.

1.10: Compare and contrast methods for optimizing workloads using cloud resources.

2.3: Summarize aspects of cloud migration.

2.5: Given a set of requirements, provision the appropriate cloud resources.

4.2: Compare and contrast aspects of compliance and regulation.

Stored data progresses through a series of states in its lifetime. For example, recent data might need to be accessed frequently while older data is rarely if ever accessed. With some data, you need to keep older versions in addition to current versions. Sometimes data is placed under legal protection and must be saved for a period. And other times data must be destroyed according to legal guidelines. To optimize storage costs and reduce the administrative load of managing this data, configure storage services to automatically transition data through these various stages.

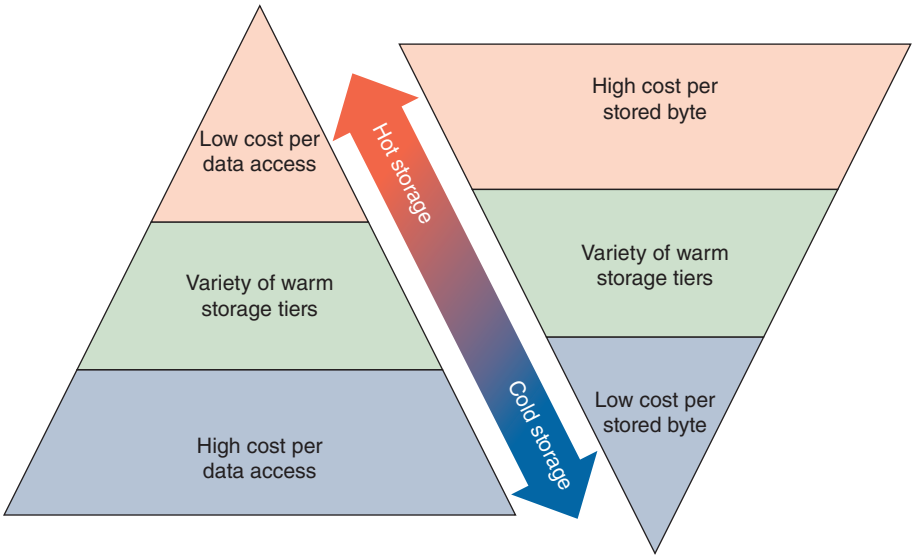
Tiering

Storage tiers, also called storage classes, can decrease the cost of storing data in exchange for increasing the cost and time of retrieving it. For example, data that is accessed frequently can be stored using a cloud service that charges more for the space used and less for the number of times you access it (Figure 8-10). This allows for low-cost, frequent access while paying a premium for the storage space itself. Because the data is expected to be accessed frequently, this top tier is often referred to as hot storage.

In contrast, data that doesn't require regular access, such as archived data or older backups that need to be kept long term, can be held in cold storage where storage space is cheaper but each request for data costs more (and takes longer). This system replicates the on-prem process of transferring old data from active systems, such as fast flash storage, to cheaper but less accessible storage media, such as tapes. The coldest—also the slowest—of these tiers, such as S3 Glacier in AWS, is designed only for archive storage and can take up to 12 hours to retrieve data after a request is made.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-10 With tiered storage, the cost of accessing data is inversely related to the cost of storing data



In between these two extremes is warm storage. At these medium tiers, data is cheaper to store than at the hot storage tier, and it's cheaper and faster to retrieve than at the cold storage tier. For example, you might store your most recent backups in AWS's Standard tier for the first 7 days, and then transfer backups to an Infrequent Access tier for the next three months. Older backups could be saved in budget-friendly S3 Glacier Archive for years.

Life Cycle Rules

Adaptive optimization of storage costs can shift data from one tier to another based on criteria the customer defines. These adaptive optimization settings are called **life cycle rules** or life cycle management. For example, in AWS's S3, you can set life cycle rules that apply to one or more objects stored in a bucket. Available life cycle rules can move current or previous versions of objects between storage classes (such as frequent access, infrequent access, archive, or deep archive) a certain number of days after the object was created or last accessed (Figure 8-11).

While not all data justifies paying for storing its older versions, versioning is an option where needed. **Versioning** refers to keeping older versions of the data as a backup or to document changes in the data, such as for compliance or regulatory purposes. Keeping older versions of files can result in significantly increased storage costs as well as increased security issues if sensitive data is held within older versions that might be moved to less secure storage configurations. However, there are times when versioning is essential, especially since the ability to restore older file versions can protect data from accidental deletion or file corruption.

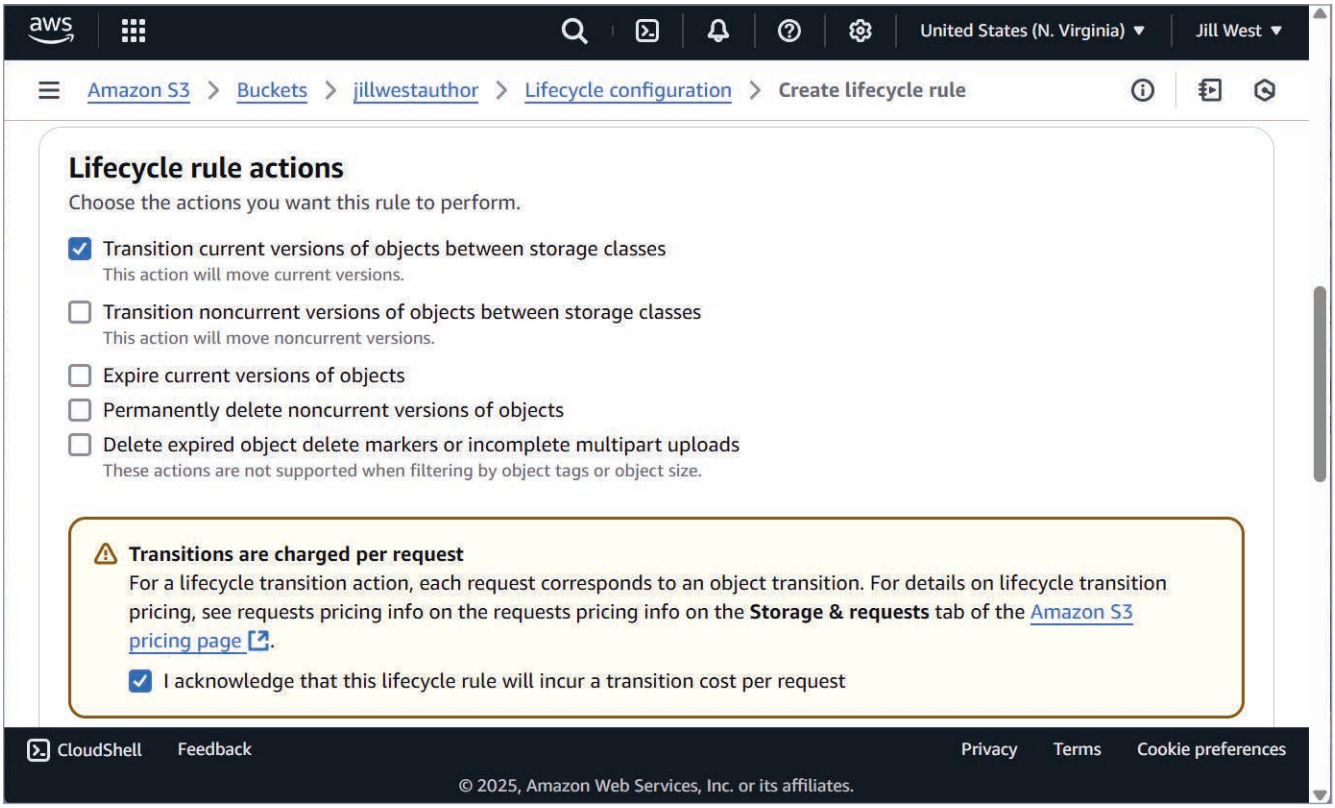
Storage Performance

Clearly, storage options can be compared with each other based on several factors. For example, the organization of data, expandability of the storage space, and required durability of the data all factor into storage decisions. Storage performance can be evaluated based on several measurable factors, three of which are described here:

- **Transfer rate or read/write throughput**—This metric specifies the speed at which data is moved through the storage infrastructure and is measured in MB/s (megabytes per second).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-11 AWS life cycle rule configuration



Source: Amazon Web Services, Inc.

- **IOPS (input/output operations per second)**—This number details how many read and write operations can be completed per second. It's typically given as an integer, such as 500 for an HDD (hard disk drive) in the cloud or 16,000 for a cloud-hosted SSD. In some cases, you might request IOPS-optimized storage media to improve database operations. For example, AWS offers Provisioned IOPS storage that allows for more predictable database performance.
- **Latency**—Latency results from bottlenecks in technology, geographical distances, or subpar optimization of data. It can be measured by TTFB (time to first byte), RTT (round-trip time), or similar metrics. You can decrease the latency of a storage solution by turning on data caching and by carefully choosing where data is located geographically. CDNs (content delivery networks), such as AWS CloudFront, are designed specifically to reduce latency when accessing stored data across large distances. Recall that CDN is a type of edge computing that temporarily caches data closer to its points of use.

Storage Migration

Shifting data storage to the cloud is a popular migration type among today's cloud users. You might have performed a similar process if you've shifted your photos, files, or contacts from a local hard drive or storage device (such as a smartphone's local storage) to cloud-based storage, such as Google Drive, Microsoft OneDrive, or Apple's iCloud. With a batch of photos, the process is relatively simple—copy the photo files from one location to the other. You might also perform some basic reorganization in the process. In a corporate environment, however, the process is more complicated, although the basic idea is the same: move data from one place to the other.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Cloud providers offer a variety of services to assist in this process. Many of these services rely on network connections, such as the Internet, for online data transfer. You learned about these connection types in earlier chapters. Services such as AWS DataSync can manage the online transfer of massive volumes of data over long periods when there is a potential for errors or other delays that can be managed automatically by the service. In contrast, some cloud providers offer offline data transfer services.

An offline data transfer consists of loading encrypted data onto a storage appliance and then physically shipping that appliance to the CSP, either by a common package-delivery service or via a dedicated transport. Upon arrival, the CSP uploads the data to the cloud, where the client can configure the data as needed. These appliances are appropriate for terabyte- or petabyte-scale data migrations. Examples include AWS's Snowball, Azure's Data Box, and Google's Transfer Appliance.

Database Migration

Migration requirements and processes for relational and nonrelational databases depend significantly on the combination of source and target database services. For example, migrating from an on-prem MySQL server to a cloud-based MySQL server is relatively straightforward. Converting from an on-prem relational database to a cloud-based nonrelational database is much more complex. Some of these migration types can be performed live with no downtime for the applications and other services that depend on these databases. Often, there will be some—perhaps a lot—of downtime. Many cloud providers offer database migration services that can ease the transition and reduce or eliminate downtime.

Data Destruction

When handling physical documents and devices on-prem, companies will often hire professional disposal services that thoroughly destroy hard copies of documents (such as by shredding) and adequately sanitize or destroy storage media or devices so no data can be recovered. Reputable vendors provide end-to-end insurance coverage from the moment they take possession of a device to be destroyed.

Destroying cloud-hosted data can present additional complications, however. Many customers also use services that duplicate data to personal devices or to on-prem storage for convenience, which is called hybrid cloud storage. Cloud providers also will often make redundant copies of data to ensure its high durability. Even the providers themselves might not know where all these so-called zombie copies of data reside within their platforms, and these copies can continue to exist undetected for years. With replicated and backed-up data stored across cloud providers' data centers, how can a customer ensure their data is destroyed in a manner that meets regulatory requirements?

The following options can help increase the security of cloud-based data disposal:

- **Use restrictions**—Limit the number of copies made of protected data, especially where extra copies are only intended to increase convenience.
- **Tenancy**—Limit data storage to a single host device dedicated to a single customer or even owned by that customer.
- **Encryption**—Encrypt all records stored in the cloud. At least for now, encryption can protect data even if it's recovered from a backup storage location (no one knows whether encryption will continue to be reliable once quantum computing emerges into mainstream markets). Whereas this process presents other difficulties, such as key management when some of the data must be destroyed but not all of it, encryption ensures the data is irrecoverable once the private key is destroyed.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

4. Which of these storage tiers would be most appropriate for archived data?
   a. Glacier
   b. Frequent Access
   c. Infrequent Access
   d. Standard
5. Which of the following metrics should be increased to improve storage performance?
   a. TTFB
   b. IOPS
   c. SSD
   d. RTT

☐ Check your answers at the end of this chapter.

Section 8-3: Cloud Storage and Database Services

Certification

1.6: Compare and contrast containerization concepts.

2.5: Given a set of requirements, provision the appropriate cloud resources.

3.4: Given a scenario, manage the life cycle of cloud resources.

Each of the major public cloud providers offers several options for meeting data storage needs. Most of these services can be categorized according to file, block, or object storage. Some services are targeted specifically to databases or VM instances or other specific use cases. Some services are also designed for cost optimization by reducing the prices on colder storage tiers while allowing increased data retrieval latency.

Let's explore some of the primary file, block, and object storage services offered by AWS, Azure, and GCP. As you read about these services, also notice different options for creating backups, which you learn more about later in this chapter.

AWS Storage Services

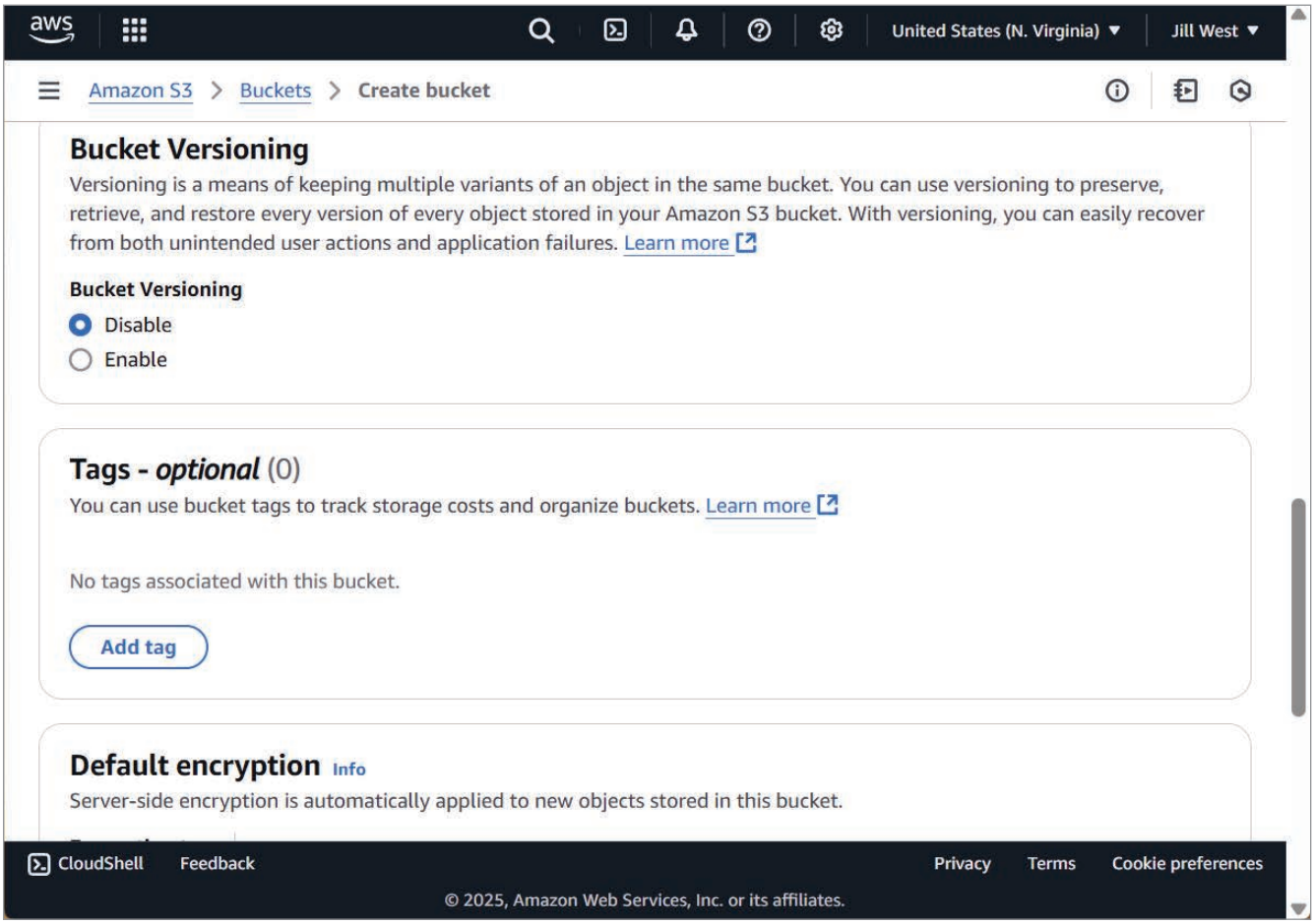
AWS continues to release new storage services to meet a variety of needs, including file system, block, object, and database storage options. AWS also offers a Backup service, which provides fully automated backup processes across multiple services—including block storage, file system storage, and database services—and can also back up on-prem data. Another popular service is AWS Storage Gateway, which can be used to bridge the gap between on-prem and AWS storage, even offering a VTL (virtual tape library) feature called Tape Gateway that presents virtual tape drives to an on-prem backup application. The following is a description of some of AWS's other popular storage services.

Object Storage

S3 is AWS's most popular and flexible cloud storage service, able to host just about any kind of data. AWS offers 11 nines durability (that's 99.999999999% guarantee of not losing a stored object) for data contained in S3 when stored across multiple AZs (availability zones). Data can be encrypted by the cloud customer (called client-side encryption) or by AWS (called server-side encryption). Figure 8-12 shows some of the options available when creating a new bucket in S3. A storage bucket is essentially a container for files, logs, static websites, or other data objects.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-12 Choose features such as versioning, logging, tags, and encryption



Source: Amazon Web Services, Inc.

Note 8-3 | By default, an S3 bucket's security configurations prevent all public access from the Internet. So why do so many security incidents occur due to S3 bucket breaches? Some of the most common causes of S3 bucket vulnerabilities include the following:

- Human error in opening bucket access too far or leaving access open after initial testing
- Misconfigured policies that allow unintended access
- Lack of encryption on stored data

There's no minimum fee to use S3, and costs are calculated based on actual storage space used, data requests (such as retrieving, scanning, or deleting data), and data transfers. S3 classes provide different storage tiers and rates to better meet customers' needs. For example, the Standard class allows for high durability, availability, and performance of data that is accessed frequently. However, data that does not need to be accessed often could be stored in the cheaper Infrequent Access class, and data that is archived for long-term storage could use the even cheaper Glacier Deep Archive storage class. This archival service is ideal for highly regulated data that must be stored for many years but is rarely accessed, such as financial or medical records. Retrieval for the coldest storage type might take up to 12 hours.

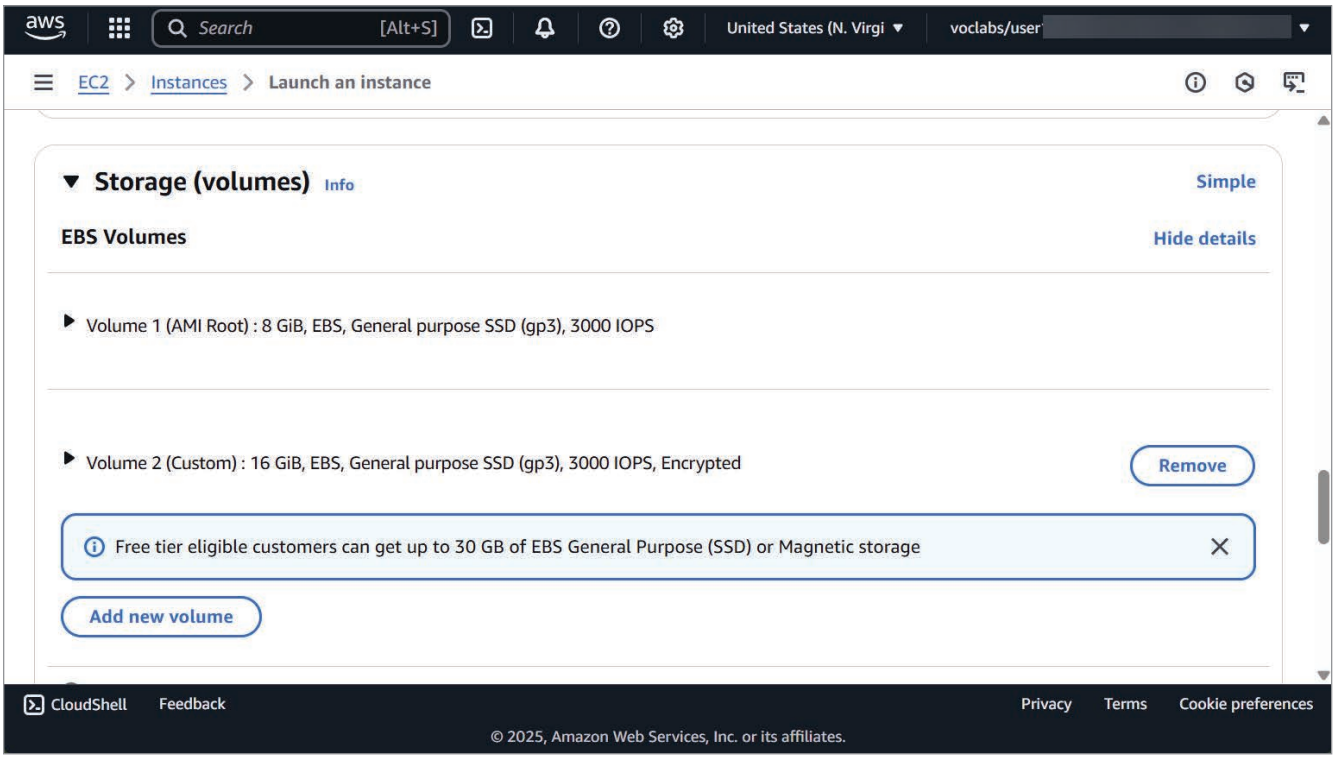
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

AI in IT | AI and ML (machine learning) services can detect stored data that is not properly secured. For example, AWS's Macie uses ML to identify sensitive data, such as Social Security numbers or credit card numbers, stored in S3 buckets that are publicly accessible or are not encrypted. Macie provides ongoing evaluation of S3-stored data. The service continuously reports its findings to guide security investigations and support compliance efforts.

VM-Attached Storage

AWS EBS is designed to work with EC2 to provide instance-level storage volumes. An EBS volume exists as a separate resource from an EC2 instance, similar to how you can remove or attach a hard drive in a physical computer. As you've already learned, this is called persistent storage because it doesn't necessarily cease to exist when you terminate the EC2 instance. In contrast, EC2 Instance Storage provides a temporary drive called an ephemeral drive that exists only while the EC2 instance exists. Ephemeral instance storage is not available for all instance types. For example, the t2.micro instance type uses only EBS storage. Figure 8-13 shows a t2.micro instance being created with its default root volume plus a new EBS volume, both of which are persistent volumes.

Figure 8-13 This instance will have two attached volumes



Source: Amazon Web Services, Inc.

Containers

You have similar options when configuring storage for your containers. Because containers are intended to exist only for short periods, their data typically exists only briefly as well. A container functions in layers defined within its image file, and most of these layers are read-only. The top layer is writable, meaning the container can create, delete, or modify files stored in that layer. But once the container is terminated, its writable layer also disappears.

To persist data beyond the life cycle of the container, that data must be stored somewhere other than the container's own writable layer. In AWS, this can be accomplished through the typical data storage services,

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

such as an EFS (Elastic File System) resource or an EBS volume. For example, you can mount an EBS volume to up to 16 containers at a time. Alternatively, you can mount an EFS file system to thousands of containers so that all their data is collected into a single location. This way, all containers in a cluster can access the same data store, and you don't have to sync data across storage volumes. The file system continues to be available to containers as they're created, run, and terminated.

File System Storage

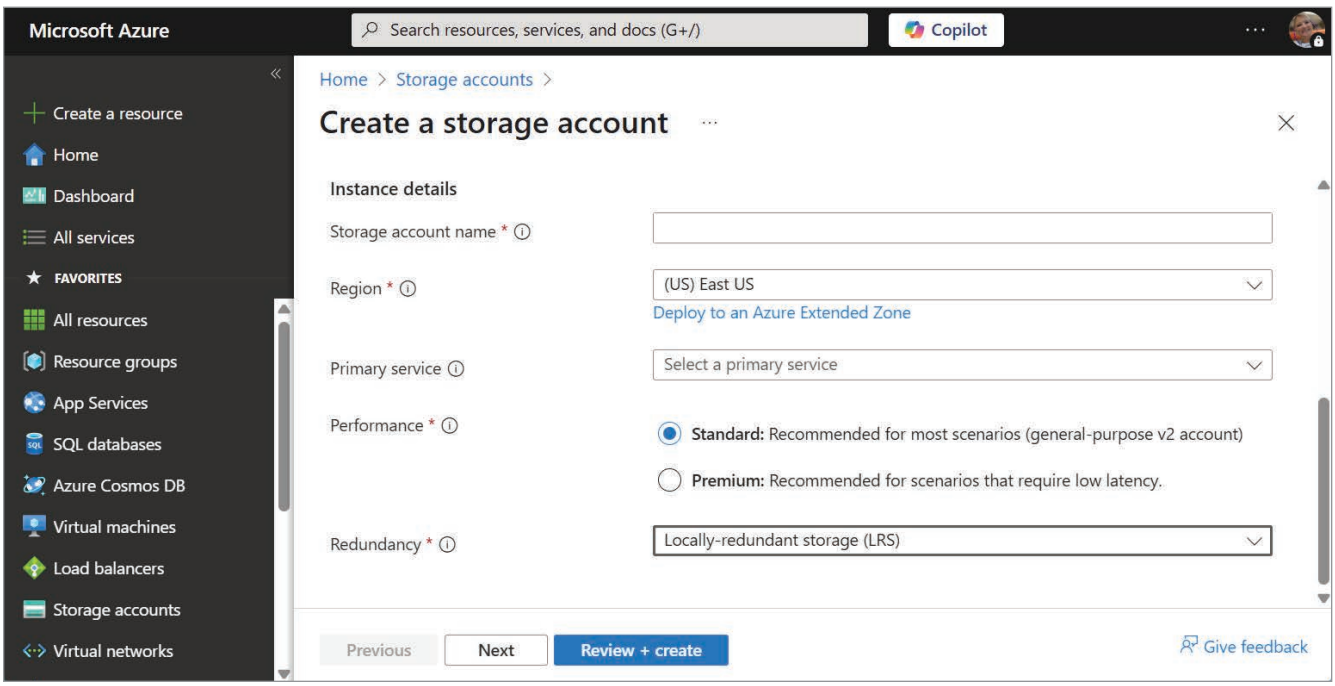
AWS's EFS and FSx (third-party file system) both provide managed file system storage. EFS supports Linux-based workloads while FSx for Windows File Server provides a native Windows file system built on Windows Server. These file systems can be mounted to many cloud-hosted VM instances or on-prem servers using NFS. This capability makes it easy to lift and shift applications that require file system access and to share files across a hybrid cloud network. The file system space can also be expanded as needed rather than being limited by the size of the physical disks hosting the file system.

As with on-prem file systems, access to files can be managed through file system permission policies that grant read, write, or root user access to files. With file system policies, you can also enforce secure transport when moving files or allow access by approved accounts only when using approved endpoints. File system policies are configured using JSON (JavaScript Object Notation). In a project at the end of this chapter, you practice editing JSON policies to manage access to stored data.

Azure Storage Services

Azure also offers object, file system, and VM-attached block storage. To use most of Azure's storage services, you must first create a storage account. This storage account contains all your Azure Storage data, including blobs, file shares, and disks. The storage account also defines a unique namespace in the Azure cloud, meaning that you must choose a name for your storage account that is not already being used by anyone else in Azure. You can associate a storage account with an existing subscription and even include it within a specific resource group for easy management. Figure 8-14 shows some of the options you have when creating a storage account. In a project at the end of this chapter, you use the redundancy option LRS (locally redundant storage) to conform to the free-tier requirements in Azure.

Figure 8-14 Create a storage account



Source: Microsoft Azure

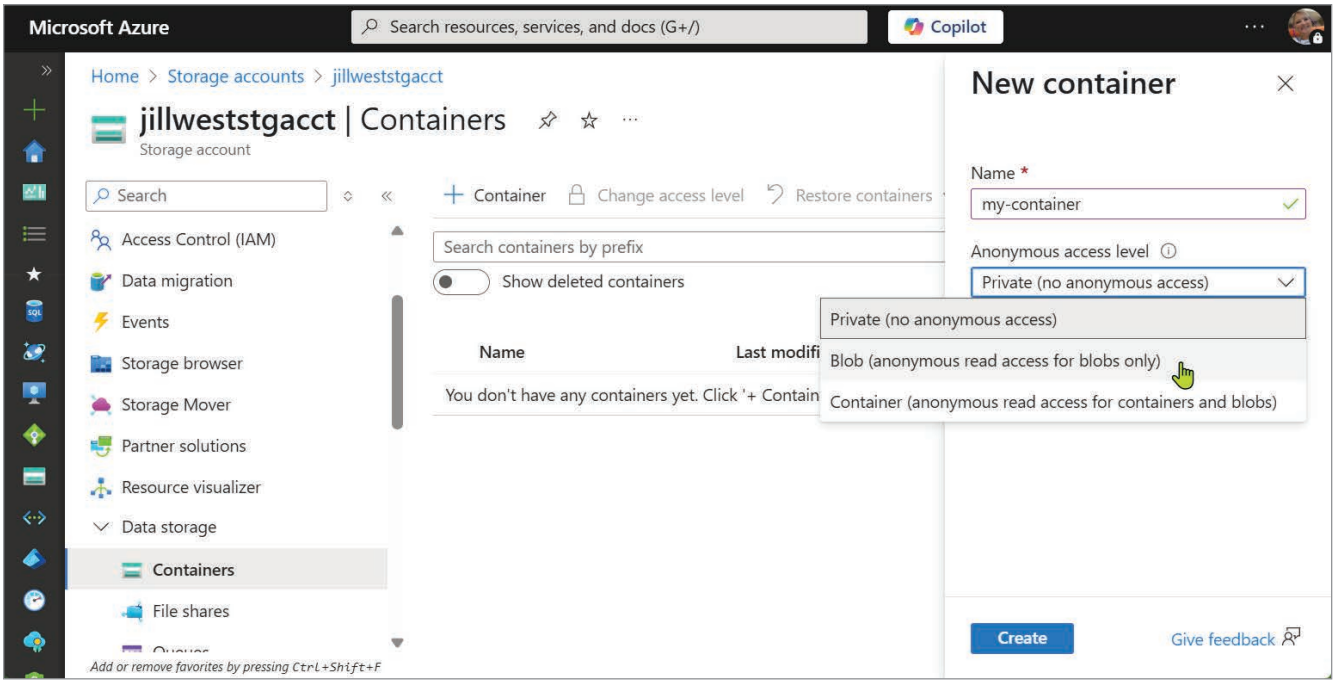
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Block and Object Storage

When managing a database, there are times when large files, such as photos or audio files, must be included with other, more structured data. Over time, these kinds of unstructured, opaque files came to be known as BLOBs (binary large objects). Microsoft borrowed this concept in naming its block and object storage service, which is called Azure Blob storage. Like other object storage services, Azure Blob supports storage of massive amounts of unstructured data. At the same time, Azure Blob includes options for block storage that supports VHDs (virtual hard drives) for VM-attached storage volumes.

In Azure Blob, you create storage containers from within an existing storage account (Figure 8-15). You can choose public access at the blob level or at the container level as you create the container. In Azure, storage containers are analogous to AWS S3's buckets, and blobs are sort of like S3 objects. You create a container, and then you upload blobs into it. However, some blob types offer object storage, and some blob types actually offer block storage. Additionally, the names of these blob types don't match the storage types they're using underneath the surface. Let's discuss this anomaly further.

Figure 8-15 Create a container for a blob, and choose anonymous read access level



Source: Microsoft Azure

There are three blob types in Azure:

- **Block blob**—Block blobs are designed for binary data, such as text, photos, or other binary files. Despite their name, block blobs are storage objects, similar in design to S3 storage objects. A block blob is made up of blocks, which are pieces of a file that are stored together. But each file is still a separate blob, regardless of the number of blocks used to store the blob, and each blob holds a separate file. Block blob storage is considered to be an object storage service. Block blobs can be any size up to about 190 TiB. Each block in a block blob can be up to 4,000 MiB.
- **Append blob**—Append blobs are similar to block blobs in that each uploaded file becomes a separate blob that is made up of blocks. However, append blobs are optimized for adding more blocks to the blob as new data is added, or appended, to the file, such as what happens with a log file as new log entries are generated and stored in the file. This is different than block blobs where uploading new data to a block blob replaces the existing file instead of adding to it. Still, append blobs are also considered to be object storage and can be any size up to about 195 GiB. Each block in an append blob can be up to 4 MiB.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Page blob**—Page blobs support frequent read/write operations, such as what is needed by disks attached to VMs or by persistent storage for databases. Page blobs, then, function as block storage, similar in design to EBS storage volumes. A page blob is made up of 512-byte pages and can be as large as 8 TiB.

VM-Attached Storage

Azure Disk Storage offers managed disks for VM-attached storage, like a hard disk attached to a physical computer. An Azure disk can provide an organized file system to the VM but can only be attached to one VM at a time. An Azure disk is a VHD stored as a page blob in Azure Blob Storage.

File System Storage

Azure Files provides cloud-hosted file share services. File shares can be mounted from Azure to Windows, Linux, or macOS systems either in the cloud or on-prem, acting as a file server or NAS device on your network. File Share can also support applications that have been shifted to the cloud from the on-prem data center, or hybrid applications that run on-prem while their data resides in the cloud.

GCP Storage Services

GCP, like the other platforms, offers object storage as its primary storage solution, file system storage to support apps and file sharing, and VM-attached storage to provide persistent storage to VM and container instances.

Object Storage

Cloud Storage is GCP's object storage solution. It offers 99.999999999% (that's 11 nines) annual durability and TTFB (time to first byte) measured in milliseconds. To use Cloud Storage, you create a bucket and then upload objects to the bucket, similar to how AWS's S3 works. Cloud Storage supports popular features such as life cycle management and object versioning; however, some of these features are only available through the CLI (command line interface). Bucket-level permissions are set using GCP's IAM service, and object-level permissions are set using an ACL (access control list). In a project at the end of this chapter, you practice editing an object's ACL to manage access to that object.

GCP's object storage is categorized by four storage classes (Figure 8-16), described next.

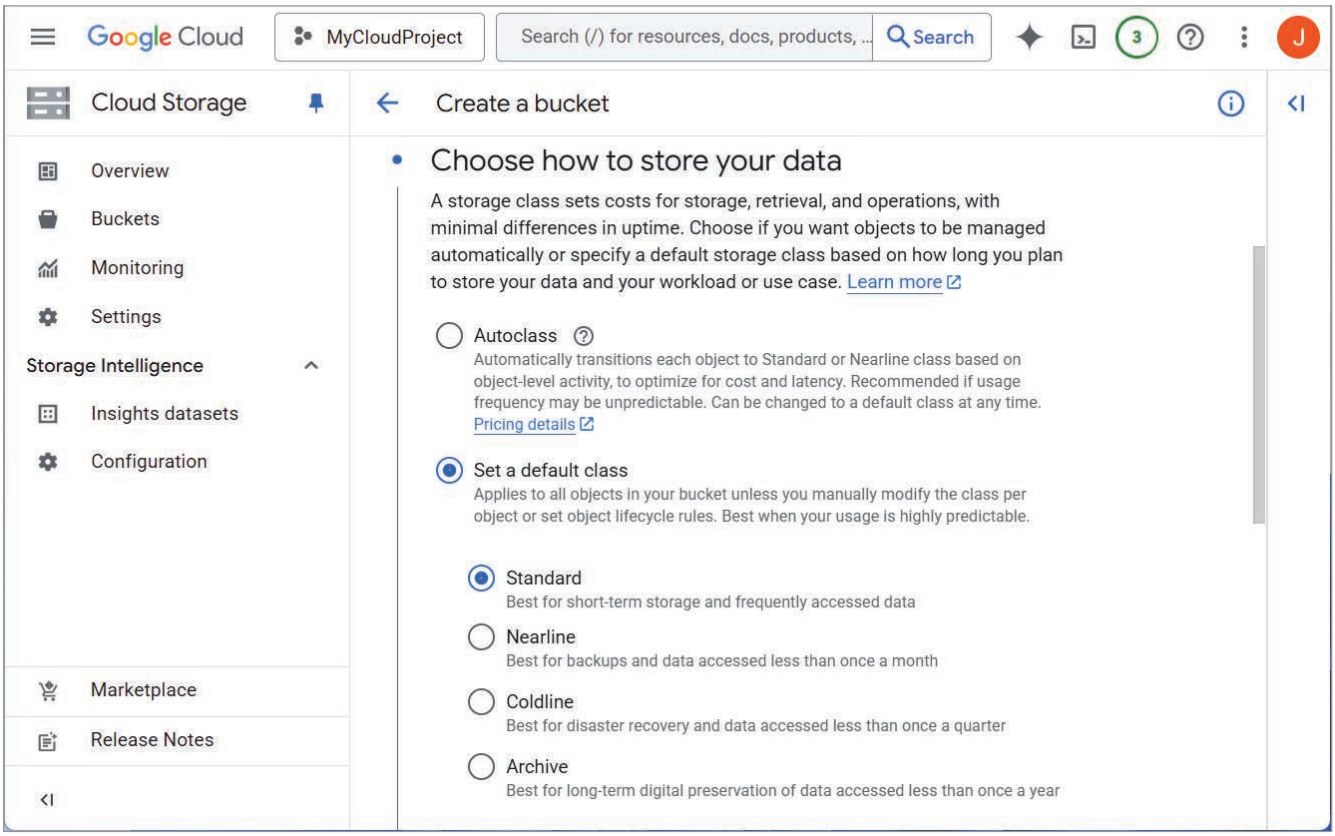
- **Standard storage** provides a balance of inexpensive storage for data that needs to be accessed frequently without requiring a minimum storage duration.
- **Nearline storage** is designed for moderate frequency access, such as less than once a month, and requires storage duration of at least 30 days.
- **Coldline storage** is designed for data that is accessed less than once or twice a year and requires storage duration of at least 90 days.
- **Archive storage** is optimized for long-term storage that is rarely accessed and requires a minimum storage duration of at least 365 days.

VM-Attached Storage

Persistent Disk and Hyperdisk offer high-performance block storage for attaching virtual SSDs or HDDs to Compute Engine or Kubernetes Engine instances. Persistent Disk is a first-generation block storage service for existing VM families while Hyperdisk is a new generation of block storage designed for high scalability and performance. You can resize, back up, or snapshot these disks, and you can even mount a disk to multiple readers. This means several VMs can read from the same disk or disk cluster, which enables easy content sharing among VMs.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-16 GCP's free trial includes limited Standard-class storage



Source: Google LLC

File System Storage

Cloud Filestore is a managed file storage service designed to support cloud-hosted applications that rely on a shared file system, similar to adding a NAS to a network. Filestore file shares can attach to Compute Engine instances, Kubernetes Engine instances, on-prem computers, and some types of external clients. Use IAM permissions to control access to Filestore operations, such as creating a Filestore instance. Use Linux permissions to control access to the file share itself, such as read, write, and execute. Filestore supports any NFSv3 client.

Applying Concepts 8-1: Research Database Services

Estimated time: 15 minutes
Objective 1.9: Explain the importance of database concepts.
Resource:

- Internet access

Description: Databases play a critical role in managing and using data in applications and websites. Think about how many times each day or each week you interact with databases: when you sign into your email account, when you post updates to social media, when you withdraw money from your bank account, when you pay for purchases at a store, and many, many more examples.

You might never be responsible for managing a database—DBAs (database administrators) require a great deal of training. However, as a cloud professional, you will likely need to support and secure database storage, or you might need to migrate databases between the cloud and an on-prem data center or between cloud services. To do this, it's important to understand the terminology involved and the business

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

context for these technologies. To learn more about database concepts, search online or use an AI chat tool (such as ChatGPT or Perplexity) to respond to the following questions:

1. What are three databases you've interacted with today? How do you know databases were involved?
2. Give an example of a database service in AWS, Azure, or GCP that supports relational databases and another service that supports nonrelational databases. What is an example of a business scenario where each of these two services would be a good fit?
3. Create a metaphor to explain to a nontechnical person what a data warehouse is and how it is different than a typical database. What is an example of a business scenario where a data warehouse would be needed?

Self-Check

6. AWS file system policies are configured using _____.
   - **a.** NFS
   - **b.** JSON
   - **c.** EFS
   - **d.** VTL
7. An object stored in Azure is called a _____.
   - **a.** storage account
   - **b.** container
   - **c.** bucket
   - **d.** blob
8. Which of the following is GCP's coldest storage tier?
   - **a.** Archive storage
   - **b.** Nearline storage
   - **c.** Standard storage
   - **d.** Coldline storage

☐ Check your answers at the end of this chapter.

Section 8-4: Creating and Storing Backups

Certification

1.4: Compare and contrast storage resources and technologies.

1.7: Compare and contrast virtualization concepts.

2.5: Given a set of requirements, provision the appropriate cloud resources.

3.3: Given a scenario, use appropriate backup and recovery methods.

4.2: Compare and contrast aspects of compliance and regulation.

As you've already learned, cloud platforms such as AWS, Azure, and GCP include built-in durability and availability features for many of their services, including keeping multiple copies of your data across hardware stacks to allow for their data recovery efforts should something go wrong. In this case, durability is something the CSP provides and manages.

That said, durability is not the same thing as a backup. A high durability rating reduces the likelihood you'll need a backup. However, critical business data still warrants backup protection. Whether data is lost through user error, such as deleting needed files or not paying a data storage bill, or through an attack or disaster, backups provide a layer of protection managed by the cloud consumer that is not covered by durability ratings.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Backup Strategies

Backup professionals suggest, at the minimum, following the 3-2-1 Rule for backups. This rule states, as illustrated in Figure 8-17, that you should have at least three copies of your data (that's one production copy and two backup copies), stored on at least two different kinds of media (such as disk and tape), with one copy stored in an alternate location geographically separate from the original data.

For an on-prem network, this means that storing one backup copy locally is acceptable; however, the other backup copy of the data should be stored in a remote location, preferably in a different geographic area where a natural disaster that affects the local network would not endanger the remote backup. For more sophisticated backup strategies that rely more natively on the cloud, consider the following variations of this rule: 3-1-2 (Figure 8-18), 3-2-1-1-0 (Figure 8-19), or 4-3-2 (Figure 8-20).

Figure 8-17 Three copies of data, stored on two different kinds of media, with one copy stored off-site

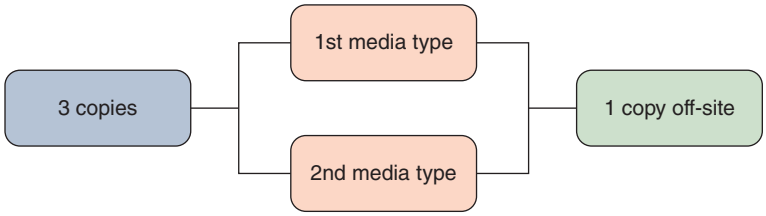


Figure 8-18 Three copies of data, stored on one kind of media (such as SSD), with two copies stored at geographically isolated locations in the cloud

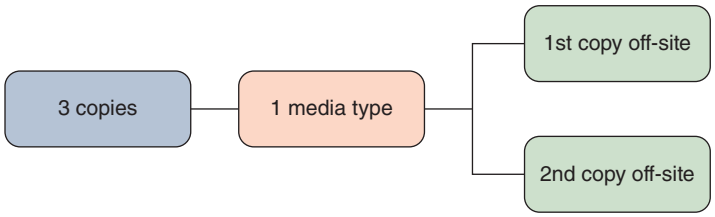


Figure 8-19 Three copies of data, stored on two kinds of media (such as SSD and tape), with one copy off-site and one offline (which protects from a malware attack), and zero errors when testing backups

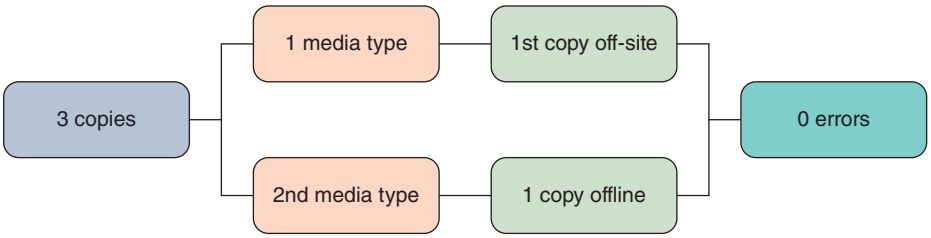
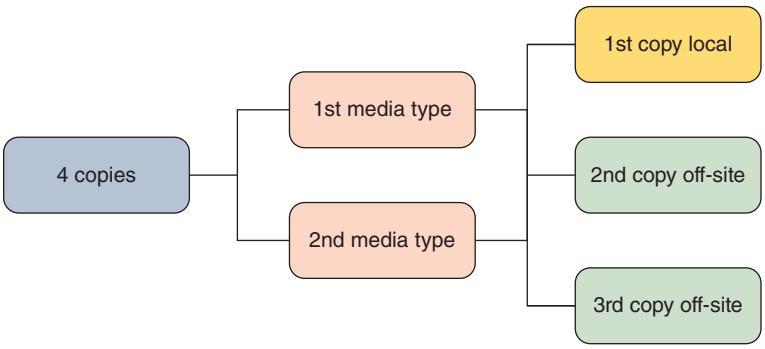


Figure 8-20 Four copies of data, stored at three geographically isolated locations, where two of those locations are off-site



It's also important to understand the difference between archived data and backup data. Archived data is cold data that is no longer used. It might be an earlier version of data that is in use, or it might be the primary copy of old data that is no longer accessed (such as medical records for former patients). Typically, archived data is stored in cheap cold storage that is not easily accessible. It's not synced to any active systems.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

While backup data is not regularly accessed, it must be frequently updated to match active data, and it should be quickly accessible should something bad happen to the active data store. Backups might include versioning, which keeps different versions of a file that has been changed multiple times.

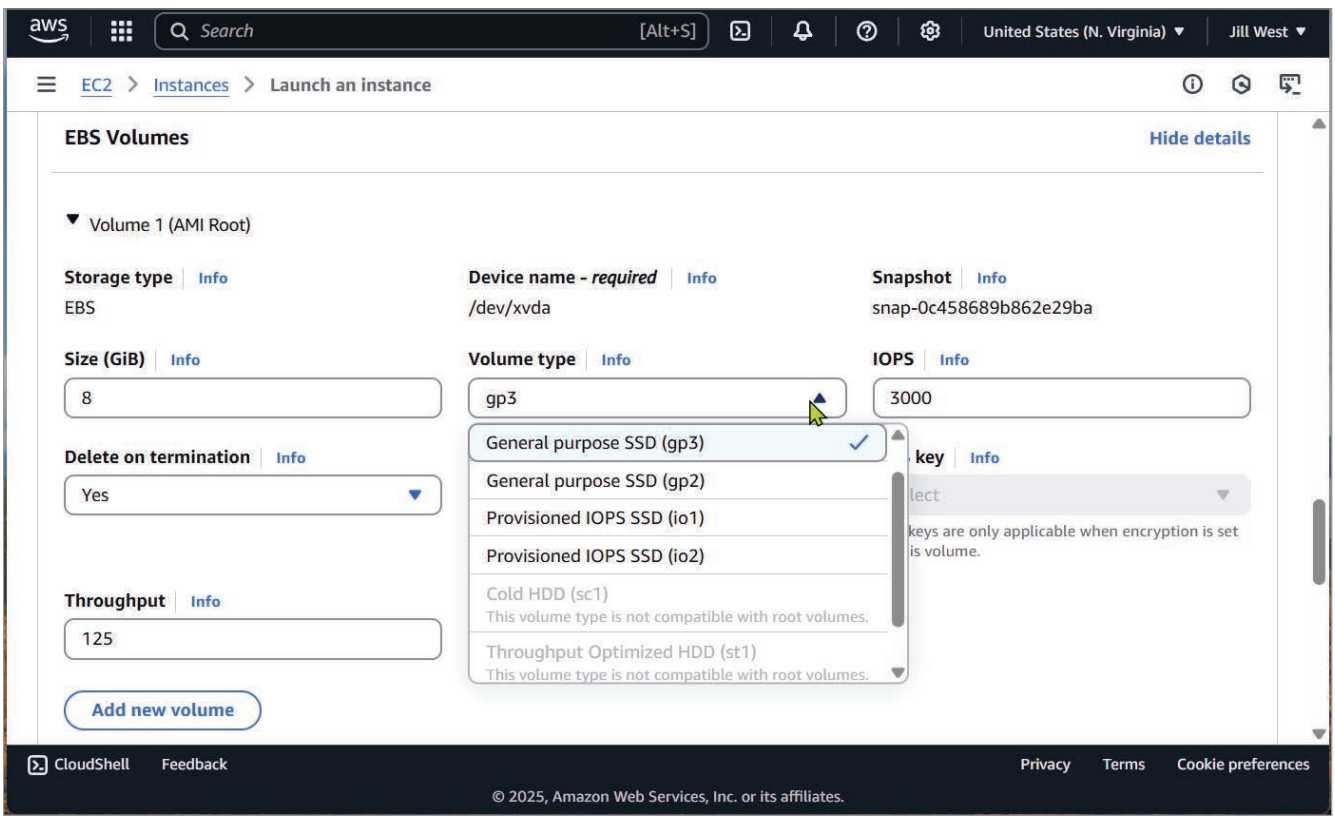
Backup Retention

Storing backups is a common use of storage space in the cloud. How long should you retain backup files? In some cases, backup retention length is dictated by regulatory standards. In other cases, you can choose what makes the most sense in your situation. Factors to consider include:

- **How extensively data changes from backup to backup**—At some point, older backup data is no longer relevant or useful.
- **How expensive backup storage is in comparison to overall budget**—Backup storage places demands on the IT budget that can be reduced by deleting older backups.
- **The existence of a legal hold**—A **legal hold**, also called a litigation hold, is a court-ordered notification requiring that identified data not be deleted. Cloud platforms, such as AWS, offer a legal hold feature that protects identified data or backups from deletion until the hold is removed, regardless of any life cycle rules that might apply to that data.

Deleting old backups can help reduce storage costs and, if using physical media for those backups, can increase available space for newer data. Using cheaper storage media, such as HDDs, can also reduce costs of backup storage. Many cloud platforms offer cheaper storage rates when using these slower storage devices. For example, Figure 8-21 shows the option to use SSDs or HDDs for the EBS volumes attached to an EC2 instance, although some options are not available for root volumes. Some automated tiering features can help optimize the transition of data from one tier to the next, taking advantage of this cheaper storage option for older backup files.

Figure 8-21 Storage volume options in EC2



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 8-4 | You can reduce data storage and backup costs by restricting the amount of data stored. For example, you might implement user quotas on file system storage to limit how much storage space any individual user can consume. As a user approaches their limit, they will receive a warning message. When they have reached their limit, they are no longer allowed to write new files to the storage space until they delete other files or obtain special permission to exceed the configured limits.

Many of the cloud services consumers configure and manage offer options for backups. Consider the following protection capabilities:

- **Network redundancy**—Availability refers to having reliable access to data. In contrast, durability refers to protection from loss. Durability is achieved by keeping multiple copies of data. If one copy is damaged, another copy can be used to restore the data. HA (high availability), however, is achieved with hardware redundancy—such as having multiple servers, cables, power supplies, and cooling solutions—so that if one of these components fails, a redundant component can take over. You've learned about how CSPs provide failover AZs where you can replicate virtual hardware across multiple physical data centers. If one data center experiences a crisis, your workloads can failover to a different AZ. This is an example of how to optimize both service and data availability.
- **Cloning and compression**—Cloning is the act of creating a new copy of existing data. The clone can replace the original with a simple swap and no further configuration. A similar concept for copying a disk drive is called imaging, where the copy is more compressed. Compression reduces the number of bits needed to store data by recoding the data in a more efficient manner.
- **Storage replication**—Storage replication supports data durability. Data is replicated across multiple data centers in a region or, for a higher price tag, across data centers in multiple regions. Storage replication to the backup target can occur immediately or periodically, as follows:
  - Synchronous replication writes data to the replica at the same time the data is written to primary storage. If a failure occurs with the primary data, the replica provides an exact, updated copy with no loss of data. Synchronous replication requires more network bandwidth, however, because data is written to two locations at the same time. And the writing process might take a little longer before the data target (such as a database) is again made available for other changes.
  - Asynchronous replication writes data to primary storage first and then periodically copies these updates to the replica. Asynchronous replication requires less network bandwidth but can also result in some data loss due to the delay in writing to the replica, depending on when a failure occurs.

Backup Types

Cloning creates a complete copy of the data at a single PIT (point in time), whereas data replication makes changes to the replica as, or soon after, the data itself changes. A clone represents a single, historical PIT and is not continually updated. In contrast, a replica is synced along with the original data as that data changes and so represents the current PIT.

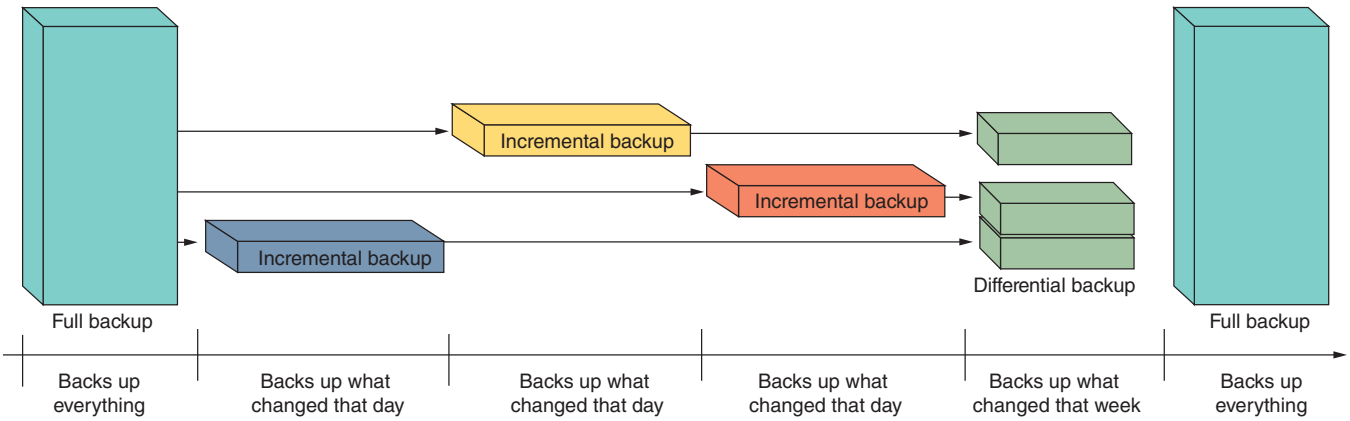
Different backup types take different approaches to copying data, which affects the PIT represented by the backup. Working with these various options can help maximize the use of available storage space, minimize the cost of storing backups, and reduce the time required to restore a backup. Consider the following backup types:

- **Full backup**—Backs up everything every time a backup is performed.
- **Incremental backup**—Backs up only data that has changed since the last backup of any kind.
- **Differential backup**—Backs up data that has changed since the last *full* backup.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-22 illustrates the conceptual differences between these backup types. It might help to imagine the different file sizes if you were to compare a full backup file, an incremental backup file, and a differential backup file side by side. The full backup would be the largest because it copies all the original data, even old data that hasn't changed in a long time. The time to recover data from a full backup, called RTO (recovery time objective), would be fairly short because all the copied data exists in one place. In other words, full backups allow for a low RTO. However, full backups take up a lot of space and require a lot of time to create.

Figure 8-22 Incremental and differential backups require fewer resources than do full backups



The incremental backup would be the smallest file. It only includes data that has changed since the last backup of any kind. For example, if you perform a full backup on Sunday, then an incremental backup would copy whatever changed on Monday, the next incremental backup would copy whatever changed on Tuesday, and then the next incremental backup would include whatever changed on Wednesday. You have one full backup from early in the week, with several smaller files containing changes for each day. To recover this data to its RPO (recovery point objective) of recovering all data, you would need to start with the full backup, then apply changes from Monday, then changes from Tuesday, and finally, changes from Wednesday. This creates a series of dependencies between backup files. Incremental backups don't take up much space and don't require much time to create. However, you must make them frequently and allow more time to recover when relying on incremental backups (that is, incremental backups have a high RTO).

The size of a differential backup (also called a delta backup or delta differential backup) falls between the other two. It includes all data changed since the last full backup (this changed data is called the delta), regardless of any other smaller backups performed since then. For example, suppose you make a full backup on Sunday. On Monday, you would make a differential backup for changes made that day. On Tuesday, you would make a new differential backup of all changes made since Sunday's full backup. On Wednesday, you would create an even larger differential backup that includes all changes made since Sunday's full backup. To recover this data, you would need to start with the full backup made on Sunday, and then apply only the latest differential backup, which would contain all changes since the latest full backup was performed. A differential backup's RTO is higher than a full backup alone and lower than a full backup with many incremental backups.

Note 8-5 | Some backup systems can create a **synthetic full backup**, which is created from the most recent full backup plus any applicable incremental or differential backups. Functionally, a synthetic full backup is equivalent to a full backup and can reduce the processing load to create the backup. As with all backup types, confirm the completion of backup generation and regularly test restoration processes.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Backup Considerations

Several factors contribute to decisions about what backup strategy best fits your company's situation. Consider the following points:

- **SLA (service-level agreement)**—You've already learned that CSPs offer SLAs to guarantee a certain level of performance and recourse options for the customer if service falls below or above certain thresholds. An IT department might also have a written SLA with the organization it belongs to. Ideally, management will play a role in defining the terms of this SLA and will budget for sufficient funds to meet those expectations. These factors will be determined by thresholds such as RTO, which is the maximum tolerable outage time, and RPO, which defines how much data loss is tolerable (a day's worth? an hour's worth? no data loss?). This kind of information will factor into decisions on backup types and backup schedule. This will also determine whether backups should be stored online and configured for automatic failover, or stored offline, which requires more time for retrieval and recovery but provides additional protection. Many backup systems use a mixture of both.
- **Resource configurations**—The types of resources on your network, whether on-prem or in the cloud, will also dictate some of the available options and decisions that must be made. For example, backing up a database or object storage will require different backup methodologies than backing up a large collection of VMs and their connected storage.
- **Authentication backups**—As with device configurations, authentication directories and configurations should also be backed up.
- **Backup rotation**—Recall that a full backup backs up everything. So how many full backups should you keep? For that matter, how many iterations of any backup files should you keep? To a degree, this decision will be made at an organizational level and will be affected by applicable government regulations or industry standards, budgets, and risk tolerance levels. Make sure you know and implement the appropriate policy and schedule for rotating backup files. On a similar note, when deleting backup files, make sure you follow any relevant standards or guidelines for adequately destroying this data. You don't want deleted backups to be recoverable by an attacker.
- **Security**—Where and how data is stored must be considered in the context of industry security standards and regulations. Data residency restrictions might dictate that data—even backup copies of data—be stored within certain geopolitical regions. Compliance standards might also require certain kinds of encryption, including the encryption of stored databases, file systems, volumes, or objects. The next section explores storage security more thoroughly.
- **Recoverability testing**—Creating backups is only part of the story. Those backups will do you no good if you can't restore data from the backups when data is lost or damaged. For example, regularly check your backup's integrity—if your backup files are corrupted or if the backup's storage medium is damaged, the backup will fail. Another major problem is needing the backup and then finding you haven't been creating backups of the data you thought you were protecting. Some professionals call this a resume-generating event in that you might find yourself unemployed and searching for another job as a result of this failure. Regularly test restoring from backups to ensure your team is familiar with the process and to make sure the backup system works reliably and is doing what it should.
- **Recovery options**—Many people think of a backup as being a singular entity. If you restore from a backup, you restore the entire thing (that is, storage volume, database, file system, and so on). However, many backup services offer more granular control rather than only being able to restore in bulk. For example, suppose only a single file or a handful of files are corrupted when a storage drive fails. Granular recovery allows you to restore only the affected files instead of the entire backup.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

9. What data characteristic ensures the data continues to exist?

   a. Consistency
   b. Availability
   c. Redundancy
   d. Durability

10. What type of copy ensures the latest changes are always included in the backup?

    a. Incremental backup
    b. Differential backup
    c. Synchronous replication
    d. Asynchronous replication

11. Which backup type creates the largest backup file?

    a. Incremental
    b. Full
    c. Differential
    d. Delta

Check your answers at the end of this chapter.

Section 8-5: Storage Security

Certification

2.5: Given a set of requirements, provision the appropriate cloud resources.

4.2: Compare and contrast aspects of compliance and regulation.

4.4: Given a scenario, apply security best practices.

4.5: Given a scenario, apply security controls in the cloud.

You've already learned about security configurations in AWS, Azure, and GCP platforms, including ACLs, zero trust design, and user or host authentication and authorization. These security tools can apply to storage services as well. For example, AWS S3 buckets can be protected by ACLs that determine which AWS accounts can access data in which buckets and what those accounts can do with that data. Data objects or volumes can reside in security zones best suited to the data's security needs. A security zone refers to the level of protection in place, for example, whether there's an emphasis on preventative controls (which provide active mitigation of threats) or on detective controls (which passively report signs of potential problems). Grouping data sets, applications, and other resources by security zone can simplify security settings. Additional approaches to securing data are described in this section.

Storage Security Techniques

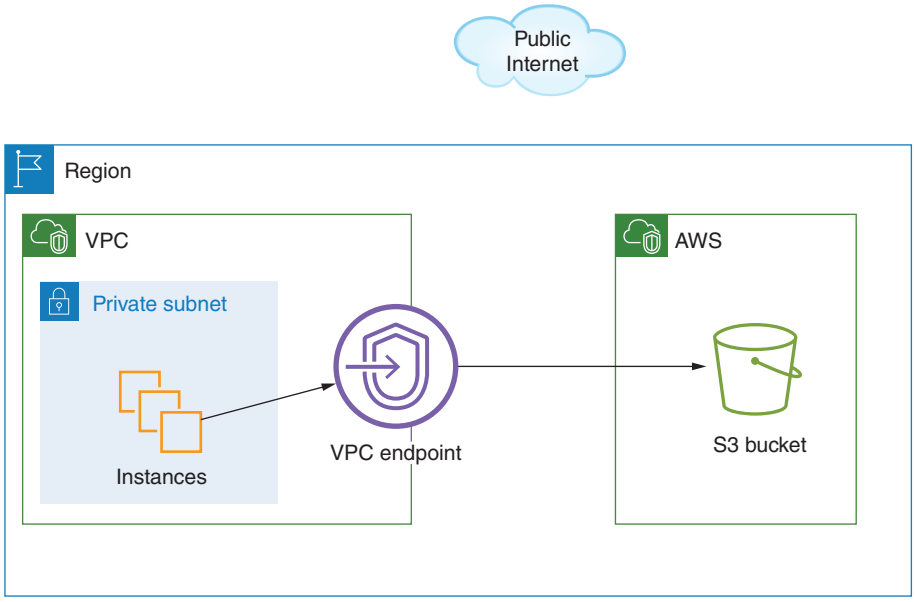
Data security techniques can vary depending on the type of storage being used. Following is a list of common storage security measures used in the cloud for object, block, and file system storage types:

- **Encryption**—Regardless of storage type, data security is greatly increased by encrypting data at rest, in transit, and in use, which you've already learned about. Typically, the most secure option is to encrypt data before it's uploaded to the cloud; however, this can make it difficult to access and use the data while it's in the cloud because the required encryption keys might not be accessible. Many configuration options are available to ensure data is encrypted as it's transferred to the cloud. Once there, different encryption processes can protect the data at rest, and key management services can help handle encryption keys in addition to managing access to those keys.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Access management**—Due to the highly configurable nature of the cloud, access control and limiting who has what permissions in the cloud are critical. This is no less true in relation to storage security. Whether data is stored in an object-based storage service or a file system service, access management through policies, permissions, and roles can help protect data from unauthorized access.
- **Network security**—You've learned a great deal already about network security in the cloud, such as security groups, firewalls, and ACLs. All these tools can help protect stored data, too. For example, any resource containing sensitive data should be placed within a private subnet, if that's possible. In the case of non-networked resources, such as AWS's S3, other configurations should be used to restrict network traffic's ability to reach these resources. For example, AWS offers VPC (virtual private cloud) endpoints that can direct traffic between VPC-hosted resources and an S3 bucket without sending that traffic across the public Internet (Figure 8-23). As you continue in your studies, take notice of specialized, proprietary services in a specific cloud platform—such as VPC endpoints—that are designed to improve accessibility while minimizing data's exposure to the Internet.

Figure 8-23 A VPC endpoint provides private access from a VPC to an S3 bucket



- **DLP (data loss prevention)**—As you've already learned, DLP is a technique that uses software to monitor confidential data, track data access and ownership, and prevent it from being copied or transmitted off the network. AWS's Macie, mentioned earlier in this chapter, is an example of a DLP service. Macie monitors S3 buckets for sensitive data that is not secured from public access and triggers security responses to these exposures. Other services, such as GCP's Cloud DLP or third party Forcepoint and Trellix, can enforce security policies that apply to sensitive data, trigger alerts in response to detecting suspicious activity, and automate responses that quarantine files or block unauthorized access.
- **Compliance auditing**—Regulations, laws, and guidelines exist to help organizations protect their data. Many organizations embrace this opportunity for optimized security by pursuing security certifications for their cloud deployments. For example, the ISO/IEC (International Organization for Standardization/International Electrotechnical Commission) 27001 standard defines cybersecurity best practices for organizations of all sizes. Earning the ISO/IEC 27001 certification is a major accomplishment for an organization and gives its customers increased confidence that the organization takes cybersecurity and data privacy seriously according to stringent industry standards.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Data Classification

Some data requires greater protection than other data. One way to do this is data classification. When designing a storage infrastructure and access rights to the data, it's easy to initially think in terms of classifying data by data type or by department or team. For example, you might classify all HR-related data together or all accounting-related data together. This kind of data classification might seem convenient at first but will create security-related headaches as you try to protect that data going forward. Instead, it might be more beneficial to think of and categorize data along different axes. For example, consider each axis listed here, how it could affect the organization of data in a storage system, and how proper organization could increase security around the data:

- **Sensitivity**—Who should have access to the data? Should the data be accessible from outside the organization or only from within the organization? Grouping data objects together that require similar levels of access control or monitoring makes it easier to manage that data effectively and securely.
- **Compliance**—What kinds of restrictions and regulations dictate how the data should be stored? Keep in mind that these requirements will vary—and might be subject to multiple layers of requirements—depending on what information is contained within the data, where the data was collected, and when it was first created. For example, patient medical records will be subject to HIPAA (Health Insurance Portability and Accountability Act) requirements but might also be subject to certain state laws or public health reporting requirements. Encryption requirements might address whether data must be encrypted at rest, in motion, in use, or all of these. Relevant regulations might also list acceptable encryption protocols or restrictions on where keys are stored and how they're accessed. There might be standards for determining who can view the data versus who can change it.
- **Life cycle**—How long will the data be active, and at what point should it be transitioned to colder storage, archived, or deleted? Use data compression and data tags to help manage large chunks of similarly life-cycled data more efficiently and enable versioning only on the data sets that need it. Also, situate data in regions that put it close to where it's being used or migrated to other storage types, such as archival storage. This will speed up data transfer times, reduce data transfer costs, and minimize the complexity of applicable governance requirements.
- **Visibility**—How should data be tracked? Logs can be configured on organized data sets that best reveal the information needed without over- or under-monitoring.

Data Obfuscation

You've already learned about encryption, which is a form of data obfuscation. **Obfuscation** refers to techniques that make data difficult or impossible to interpret even if it is discovered or stolen. For example, data that is encrypted appears to be gibberish and must either be decrypted using the correct key, or you must have knowledge of encryption technology as well as significant processing power to break the encryption and convert the data back into plaintext. One weakness of encryption in the context of data obfuscation, however, is that encryption is reversible. The evolution of quantum technologies might, at some point, provide such highly sophisticated and efficient processing power that quantum computers could easily break through today's encryption protocols, leaving vulnerable massive amounts of sensitive data.

An alternative to encryption is **tokenization**. With this technique, sensitive data is replaced by a token, which is random data unrelated to the original data. It might be in the same format as the original data; for example, it might have 16 digits like a credit card number does. This is called a format-preserving token; it can be presented to the database and other tools in a way that doesn't cause data validation issues. Alternatively, a nonformat-preserving token doesn't resemble the original data at all.

The token, then, sits in place of the sensitive data so that, if the company's stored data is hacked, the attacker won't obtain the sensitive data itself and would have no way of reversing the tokenization, even with very powerful tools. For example, suppose the company has stored tokens in place of credit card numbers for purchases from customers. The company sends the tokens to the payment processor, who can then exchange

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the token with a third-party tokenization service that provided the token to retrieve the sensitive data. That data is stored in the tokenization service's data vault, which is the only place where each credit card number is associated with a specific token. Additionally, these credit card numbers are then stored separately from related data, such as the owner's name or the card's expiration date. Even if someone were to access the data vault, a list of credit card numbers by themselves is not particularly useful.

This process is similar to checking baggage when flying on a commercial airliner. Each passenger receives a tag associated with their luggage. The luggage is managed and transported by airline staff while each passenger holds the tag associated with their bag. As passengers arrive at their destinations, they can exchange their tags for their luggage. Similarly, the data itself is stored at the tokenization provider's site in a way that would be useless to an attacker. Only when redeemed by the token and placed back into its original context does the data provide any valuable information.

Regulations for Storage Security

Government regulations and industry standards address many aspects of data security. Some of these factors exist in the more abstract business environment or political environment, not just in a technical environment. The rules that apply to one instance of data might be very different than rules that apply to similar data in a different context.

Generally, data is owned by someone, whether that is a person or an organization. **Data ownership** refers to the possession of and responsibility for data. The question of who owns what data can sometimes be tricky once data is stored or used by someone other than the entity who created the data. Sometimes, you knowingly sign away rights to data ownership, such as when you post information on a public discussion forum and agree that you no longer own the copyright for anything you post. Other times, the law defaults to the position that the creator of the data owns that data. In some scenarios, the term "ownership" is more relevant to a team within an organization that is using data, even though the organization itself or even its customers might own the data from a legal standpoint. And sometimes, the right to make decisions about what happens to data about you belongs to you regardless of whether you created the data, possess the data, or have direct access to the data. The question of data ownership is determined by many factors—who the data is about, who created the data, who is using the data, where the data is stored, what kinds of information the data contains, and the context in which ownership is being discussed.

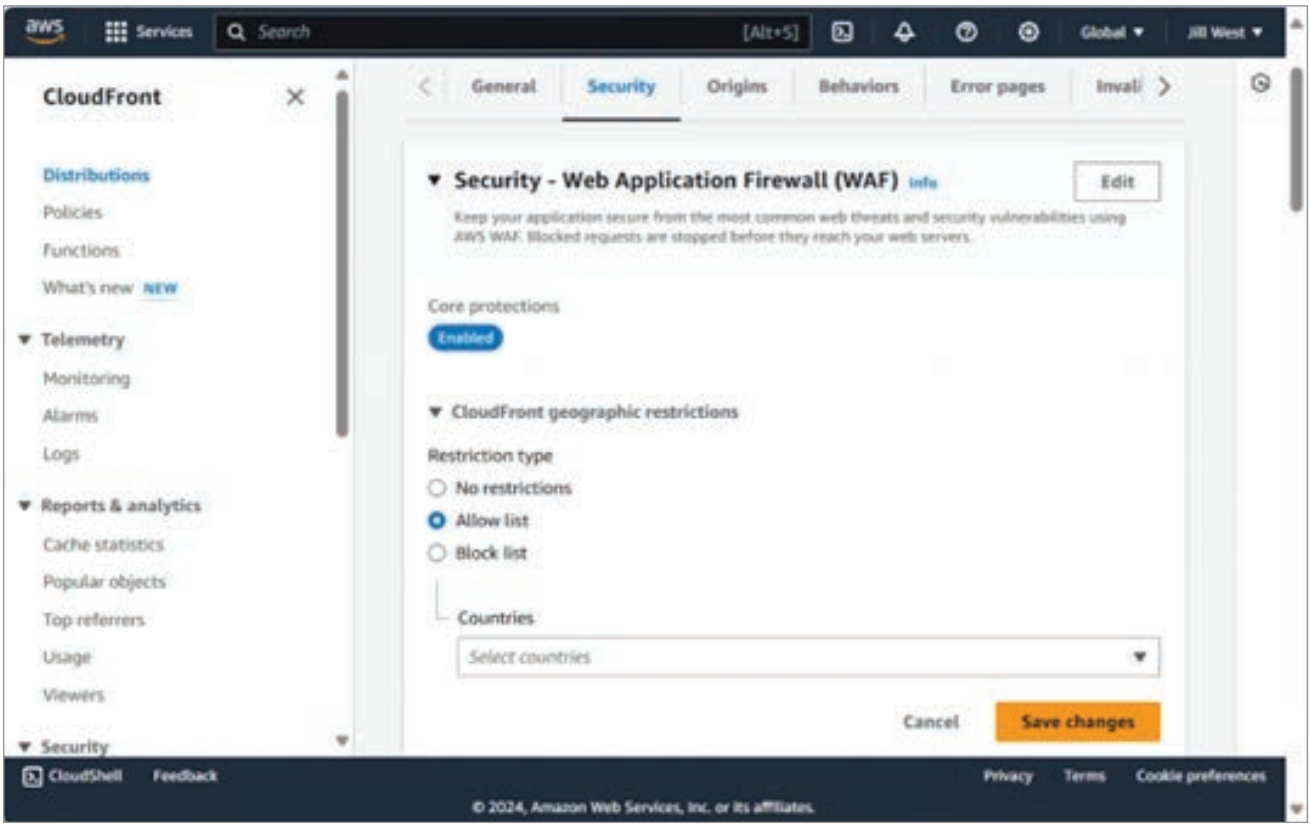
Determining data ownership can be especially relevant when discussing data sovereignty. **Data sovereignty** refers to the right to determine what laws apply to data. Most countries dictate that data residing within their geopolitical boundaries is governed by their own laws. Some regulations, such as the GDPR (General Data Protection Regulation) in the European Union, define strict requirements organizations must obey when handling data about the citizens of protected countries, even if the organizations themselves are not located in those countries. Some laws prevent the movement of certain types of data across international boundaries. And other laws require that data entering their country be subject to inspection. For this reason, you should always be careful about what sensitive data you bring with you when traveling internationally. If you have customer, patient, or student data encrypted on your laptop, for example, and you enter a particular country with these kinds of laws, you might be required to decrypt your data and subject it to inspection. This can cause serious issues with data privacy laws in your own country.

Regarding the cloud, data sovereignty requirements often dictate data residency, or data locality. This means that certain data must be stored within specific geopolitical boundaries. Increasingly, cloud providers are offering options for customers to manage the physical location of their data in order to comply with these requirements. For example, you might choose a particular region for a data storage resource that ensures your data is not transported beyond a certain country's borders. Alternatively, you might choose a data storage location with the intent of keeping the data close to its point of use, which reduces its exposure while being transported across networking infrastructure.

Figure 8-24 shows where you can configure countries that are allowed or blocked from accessing data stored in an AWS S3 bucket using CloudFront and WAF (Web Application Firewall). Other services can help keep data about that country's customers isolated within that country's borders and not duplicated to other regions.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-24 Allow or block access to stored data based on the user's country



Source: Amazon Web Services, Inc.

In addition to data's physical locality, laws and regulations often address requirements for **data retention**, which refers to the storage of data for a defined period. For example, in the United States, state laws generally require that pediatric medical records be retained for at least 10 years or until the child reaches the age of majority (typically age 18), whichever is longer. These rules can be complex and often, different governing bodies—such as the federal government and various state governments—create their own laws that all must be considered when determining data retention requirements for any specific instance of data.

Business contracts can apply their own requirements for data retention policies. In some cases, these contracts define minimum requirements for data retention similar to the laws discussed previously, such as 5 years. However, sometimes contracts define *maximum* thresholds for data retention. For example, a contract might require that personal or proprietary data be retained only as long as that data is needed to fulfill the terms of the contract, after which the data copies must be destroyed. There might also be requirements on how data is stored, the security applied to data storage, and allowances for who can access the data.

Whatever industry you work in as an IT professional, make sure to research regulations, laws, and industry standards that apply to the data you work with. For example, elementary school records must be handled very differently than medical records at a dentist's office. Your employer should provide adequate training that ensures you're aware of the requirements applicable to your work. But keep in mind that not knowing what laws apply to you is not considered an acceptable defense if your actions or inaction results in the exposure of protected data.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

12. Which of the following is the most secure form of data obfuscation?

a. Classification
b. Encryption
c. Authorization
d. Tokenization

13. Which of the following aspects of data security would be at risk with a CDN service such as AWS CloudFront?

a. Data locality
b. Data sovereignty
c. Data ownership
d. Data retention

☐ Check your answers at the end of this chapter.

Summary

Section 8-1: Storage Types

- Data types stored in the cloud include office and media files, applications, websites, logs, and databases.
- Three basic networking approaches to storing data on-prem include DAS, NAS, and SAN.
- Cloud storage services, also called STaaS, offer similar storage categories plus one more type of storage: file system storage, block storage, and the cloud-native object storage.
- Relational databases incorporate a specific structure to the data, while nonrelational databases are more flexible in their design.

Section 8-2: Storage Configuration

- Storage tiers—such as hot storage, warm storage, cold storage, and archive storage—can decrease the cost of storing data in exchange for increasing the cost and time of retrieving it.
- Storage performance can be evaluated based on several measurable factors, including transfer rate or read/write throughput, IOPS, and latency.
- Cloud services such as AWS DataSync can manage the online transfer of massive volumes of data over long periods when there is a potential for errors or other delays, while offline data transfer consists of loading encrypted data onto a storage appliance and then physically shipping that appliance to the CSP, either by a common package-delivery service or via a dedicated transport.
- Use restrictions, limited tenancy, and enforced encryption can increase the security of cloud-based data disposal.

Section 8-3: Cloud Storage and Database Services

- S3 is AWS's most popular and flexible cloud storage service, able to host just about any kind of data.
- To use most of Azure's storage services, you must first create a storage account.
- GCP Cloud Storage supports popular features such as life cycle management and object versioning; however, some of these features are only available through the CLI.

Section 8-4: Creating and Storing Backups

- Backup professionals suggest, at the minimum, following the 3-2-1 Rule for backups; more sophisticated backup strategies that rely more natively on the cloud include the following variations of this rule: 3-1-2, 3-2-1-1-0, and 4-3-2.
- In some cases, backup retention length is dictated by regulatory standards; in other cases, you can choose what makes the most sense in your situation.
- Choosing an appropriate backup type—such as full backups, incremental backups, or differential backups—can help maximize the use of available storage space, minimize the cost of storing backups, and reduce the time required to restore a backup.
- Backup strategy design relies on several factors, such as relevant SLAs, resource configurations, authentication backup needs, backup rotation schedule, security needs, recoverability testing, and other needed recovery options.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 8-5: Storage Security

- Common storage security measures used in the cloud for object, block, and file system storage types include encryption, access management, network security, DLP, and compliance auditing.
- Data classification might be performed along a variety of axes, such as sensitivity, compliance requirements, life cycle, and visibility needs.
- Tokenization is an irreversible form of data obfuscation.
- The question of data ownership is determined by many factors—who the data is about, who created the data, who is using the data, where the data is stored, what kinds of information the data contains, and the context in which ownership is being discussed.

Key Terms

For definitions of key terms, check the Glossary.

big data
DAS (direct attached storage)
data ownership
data retention
data sovereignty
durability
graph database
key-value database
legal hold
life cycle rule
local storage
NAS (network attached storage)
nonrelational database
obfuscation
RAID (Redundant Array of Independent Disks or Redundant Array of Inexpensive Disks)
relational database
SAN (storage area network)
storage tier
synthetic full backup
tokenization
vector database
versioning

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

CDN content delivery network
IOPS input/output operations per second
iSCSI Internet Small Computer System Interface
JSON JavaScript Object Notation
LUN logical unit number
NAS network attached storage
NoSQL not-only Structured Query Language
NVME nonvolatile memory express
RAID redundant array of independent disks or redundant array of inexpensive disks
SAN storage area network
SQL Structured Query Language

Review Questions

1. Which type of cloud storage would be the best fit for storing medical X-rays?

   a. File system
   b. Relational database
   c. Block
   d. Object

2. Which language is used to ask questions of a relational database?

   a. SQL
   b. FC
   c. LLM
   d. JSON

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. What kind of data is best stored in a nonrelational database?

   a. Metadata
   b. Unstructured data
   c. Versioned data
   d. Encrypted data

4. What effect does enabling versioning have on your cloud storage?

   a. Decreases durability
   b. Increases retention-based restrictions
   c. Increases cost
   d. Requires secure data disposal

5. In which cloud storage option can you store persistent data from a container that is then available to thousands of other containers?

   a. Block storage
   b. Writable layer
   c. Readable layer
   d. File system

6. Which data life cycle state is complicated by data redundancies?

   a. Frequent access
   b. Archive
   c. Destruction
   d. Creation

7. What kind of storage volume exists only when it is connected to a VM instance?

   a. Distributed
   b. Persistent
   c. Ephemeral
   d. Elastic

8. Which of the following statements are true? **(Select two.)**

   a. Durability is not the same thing as a backup.
   b. All backup copies should be stored together.
   c. Archive data and backup data are the same thing.
   d. Storage replication supports data durability.

9. Under court order, you applied a legal hold to a backup copy of a relational database for a court hearing that will occur in three months. This version of the backup is scheduled for deletion under its normal life cycle at the end of the week. When will the backup copy be deleted?

   a. When the backup copy is moved to the archive tier.
   b. When the legal hold is removed from the backup copy.
   c. When the current life cycle phase ends.
   d. When the backup copy is marked for deletion.

10. Suppose you make a full backup the first Sunday of each month, a differential backup on each of the other Sundays, and an incremental backup every Tuesday and Thursday. Today is the 12th, which is the second Wednesday of this month. Which files (using the latest version available for each) need to be combined to create a current synthetic full backup?

    a. One full backup
    b. One full backup and two differential backups
    c. One full backup, one differential backup, and one incremental backup
    d. One full backup, one differential backup, and two incremental backups

Scenario-Based Questions

Question 8-1

Emile works for a start-up tech firm that caters to financial advisors and consultants in a thriving metropolis. His firm's existing network architecture worked fine for a two-person team forming the concept of their business and wooing investors. Now they're ready to expand to a fully functional virtual environment that hosts applications on several VM and container instances and supports shared storage among all the instances. However, Emile must keep in mind that his company still cannot afford specialized expertise for managing their network storage. Which storage architecture is the best fit for Emile's needs?

a. SAN
b. DAS
c. JBOD
d. NAS

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Question 8-2

Amber is restoring a database backup after a server failure caused by flood damage from a leaky pipe. Her company creates a full backup on the first weekend of each month, a differential backup each weekend between those full backups, and an incremental backup each weeknight. Today is Wednesday in the third week of the month. Which backup files will Amber need to apply to this month's full backup to restore all backed-up data?

a. The most recent differential backup and the most recent incremental backup
b. The second differential backup and the two most recent incremental backups
c. Both differential backups and the most recent incremental backup
d. Both differential backups and the two most recent incremental backups

Question 8-3

Mia's company is moving their video database—about 34 TB—to a cloud-based storage solution. The company is located in a fairly small town, at least an hour's drive from the nearest large city, and doesn't do a great deal of data transfer online for normal business operations other than regular contact with customers through email or video conferences. Which data transfer option is the best fit for Mia's company?

a. Private connection through a colocation facility
b. Storage appliance shipped via UPS
c. Existing public Internet connection
d. Dedicated VPN connection

Question 8-4

Mateo works for a mid-sized healthcare company. He's working on a project to migrate some of the company's data to cloud storage, and Mateo is specifically in charge of planning security configurations to keep the data safe, according to all relevant government regulations and industry standards for each type of data being migrated. The team has decided to take a classification approach to defining security requirements and, eventually, access control policies in the cloud. Which of the following data types would most likely be subject to HIPAA regulations, requiring one of the most sensitive classification levels? **(Select two.)**

a. Employee HR files
b. Patient medical records
c. Customer service emails
d. Marketing materials

Hands-On Projects

Note 8-6 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Project 8-1: Manage Storage in AWS

Estimated time: 30 minutes (+10 minutes for group work, if assigned)
Objective 1.4: Compare and contrast storage resources and technologies.
Group work: This project includes enhancements when assigned as a group project.
Resources:

- AWS account
- A photo or screenshot saved as a .jpg file on the local computer
- Internet access

Description: Storage management responsibilities include controlling who has access to data and what kind of access they have. It also includes keeping sufficient backups to restore any data that is lost, damaged, or accidentally deleted. In this project, you'll practice working with access policies for storage buckets and objects in S3. You'll also practice restoring an object after deletion.

Note 8-7 | At the time of this writing, all the services used in this project offered sufficient free tier limits to complete this project with no fees so long as the following are true:

- You stay within 5 GB of monthly storage space used.
- You are within the first 12 months of when your account was created.

To view a current list of what services are included for free during your AWS free trial, visit aws.amazon.com/free.

All project steps can also be completed using an AWS Academy account with no additional fees.

Complete the following steps:

1. In your AWS Management Console, go to the **S3** dashboard. S3 is listed under the **Storage** service category.
2. Click **Create bucket** to create a storage bucket. Give the bucket a unique name. This name must be unique across all of AWS S3, so you might have to get a little creative with the bucket name. What name did you give your bucket?
3. Scroll down to the "Block Public Access settings for this bucket" section. In this section, leave all default public access restrictions in place. Notice that, by default, your bucket's configurations will prevent any public access to your bucket or its contents.
4. In the bucket versioning section, enable versioning. In the encryption section, what is the default source of encryption keys for your bucket?
5. Leave all other default settings and click **Create bucket**.
6. In your list of buckets, click the bucket you just created.
7. Upload a photo from your local computer to your bucket. If you don't have a photo on your computer, you can take a screenshot of part of your screen. Save the screenshot as a .jpg file, and then upload the image to your bucket. Scroll through the available options for destination, permissions, and properties, but don't change any default settings for the upload.
8. Return to the Objects tab for your bucket and click the file you just uploaded. Copy and paste the object URL in a new tab in a different browser or in an incognito window for your current browser. What output do you get on the new tab?
9. Keep this tab open, as you will refresh it throughout this project to test public access to the object.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10. Return to the bucket's page and click the bucket's name in the file path at the top of the screen, and then go to the bucket's **Permissions** tab. In the Block public access section, click **Edit**. Click to deselect all the public access restrictions for your bucket and save your changes. Confirm the changes.
11. Return to your separate browser tab with the Object URL and refresh the page. Does it show the image this time? Why do you think this is?

AWS has changed the way public access is granted to a bucket and its objects. Previously, users would create ACLs to manage access to the bucket or to individual objects. While this functionality is still available in S3 at the time of this writing, AWS recommends avoiding the use of ACLs when possible and, instead, managing access to objects at the bucket level using bucket policies. At the time of this writing, bucket policies are managed using JSON. To change the bucket policy, you'll need to edit the Policy JSON Document. Complete the following steps:

12. Return to the object's page. On the object's **Properties** tab, copy the object's Amazon Resource Number (ARN).
13. Return to the bucket's page. On the bucket's **Permissions** tab in the Bucket policy section, click **Edit**. In the "Edit statement" pane on the right, click **Add new statement**. In the "Add actions" section, select **S3**. Under "Access level - read," select **GetObject**. Notice the JSON code changing in the Policy pane as you make your selections. What code lines reflect the selections you've made so far?
14. Next to "Add a resource," click **Add**. The S3 service should be preselected. For resource type, select **object**. For resource ARN, paste your object's ARN that you copied in Step 12. Click **Add resource**. What code line reflects this selection?
15. To make the object accessible to the public (that is, to all principals), you'll need to edit the Principal field manually. In the line that defines the Principal, replace the curly brackets `{}` with `"*"` (that is, open quote, asterisk, close quote) to use the wildcard `*` in reference to all principals. The line should read

    ```
    "Principal": "*",
    ```

    without changing any other lines.
16. At the bottom of the page, click **Save changes**. If you get any error messages, troubleshoot the problem.

> **Note 8-8** | AWS Academy users might notice a warning saying they do not have permission to use the policy validator. This is okay. Ignore the warning and continue with the steps as instructed.

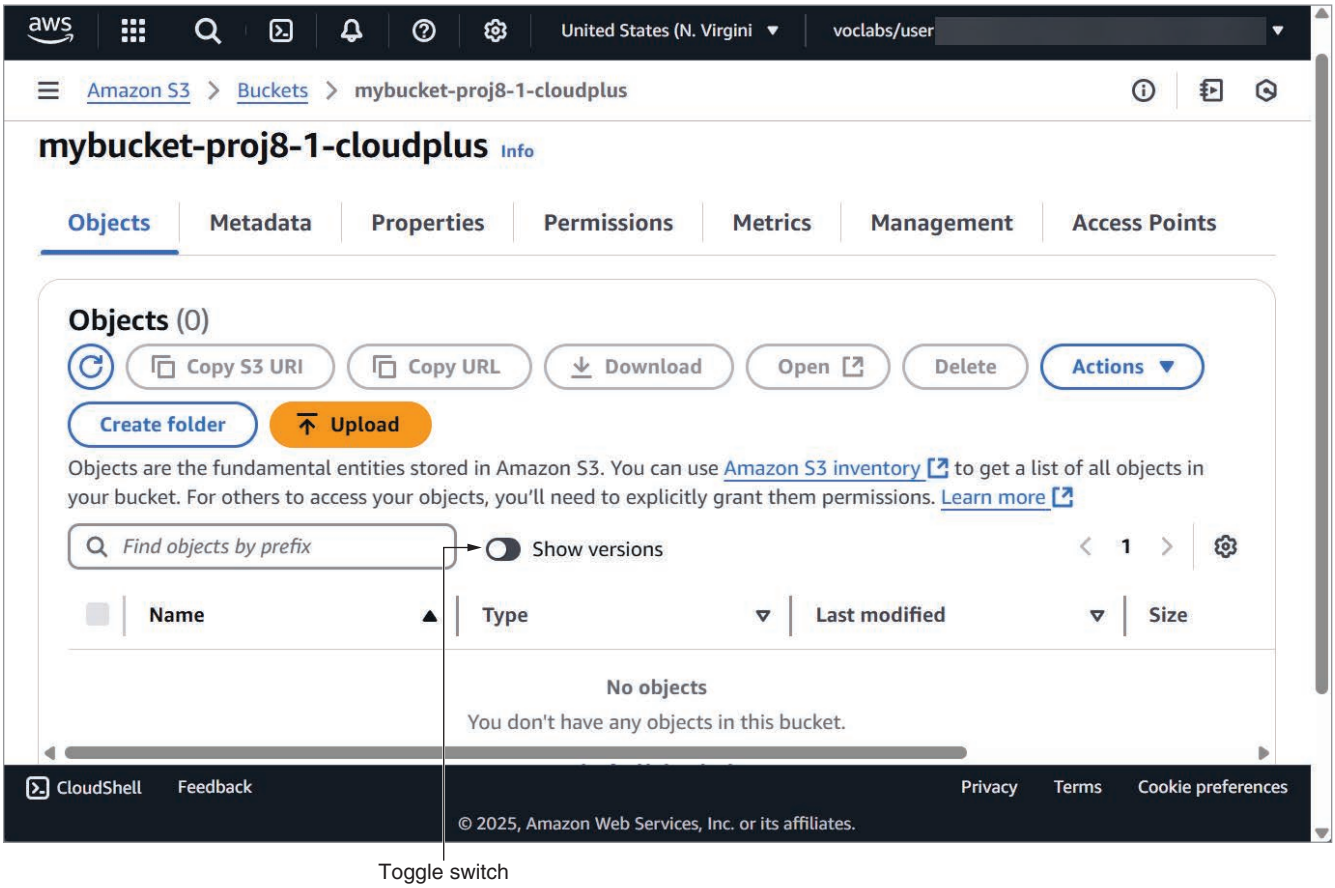
17. When the bucket policy is successfully edited, **take a screenshot** of your bucket policy; submit this visual with your responses to this project's questions.
18. Try again to access the file through the Object URL on your incognito tab. Does it work? Why do you think this is?
19. **For group assignments:** Each group member should edit access to their object so only one other group member, using their AWS account, can access the object over the Internet. **Take a screenshot** of the access settings that allow access to your group member but to no one else; submit this visual with your responses to this project's questions.

Now you're ready to practice deleting the object and recovering it. Complete the following steps:

20. Return to the bucket's **Objects** tab where the file is listed in the bucket. Click the check box to select the file, and then click **Delete**. Confirm the deletion. How many objects are listed on your bucket's Objects tab now?
21. Toggle the switch next to "Show versions" (Figure 8-25). Now how many items are listed in your bucket?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8-25 Show versions toggle switch



Source: Amazon Web Services, Inc.

22. One of these items is the file itself, and the other is a delete marker on the object that suppresses the file so it won't show up in a list of objects contained in the bucket. The file itself is not deleted, only hidden. To recover the file, click to select the Delete marker item in the list and click **Delete**. Confirm the deletion of the delete marker (not the object itself), and then return to the bucket's Objects tab. Toggle off the "Show versions" switch. How many items are listed in your bucket now? Why do you think this is?
23. Delete all the resources you created in this project, including the object and the bucket. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
24. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 8-9 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Project 8-2: Manage Storage in Azure

Estimated time: 30 minutes
Objective 1.4: Compare and contrast storage resources and technologies.
Resources:

- Azure account and subscription
- Internet access

Description: Storage management responsibilities include controlling who has access to data and what kind of access they have. It also includes keeping sufficient backups to restore any data that is lost or accidentally deleted. In this project, you'll practice working with storage accounts, storage containers, and blobs. You'll also create and restore a blob snapshot.

Note 8-10 | At the time of this writing, all the services used in this project offered sufficient free limits to complete this project with no fees so long as the following are true:

- You stay within 5 GB of monthly storage space used.
- You are within the first 12 months of when your account was created.
- You choose the LRS (locally-redundant storage) option when creating your storage account.

To display a current list of what services are included for free during your Azure free trial, in the navigation pane, click All services and search for Free services.

All project steps can also be completed using an Azure for Students subscription and your available free credit.

Complete the following steps:

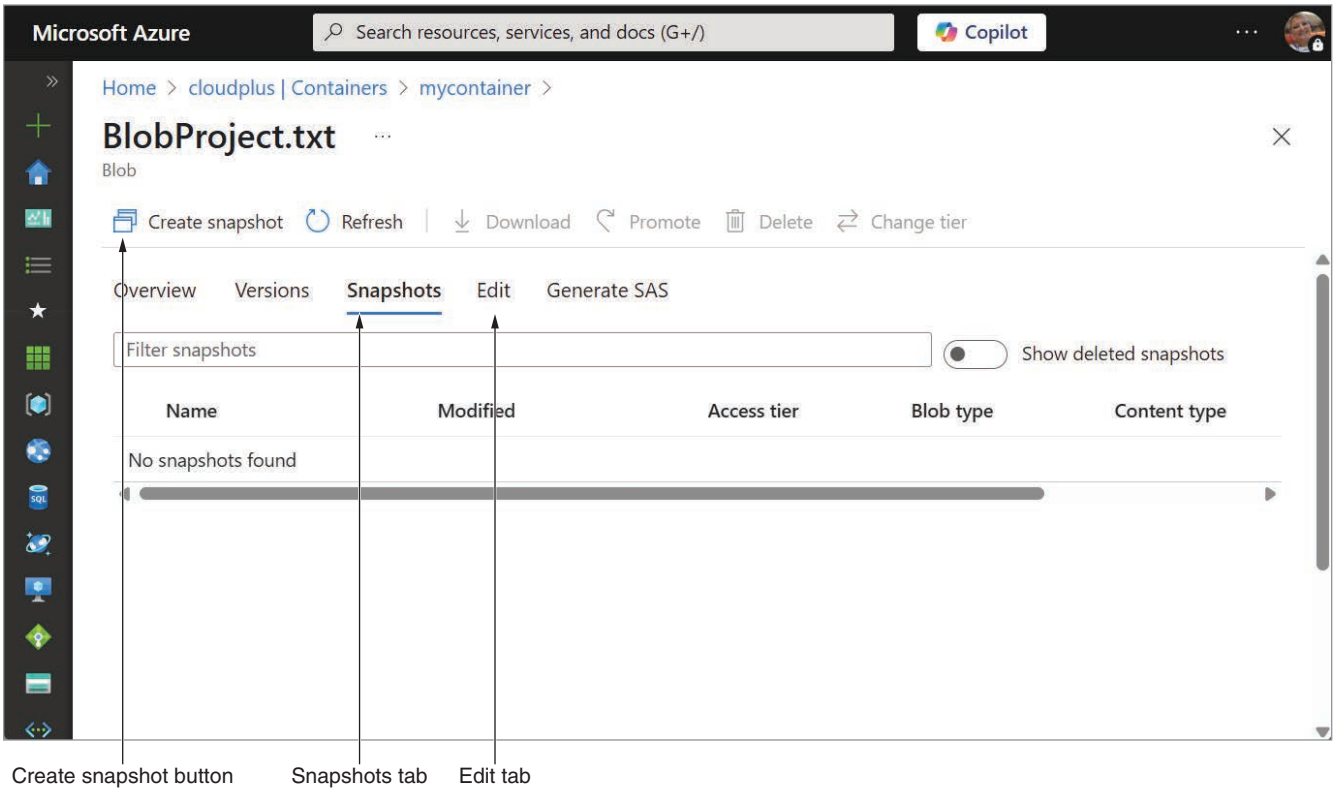
1. First, you need to create a text file to upload and edit later in the project. These instructions are written for Windows and can be adapted for other OSs. Create a new text document in Notepad. Type some text, such as This is a test. Save the document and name it something easily identifiable, such as BlobProject.txt. Close the document. What is the name of your document, and where did you save it?
2. You need a storage account to manage storage in Azure. In your Azure portal on the navigation menu, click **Storage accounts**. Click **Create storage account**. Use the following information to configure your storage account:
   a. Make sure the desired subscription is selected–for students using the Azure education account, make sure Azure for Students is selected for the subscription.
   b. Click **Create new** to create a new resource group. Give the resource group a name.
   c. Name your storage account something that is unique across all of Azure.
   d. Change Redundancy to **Locally-redundant storage (LRS)**–this setting is required to stay within free tier service limits. Click **Next**.
   e. On the Advanced tab, check the box to **Allow enabling anonymous access on individual containers**.
   f. Leave all other options at their default selections. Click **Review + create**, and then click **Create**.
3. After deployment is complete, on the Storage accounts page, click your new storage account. What types of data storage are available, as listed in the blade's navigation pane?

Note 8-11 | Recall that storage containers in Azure are similar in concept to buckets in AWS and GCP.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. In the blade's navigation pane, under Data storage, click **Containers**. Create a new container and give it a name, such as mycontainer. Without changing the Public access level setting, examine the anonymous access level list. What options are available?
5. Leave the Public access level set to **Private (no anonymous access)** and click **Create**.
6. Click the container you just created. To upload the text file you created in Step 1, click **Upload** and select the file. Click the **Advanced** dropdown menu. By default, the blob type is set to Block blob. Without changing the setting, check what other types are available. What options are listed?
7. Leave the Blob type set to **Block blob**, and then click **Upload**.
8. Click the file you just uploaded. Copy the URL and paste it into a new tab in your browser. What happens?
9. Close the browser tab and return to your Azure portal. Close the blob blade. On the container blade, click **Change access level**. Change the Public access level selection to **Blob (anonymous read access for blobs only)** and click **OK**. Click the filename again, and try again to open the text file's URL in a new tab in your browser. What happens this time?
10. Close the browser tab and return to your Azure portal. On the blob's blade, click the **Snapshots** tab, and then click **Create snapshot** (Figure 8-26). The new snapshot appears in the list. **Take a screenshot** of the list containing the snapshot; submit this visual with your responses to this project's questions.

Figure 8-26 Blob's Snapshots tab



11. Click the **Edit** tab. Change the text to something noticeably different, such as Beam me up, Scotty. Save your changes. Return to the **Overview** tab, and open the blob's URL in a new browser tab. What happens this time?
12. Close the browser tab and return to your Azure portal. On the blob's blade, click the **Snapshots** tab. Select the snapshot and click **Promote**. Open the blob's URL in a new browser tab again. What happens this time?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

13. Delete all the resources you created in this project, including the storage account, container, and blob. In what order did you delete these resources? What error messages did you encounter? How did you handle these problems?
14. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

> **Note 8-12** | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 8-3: Manage Storage in GCP

Estimated time: 30 minutes
Objective 1.4: Compare and contrast storage resources and technologies.
Resources:

- GCP account
- A photo or screenshot saved as a .jpg file on the local computer
- Internet access

Description: Storage management responsibilities include controlling who has access to data and what kind of access they have. It also includes keeping sufficient backups to restore any data that is lost or accidentally deleted. In this project, you'll practice working with storage buckets and objects in Cloud Storage. You'll also practice setting up life cycle rules.

> **Note 8-13** | At the time of this writing, all the services used in this project offered sufficient free tier limits to complete this project with no fees so long as the following are true:
> - You stay within 5 GB of monthly storage space used.
> - You create storage buckets in the us-east1, us-west1, or us-central1 regions.
>
> To check a current list of what services are included for free during your GCP free trial, visit cloud.google.com/free.

Complete the following steps:

1. In your GCP console, create a new project.
2. In the navigation menu, point to **Cloud Storage**, and click **Buckets**. If it opens, close the LEARN pane on the right. To create a bucket, click **Create**. Give the bucket a name that is unique across all of Cloud Storage and click **Continue**.
3. Change the location type to **Region**. Make sure the Location is set to one of the regions that supports free tier services: us-east1, us-west1, or us-central1. Click **Continue**.
4. Leave the default storage class as **Standard** and click **Continue**.
5. Deselect the option to **Enforce public access prevention on this bucket**. Change the access control settings to **Fine-grained** and click **Continue**.
6. Deselect the **Soft delete policy** and click **Create**.
7. On the pop-up message regarding preventing public access, click **Confirm**.
8. On your new bucket's page, click **Upload** and then click **Upload files**. Select a photo file from your local computer to add to your bucket. If you don't have a photo on your computer, you can take a screenshot of part of your screen. Save the screenshot as a .jpg file, and then upload the image to your bucket. Click **Open**. The object is added to the list near the bottom of your screen. What is the public access setting for your object?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9. Click the ellipsis icon for your object, and then click **Edit access**. How many permissions are listed here?
10. To make the object publicly accessible, click **Add entry**. Change the Entity field to **Public**. The Name field should auto-populate with **allUsers**. Leave the Access field as **Reader**. **Take a screenshot** of the entry configuration; submit this visual with your responses to this project's questions.
11. Click **Save**. Point to the question mark icon next to the Public access column and read the information. What is the Public access setting for your object now?
12. To truly test public access to this object, you'll need to attempt to open the URL in a browser where you're not signed in to GCP or any other Google site or browser. Copy the object's URL. Open a different browser where you're not signed into Google, or open an incognito window in your current browser, and paste the link in that window. What happens?
13. Do not close this browser tab or window. Test your control over the permission settings for this object by "breaking" the public access configuration. Edit access to this object again, and then delete the **Public** item you created in Steps 10 and 11. Save your changes.
14. The paper clip icon next to the object's access setting disappears; however, you should still have the URL showing in the other browser's tab or window. Return to that page and refresh the page. What happens?
15. Click the bucket's **Protection** tab. Under Object versioning, click **Object versioning off** to enable object versioning. Check the box to **Add recommended lifecycle rules to manage version costs**. Click **Confirm**.
16. This action adds two life cycle rules. To explore other life cycle rule options, click **Manage rules**, and then click **Add a rule**.
17. Suppose objects in this bucket will be accessed maybe three or four times a year. Which life cycle action would be the best fit for this scenario?
18. The wait time for a life cycle rule to take effect is longer than what is appropriate for this project. Therefore, click **Cancel**.
19. Go to the list of projects and shut down the project you created in Step 1. Confirm the project is shut down and scheduled for deletion.
20. Make sure to properly sign out of your account.

> **Note 8-14** | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Challenge Projects

> **Note 8-15** | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 8-1: Host a Static Website

Estimated time: 45 minutes (+10 minutes for group work, if assigned)
Objective 2.5: Given a set of requirements, provision the appropriate cloud resources.
Group work: This project includes enhancements when assigned as a group project.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Resources:

- Public cloud account (such as AWS, Azure, or GCP)
- Internet access

Description: In the Hands-On Projects for this chapter, you practiced creating storage buckets, adding objects to the buckets, and configuring access permissions for those objects. In this project, you will apply those skills to host a static website in a public cloud storage service.

Imagine you are building a small website for a local charity. You've decided that one way to keep costs down for maintaining the website is to design a static website that can be hosted in a public cloud storage bucket. This means there can be no server-side processing, such as accessing information from a database or managing login tasks. As you research the possibilities for static websites, you find there are some beautiful templates you can use to design an impressive website that meet these requirements and will easily make a good impression on visitors to the site. But first, you need to make sure you have the infrastructure set up correctly in your public cloud account. You'll start with a very basic index file for your site as you experiment with access permissions.

Complete the following steps:

1. You need an index.html file to serve as the initial home page for your website. To do this, create a text document in Notepad or a similar simple text editor (not Microsoft Word or Google Docs) and save it as **index.html**.
2. Choose one of the following options to write the HTML code:
   - **Prompt an AI chat tool**–Use an AI chat tool to generate code for a basic index.html file that will create a simple static website. Copy the generated code into your text document. Do not use any text formatting in your file. Replace or add to some of the page content text, inserting information that identifies you as the author (such as your name and the name of this course).
   - **Copy the provided text**–Include the following text in your file. Do not use any text formatting, such as bold or italics. Instead, replace the bold, italicized text with your own text that is personalized and will identify you as the author (such as your name and the name of this course).

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="utf-8">
  <title>My Website</title>
  <style>
  body {
    font-family: Arial, Helvetica, sans-serif;
  }
  </style>
</head>
<body>
  <h1>My Static Webpage</h1>
  <p>This webpage is hosted in my cloud storage.</p>
</body>
</html>
```

Note 8-16 | When you save your file, make sure to name the file **index.html**. To do this, you might need to save the file first as a .txt file, and then do a Save As function and type the .html file extension in place of the .txt file extension.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. Do some research for your public cloud platform to find instructions for hosting a static website in that cloud. Some platforms require special settings for hosting a static website. For example, you might need to use a particular name for your bucket or container, you might need to enable static website hosting for the bucket, or you might need to eliminate pieces of the HTML code (such as the header). As you do your research, keep in mind the following details for this project:
   - You do not need to set up resources required to support a custom domain. For example, GCP requires a load balancer to use a custom domain, and this configuration is not required for this project.
   - You do not need to create an error file, also called a 404.html file, for this project.
   - You do not need to configure redirection rules.
   - You do not need to use a more specialized service that might be better designed for hosting static websites, such AWS Amplify Hosting.
4. If you need to create a project or resource group to manage project resources when you're finished, do so now.
5. In your selected public cloud platform, create a storage bucket or container and enable public access to this bucket or container. What is the name of your storage bucket or container?
6. Upload your index.html file to your bucket or container. If needed, enable public access to this object or blob. You might need to first enable public access for the bucket or container, if you did not do so earlier. What steps did you need to take to enable public access to the file?
7. Visit the website's public URL in an incognito window to ensure anyone can access your website. If you receive any error messages here, troubleshoot and fix the problem. **Take a screenshot** of your website showing the public URL in the browser's address bar and the personalized text on the webpage; submit this visual with your responses to this project's questions.
8. **For group assignments:** Sometimes changes to a website's content or design can take a little while to propagate to all users. Share your website's URL with a group member. Make sure they can view your website on their own device. Make some edits to your website content, such as adding a message to your teammate. Start a timer as soon as you save your changes, and have your teammate refresh their browser several times until the changes are visible on their end. How long did it take for the changes to appear to your group member?
9. Delete all the resources you created for this project, including any resource groups, storage accounts, buckets, containers, objects, or blobs.
10. Check through your account to confirm you permanently deleted or terminated any resources created in this project. Make sure to properly sign out of your account.

> **Note 8-17** | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Solutions to Self-Check Questions

Section 8-1: Storage Types

1. What type of storage is the hard drive in a laptop?

   Answer: c. DAS

   Explanation: DAS (direct attached storage), the most familiar storage type to the typical user, is directly connected to the computer or server that uses it, such as the hard drive in a laptop. A SAN (storage area network) is a specialized portion of an organization's network dedicated

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

entirely to supporting storage hardware. Each block in a SAN contains data that is referenced by the block's LUN (logical unit number). NAS (network attached storage) is attached to network infrastructure and relies on this network connection to provide file access to clients.

2. Which type of cloud storage would be the best fit to host a relational database?

   Answer: b. Block

   Explanation: Block storage works well for hosting databases where data changes frequently and requests for data must provide low-latency performance. Object storage, a cloud-native storage infrastructure, is best for data that is written once and left alone or only read. File system storage is best for scenarios where a folder hierarchy is required for direct connections to a larger number of users.

3. Which database type supports RAG?

   Answer: d. Vector databases

   Explanation: LLMs calculate vector embeddings for RAG-provided documentation and store these vectors in **vector databases**. A key-value database is essentially a list of key-value pairs, such as tags used to identify cloud resources. Graph databases store data according to entity relationships, such as connections between social media accounts. A relational database stores data in rigid table structures that must contain the same kinds of information for every record in a table.

Section 8-2: Storage Configuration

4. Which of these storage tiers would be most appropriate for archived data?

   Answer: a. Glacier

   Explanation: AWS's **Glacier** storage tiers provide the coldest and cheapest storage for archived data that rarely needs to be accessed. Infrequent Access and Standard tiers cost more to store the data but don't charge (or don't charge as much) for more frequently accessing the data, which is not appropriate for archived data that rarely needs to be accessed.

5. Which of the following metrics should be increased to improve storage performance?

   Answer: b. IOPS

   Explanation: IOPS (input/output operations per second) details how many read and write operations can be completed per second—the higher this number, the better performance a storage resource can provide. TTFB (time to first byte) and RTT (round-trip time) both measure latency, which should be as low as possible. While SSDs (solid state drives) provide much higher IOPS than HDDs, IOPS is the metric used to measure this performance.

Section 8-3: Cloud Storage and Database Services

6. AWS file system policies are configured using _____.

   Answer: b. JSON

   Explanation: To change AWS file system policies, you must edit **JSON (JavaScript Object Notation)** code. NFS (Network File System) is a protocol to give Linux- or UNIX-based systems access to files across a network. EFS (Elastic File System) provides a file system to Linux-based workloads. VTL (virtual tape library) is a service that presents virtual tape drives to an on-prem backup application.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7. An object stored in Azure is called a _____.

Answer: d. blob

Explanation: In Azure, storage containers are analogous to AWS S3's buckets, and **blobs** are like S3 objects. You create a storage account, create a container, and then you upload objects, called blobs, to it.

8. Which of the following is GCP's coldest storage tier?

Answer: a. Archive storage

Explanation: Archive storage is optimized for long-term storage that is rarely accessed and requires a minimum storage duration of at least 365 days. Nearline storage is designed for moderate frequency access, such as less than once a month, and requires storage duration of at least 30 days. Standard storage provides a balance of inexpensive storage for data that needs to be accessed frequently without requiring a minimum storage duration. Coldline storage is designed for data that is accessed less than once or twice a year and requires storage duration of at least 90 days.

Section 8-4: Creating and Storing Backups

9. What data characteristic ensures the data continues to exist?

Answer: d. Durability

Explanation: A CSP's guaranteed **durability** refers to the protection against losing a stored object. Consistency refers to how well copies of data are consistent with each other. Availability refers to having reliable access to data. High availability is achieved through redundancy, which ensures multiple copies of the data.

10. What type of copy ensures the latest changes are always included in the backup?

Answer: c. Synchronous replication

Explanation: Synchronous replication writes data to the replica at the same time the data is written to primary storage. If a failure occurs with the primary data, the replica provides an exact, updated copy. Asynchronous replication writes data to primary storage first and then periodically copies these updates to the replica, which can result in some data loss due to the delay in writing to the replica. An incremental backup backs up data that has changed since the last backup of any kind. A differential backup backs up data that has changed since the last full backup.

11. Which backup type creates the largest backup file?

Answer: b. Full

Explanation: A **full backup** creates the largest backup file because it copies all the original data, even old data that hasn't changed in a long time. Incremental and differential backups only back up what has changed since the last applicable backup. Delta identifies what data has changed since the previous backup—differential backups are also called delta backups.

Section 8-5: Storage Security

12. Which of the following is the most secure form of data obfuscation?

Answer: d. Tokenization

Explanation: With **tokenization**, sensitive data is replaced by a token, which is random data unrelated to the original data and cannot be reversed to reveal the original data. Classification simply identifies how sensitive data is, which can be used to inform access control methods. Encryption can be reversed. Authorization helps enforce access controls to protect data but is not a form of data obfuscation.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

13. Which of the following aspects of data security would be at risk with a CDN service such as AWS CloudFront?

Answer: a. Data locality

Explanation: CDN (content delivery network) services cache copies of data nearer to the points of use, which directly impacts any potential **data locality** restrictions. Data locality might be dictated by data sovereignty requirements. Data ownership doesn't change when a CDN caches copies closer to users. Data retention requirements might be indirectly impacted by copying data for a CDN cache, but because these caches are so short-lived, a significant impact is unlikely.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Observability and Monitoring

Chapter 9 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Describe cloud observability and monitoring tools and techniques.
2. Compare monitoring with events, logs, and traces.
3. Configure cloud monitoring tools.
4. Troubleshoot cloud problems and performance.

Chapter 9 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 9-1: Monitoring Resources
- Targets to Monitor
- Monitoring Tools
- Dashboards

Section 9-2: Events, Logs, and Traces
- Events
- Logs
- Traces

Section 9-3: Cloud Monitoring Services
- Monitoring in AWS
- Monitoring in Azure
- Monitoring in GCP

Section 9-4: Troubleshooting Problems and Performance
- Quotas and Throttling
- Utilization Issues
- Scaling and Sizing Issues
- Restoration from Backup

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 9 Scenario "Hey, Sam, how's it going?" you ask as Malik's 15-year-old son walks in the door. "Your dad will be back in a few minutes," you add. "He's fixing a problem with a printer."

"Hi," Sam mumbles as he shuffles over to Malik's desk and flops into his office chair.

"What's wrong? You seem mopey," you observe. Sam is a bright kid and normally very excited about one thing or another.

"Dad's going to be so mad at me," he groans. "I failed my English test today."

Just then, Kendra walks in and overhears Sam as he continues, "We've been reading short stories and talking about symbolism. Colors, weather, whatever. I don't know. Apparently, those authors always mean something more than what they say. It doesn't make sense, and then I flunk the test."

"Oh, I remember those discussions in high school English class," Kendra says as she smiles to herself. "I loved those stories! So much depth and meaning and significance. Math, however–I failed algebra . . . more than once," she confesses with a sheepish grin and an eye roll.

"But algebra is easy!" Sam argues. "You learn the rules and follow the steps. It's logical. Symbolism–how are you supposed to know what the author was thinking, when they say one thing but mean something else? It doesn't make sense."

"Ah, well, our minds all work differently, I guess," replies Kendra. "Algebra didn't make sense to me–how are you supposed to use letters that aren't really numbers when you're calculating things?"

"Will your dad really be mad?" you ask Sam. "He seems like a pretty laid-back guy to me."

"Well, he won't ground me or anything," Sam replies. "Dad says low grades are indicators of where you need to work harder. But I'd rather work on learning more about computers. That's a lot more interesting than old stories."

"Hmm, your dad's right," nods Kendra. "Grades show you what you're doing well and what needs more attention. But computers tell stories, too. And sometimes, you need to be able to explore underneath the plain meaning of things to discover what's really going on with a computer system."

"Oh, like when you're monitoring the network," you add, as you think about the detective work it took to identify the printer problem Malik's working on.

"Yes, and monitoring the cloud," adds Kendra. "Sometimes the surface problem isn't really the problem. You have to think about the bigger context, the clues you have, and even consider the information that you're missing, not just the information you have."

The next morning, as you, Kendra, and Malik meet to go over your cloud migration plans, Malik mentions to Kendra, "Thanks for chatting with Sam yesterday. He was really bummed about his English grade. But he was also a little bit excited to find out how studying symbolism can help him work better with computers."

"My pleasure," says Kendra as she settles into her chair. "He's a good kid."

Before the conversation could turn to the meeting agenda, you wonder aloud, "Kendra, something you said yesterday has been stuck in my head. You said, 'Sometimes the surface problem isn't really the problem.' With our on-prem network, we can go examine devices and cabling and everything to find out what the real problem is. How do we do that in the cloud?"

"Ah, good question," replies Kendra. "In the cloud, we have to rely on monitoring tools to help us find out what's going on. It's an awful lot of data, so we also have to rely on tools to help analyze the information and even respond automatically to some of the problems."

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

"Okay, just give us the research questions already," jokes Malik.

"Ha—right," grins Kendra. "Let's find some information on these questions," she adds as she scribbles on a page in her notebook.

- How are cloud resources monitored?
- What kinds of information can be collected about cloud resources and activities?
- What cloud services are designed to support observability and monitoring?
- How are problems and performance issues identified?

Section 9-1: Monitoring Resources

Certification

1.2 Explain concepts related to service availability.

1.8 Summarize cost considerations related to cloud usage.

2.3 Summarize aspects of cloud migration.

3.1 Given a scenario, configure appropriate resources to achieve observability.

Throughout this course, you've learned about different types of clouds, various services you can run in the cloud, and how to create resources in the cloud. You've learned how to organize these resources in virtual networks and subnets, and how to connect resources in the cloud and back to your on-prem data center. All of this is great until something goes wrong. How can you monitor your cloud for problems? How can you make sure that your users have an optimal experience interacting with your cloud resources? And how can you ensure this all stays within your budget?

In this chapter, you learn about how to manage your cloud's transparency, or observability, to monitor resources and respond to problems. You also learn how to manage your cloud's capacity to support services and resources and how to optimize both performance and cost efficiency. In Chapter 10, "Cloud Automation," you learn more about how to automate maintenance processes to minimize the demand on IT staff's time and further optimize cloud performance and security. But first, you need to explore what resources can be targeted by your monitoring tools as well as what information can be collected.

Targets to Monitor

To balance performance and costs, many of the cloud resources you've been working with—such as VM instances, storage buckets and volumes, databases, gateways, and load balancers—generate API (application programming interface) messages that report various metrics. Monitoring tools, whether native to the CSP's platform or installed from a third-party monitoring service, capture information from these APIs and organize the data to report on the health, activities, and expenses of the monitored resources. Specific KPIs (key performance indicators) are analyzed by these monitoring services. Together, this systematic design, called **observability**, ensures transparency of activities and resources in a distributed system that enables monitoring and automated responses.

For example, consider a running VM instance in the cloud. A common KPI for a VM is CPU utilization, which shows how much of a CPU's capacity is being used. With this information, you can increase the number of vCPUs for that server when needed (called vertical scaling). Alternatively, you could take the more common approach and spin up additional VMs in the server cluster (called horizontal scaling) if existing VMs

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

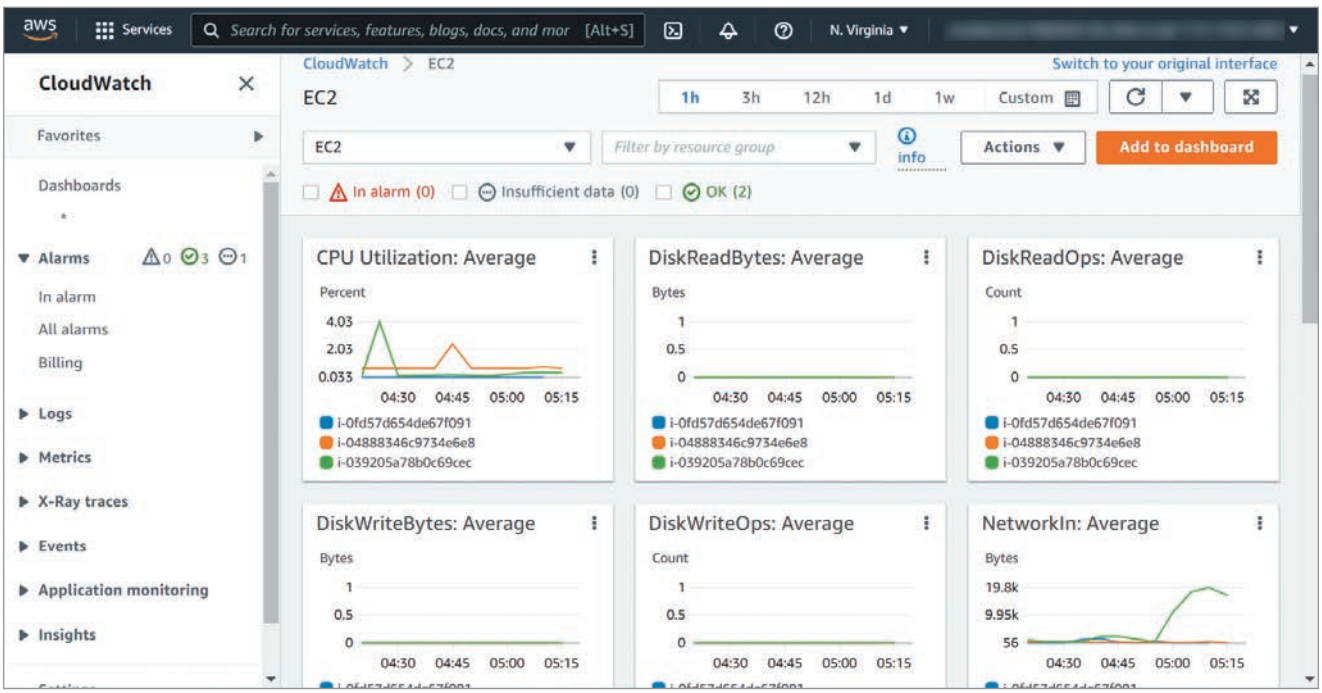
show signs of strain. Other metrics typically include data in or out of a network interface and disk read or write operations on a storage volume. All these metrics help indicate the VM's health, performance, and cost.

Many variables contribute to the fluctuation of cloud performance and costs. For example, a VM server will perform best if its CPU utilization remains below 80 percent. However, if CPU utilization falls below 10 percent for an extended period, this might indicate the VM instance is no longer needed and should be deallocated so it no longer costs you money. Note that managing the capacity of your cloud resources can help you optimize both cloud performance and expenses.

Note 9-1 | This concept of continuously monitoring a server's hardware (or virtual hardware) exertion to track its overall health is borrowed from the on-prem data center. For example, an on-prem server monitoring solution might use IPMI (Intelligent Platform Management Interface) standards to monitor and configure a server's hardware status, including power consumption, temperature, and fan speeds.

Figure 9-1 shows some metrics monitored by default in AWS for three running VM instances in EC2. Notice the different color lines for each instance and the various KPIs being monitored, such as CPU utilization, disk reads and writes, and network traffic. This figure also shows that three alarms are helping to monitor these resources. If one of the alarms is triggered, a cloud admin would need to dig into the alarm's information to determine what went wrong and whether the related resources need additional attention.

Figure 9-1 EC2 KPIs

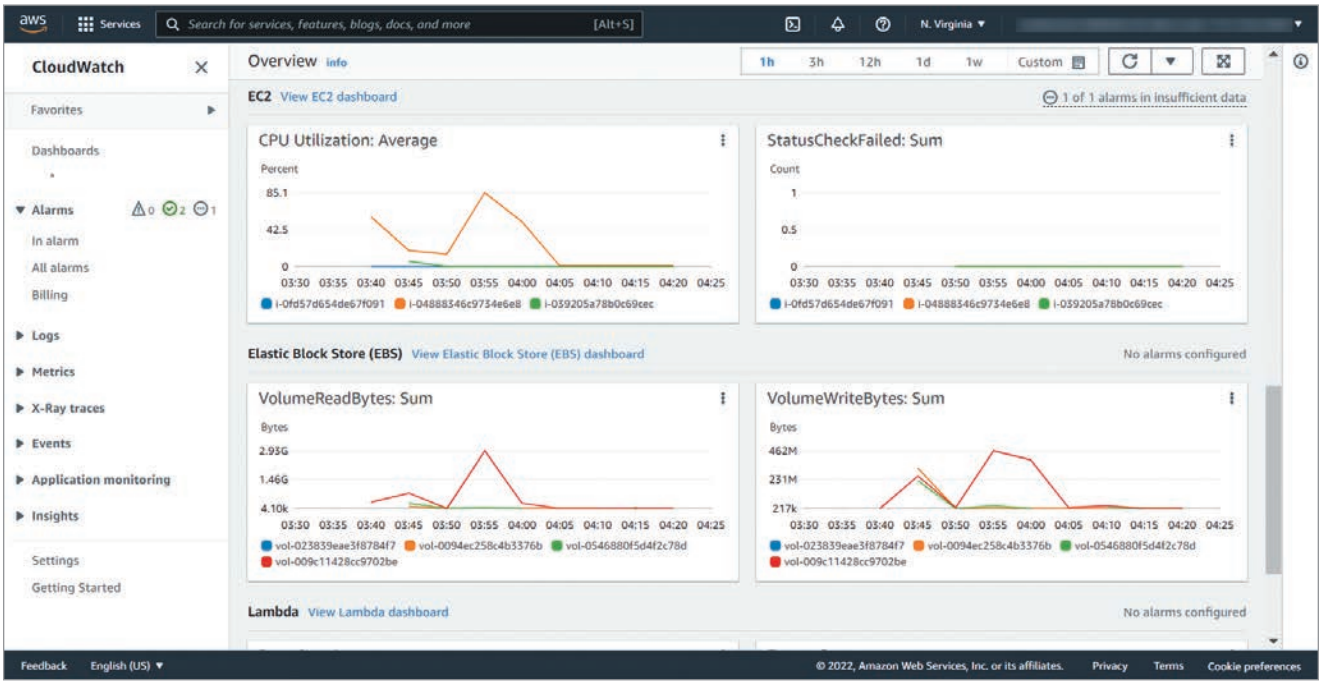


Source: Amazon Web Services, Inc.

Data collected on these metrics can be monitored on a service's or resource's own page, such as the page for a specific VM instance, or data from many services can be collected in a central location, giving a bigger picture of a cloud deployment's overall health. For example, AWS's CloudWatch service allows cloud technicians to monitor metrics across AWS services and collate this data in one or more dashboards, track events, and generate alerts. Figure 9-2 shows a cross service dashboard in CloudWatch. Azure Monitor and GCP's Cloud Monitoring perform similar functions in their respective platforms.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9-2 The cross service dashboard includes data on EC2 and EBS



Source: Amazon Web Services, Inc.

Common metrics to track in the cloud include the following:

- **Overall utilization**—How much storage space is used? How much memory or CPU capacity is being used consistently, and are those numbers pushing the boundaries of available resources or budget? Are resources bursting beyond normal thresholds? Is throughput peaking near available limits? These KPIs will indicate whether system load is increasing and your cloud needs increased capacity, if everything is stable, or if you should decrease its available capacity. If you've configured automated processes to adjust capacity, overall utilization will not only inform those processes, but can also give you insight into whether your automations are functioning as desired.
- **Elasticity usage**—You already know that scalability refers to a cloud resource's ability to scale up, down, in, or out to adjust to variable demand. Building on this concept, elasticity refers to a cloud resource's ability to increase or decrease in capacity automatically without requiring intervention from a cloud admin. For example, if more storage space is needed, does the storage service adjust automatically and dynamically to this demand, or does scaling require manual input from an admin? Elastic configurations might be appropriate in some scenarios but counterproductive in others. This could depend on how often the resource's needs fluctuate, how much money is saved by minimizing resource allocation at any given time, and whether the resource can handle a fluctuating environment or if it requires a more stable allocation.
- **Performance**—Cloud performance is determined by latency measurements and relative health of cloud resources while minimizing disruptive incidents. By reporting data based on metrics detailed in any relevant SLAs you're responsible for meeting, you can track whether you're achieving expected performance standards.
- **System or service availability**—An individual system's or service's availability is measured by uptime versus downtime, connectivity to other services, and the health and performance of its various components. Many times, an SLA also requires a certain minimum uptime for the overall monitored system. This is similar in concept to the on-prem KPIs of MTBF (mean time between failure), MTTF (mean time to failure), and MTTR (mean time to repair) that are reported for hardware such as disk drives and routers. The cloud can use these same terms for similar purposes with virtual systems. A related acronym is MTTD (mean time to detect), which refers to how long a failure exists before a monitoring

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

system identifies the problem. Yet another acronym, MTTA (mean time to acknowledge), is an important indicator of how responsive your procedures are to the occurrence of a problem. Just because a text message is sent to a cloud admin does not mean the problem is addressed right away if it's the middle of the night and no one is "on call" to handle the problem off-hours. Problems in any of these areas can indicate the need for further investigation to identify the cause of the problem and try to fix it or to evaluate your response protocols, staff scheduling, and distribution of responsibilities across your team.

- **Cost**—Budget limitations, business needs, and company policies require that costs of cloud resources be adequately tracked. This is not just a budget-limiting feature—tracking cloud costs can impact business decisions and might be necessary for billing to various budgets across projects, departments, and customer accounts. Cloud cost monitoring and optimization provide an opportunity to granularly monitor cloud consumption per user, customer, team, project, or department and make adjustments where needed. These costs then can be charged to the project responsible for some cloud resources, to the department that requested a particular system, or to the customer whose account must fund a product or service. Consider the following arrangements:
  - **Chargeback model**—Many companies use a chargeback model where expenses for specific resources or resource groups are charged back to a specific department in the organization according to who is using those resources. For example, if an accounting application is running in the cloud, the charges for that application's resources could be tracked separately and charged to the Accounting Department.
  - **Showback model**—In contrast, a showback model tracks usage of cloud resources without attributing specific costs to that usage. This serves the purpose of increasing cost awareness without introducing the complexity of separately tracking funding for cloud services.

All this information can be collected by monitoring tools that observe cloud services from the cloud platform's perspective. In other words, these tools don't function within the individual resources but collect information from the outside, like security cameras watching the outside of a building.

Alternatively, you can also install a software agent inside, say, an instance's OS to collect metrics from an insider's perspective. An **agent** is a lightweight application that runs in the background, automatically performing tasks on behalf of a user or another program. This would be similar to installing security cameras throughout the interior of a building. For example, AWS's CloudWatch Agent can be installed on an EC2 server instance running either Linux or Windows, or it can be installed in on-prem servers for hybrid cloud monitoring. The agent collects and reports metrics associated with counters in Windows Performance Monitor for a Windows server, which includes performance information on the OS and its applications, services, and drivers. On Linux systems, the agent reports detailed performance data on CPU, disk, memory, and network activity.

Resource Metering

One particular type of monitoring you might need in your cloud deployment is **resource metering**, which is a way to measure the usage of individual resources, such as network bandwidth or CPU cycles, and enforce a pay-as-you-go pricing model. You just learned that a chargeback model can be used to track expenses for specific resources to individual departments in your organization. Resource metering is a similar idea but can be implemented on your cloud-hosted services to document your customers' usage of your resources. With this information, you can then bill customers for their usage of your services based on the same kind of pay-as-you-go billing model that you, as a cloud customer, have with your cloud provider. For example, in AWS, you can use AWS Marketplace Metering APIs to have AWS bill your customers for their usage of your SaaS (Software as a Service) offerings.

Tagging Cloud Resources

Tracking various resources across many cloud-hosted systems can be made easier with tags. **Tags** are key-value pairs you create to track resources with whatever organization you need. For example, you might use tagging to categorize resources by project or department. Tags can track cloud expenses and correlate these

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

expenses to departmental budgets within your company. In a cloud environment where you might have difficulty taking inventory of all cloud resources and tracking resources throughout their various life cycle phases, tagging can help identify and monitor this expansive, abstract infrastructure.

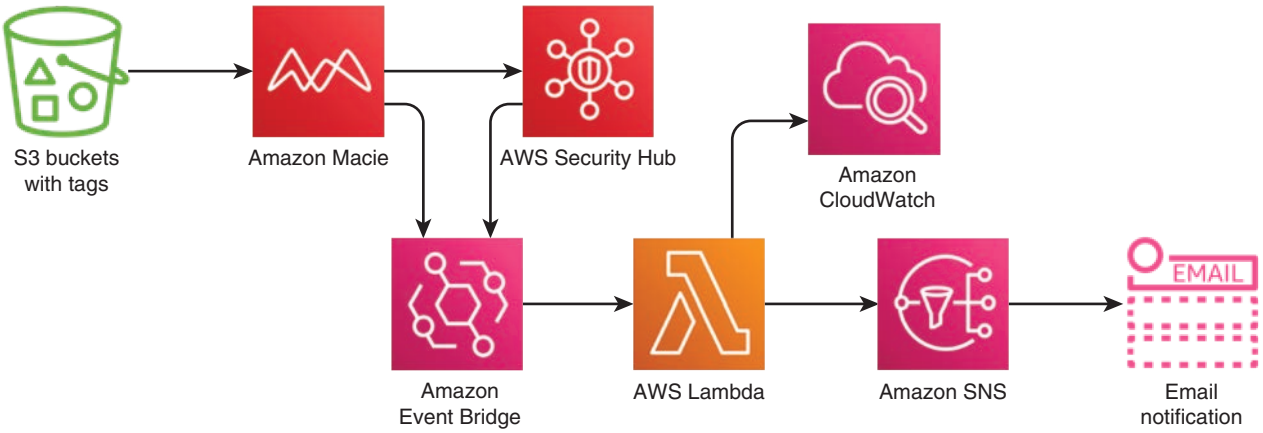
To use tags effectively, it helps to create a comprehensive tagging policy early in the cloud deployment process. Designed well, tags can increase the business value of cloud resources as well as make the cloud admins' work easier. Two layers of tag design include the following:

- **Explicit tagging**—Use key-value pairs to create useful resource metadata fields for identifying and tracking resources. For example, you might create a tag named "Department," and a possible value for this field could be "Accounting."
- **Implicit tagging**—Use naming conventions to increase consistency across resources and to increase saturation of information while requiring fewer metadata fields. For example, you might begin the names of all EC2 instances used by Human Resources with the letters "HR" followed by other needed information (such as project, purpose, location, etc.).

Many tags can be designed around naming conventions already in place for an on-prem network. Also consider which costs need to be tracked by tags and how these costs are allocated to various budgets by department, project, or initiative. You might also use tags to identify access policies for resources and services. For example, identify sensitive resources as "internal" or "classified," and tag less secure resources as "public" or "open."

AWS's Macie service works in partnership with AWS SecurityHub to validate data classifications using ML (machine learning) techniques, which search for sensitive data in S3 buckets. Macie regularly scans all S3 buckets and can accurately identify sensitive data such as bank account numbers, credit card information, authentication credentials, and PII (personally identifiable information). If any of this data is found to be improperly exposed, Macie can trigger an alarm that notifies cloud admins of the problem. Figure 9-3 illustrates a possible validation system where data is classified in S3 buckets using tags. Macie scans the buckets and triggers a Lambda function to send a notification if a data misclassification is found.

Figure 9-3 A validation system where data is classified in S3 buckets using tags



Note 9-2 | Tagging can also be performed automatically to ensure new resources—such as those created during auto-scaling operations—receive appropriate tags for tracking.

Monitoring Tools

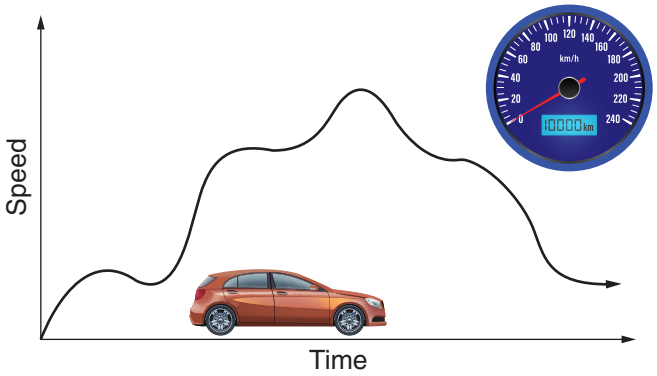
Now that you know what you're monitoring, let's explore what kinds of data you can collect. Some monitoring processes, such as logs, offer retroactive analysis of past events. Think about the grades you earn on quizzes. Each grade is recorded in the gradebook. At the end of the semester, you can observe the progress

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of your learning throughout the semester. Other monitoring processes provide intermittent insights into a resource's state or performance. For example, a traffic camera that monitors traffic volume on a highway might take a photo every two minutes and count the vehicles in each photo. The data from these snapshots provide insight over time for traffic conditions.

In contrast, some monitoring tools support continuous monitoring, which means the tool can provide near-immediate feedback on performance or security metrics in a continuous, unbroken manner. For example, consider the speedometer in a car. This gauge provides a continuous indication of the car's speed. You could map the car's speed on a line graph—the line would go up and down as the car accelerates and decelerates (Figure 9-4), but it would never be broken as long as the car is turned on. In a cloud network, a line graph showing data on inbound network traffic also could provide continuous information instead of intermittent measurements.

Figure 9-4 Vehicle speed mapped on a line graph



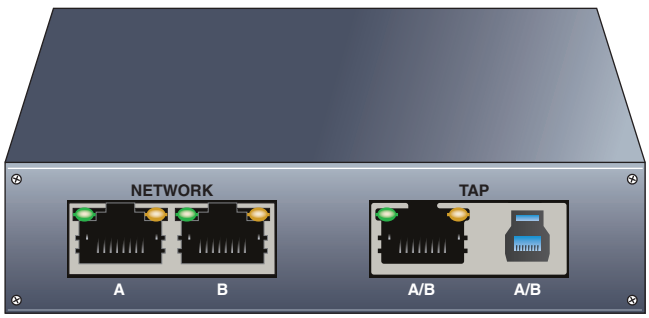
Alerts and alarms can inform staff when a threshold is breached or when a concerning event occurs. For example, if sensitive data is accessed outside of expected hours, an alarm can require attention to respond to the breach quickly. When implementing monitoring tools, verify continuous monitoring is in use where it's needed and consider the differences in timeliness for each monitoring strategy as you read about them in this chapter.

Capturing Network Traffic

Monitoring a network involves more than observing each individual resource. You also need to capture traffic between these resources and analyze these communications for signs of problems, such as network congestion or security issues.

If you've taken a networking class, you've probably learned about many of the methods used to collect network traffic on-prem. For example, the Wireshark application can capture traffic crossing a specific network interface. Captured traffic can be analyzed in-depth to determine what information is being transmitted, which protocols are in use, and what security gaps need to be addressed. You can even install a device called a network TAP (test access point), or packet sniffer, in line with network traffic to capture traffic flowing between various devices. Figure 9-5 shows a basic TAP. A packet broker near the TAP can sort this traffic, decrypt TLS traffic, mask sensitive data, and send different types of traffic to devices designed to further monitor and analyze traffic. Common targets include a WAF (web application firewall), a SIEM (security information and event management) tool, or an IDS/IPS (intrusion detection system/intrusion prevention system). Many network attacks can be detected with these tools.

Figure 9-5 Network TAP



What about cloud networks? You can't exactly insert a TAP device on the cable between two switches when you don't manage switches in the cloud. Some vendors have created solutions to help bridge the gap between traditional packet brokers and cloud infrastructure. These solutions capture cloud-based traffic related only to your resources and direct this traffic to an on-prem or cloud-hosted packet broker for monitoring and analysis. The tools are designed to distinguish between your cloud's traffic versus someone else's cloud traffic, even though both networks might use the same underlying physical infrastructure. Other, newer packet brokers are designed for hybrid or cloud deployments and can work across your entire network, whether physical, virtual, or cloud.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Monitoring User Experience

All monitoring is designed, in the end, to provide a satisfactory user experience, including ensuring security, performance, and availability of network resources. For example, as user density increases, capacity of resources (such as virtual desktops) needed by these users should be able to increase to meet fluctuating demand. However, this is only one layer of the monitoring and responsiveness needed to provide a positive user experience. Some tools are designed specifically to collect information on the user's actual experience of IT services. **ITSM (IT service management)** tools help IT teams monitor several customer-facing processes to better meet user needs and expectations. Some of these core practices include the following:

- **Service request management**—Requests for permissions, updates, or enhancements
- **Knowledge management**—Organization of accumulated knowledge specific to the organization
- **IT asset management**—Oversight of asset life cycles and configurations
- **Incident management**—Responses to unplanned disruptions to restore services
- **Problem management**—Discovery and mitigation of underlying problems that lead to incidents
- **Change management**—Oversight of changes to any IT system while minimizing disruptions

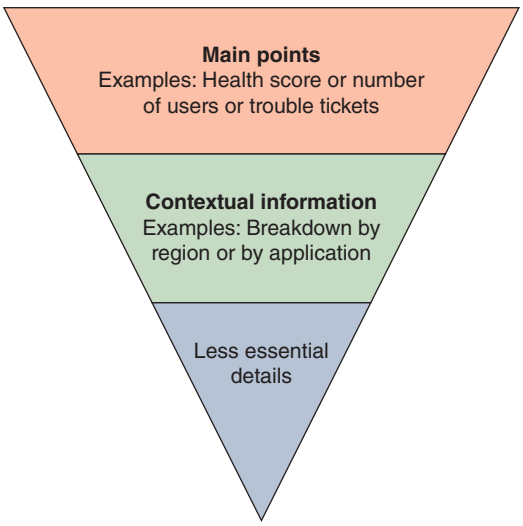
All these service management tools can benefit from incorporating and integrating a company's cloud systems and automation capabilities. Newer ITSM tools revolve around the core concepts of DevOps that offer cross-department cooperation and collaboration.

Dashboards

Dashboards play a crucial role in monitoring cloud resources, as this is one of the primary interfaces between human admins and monitored data about the underlying technical systems. A well-structured dashboard can give a cloud technician a quick overview of critical services and draw attention to issues that might indicate problems. Too much information in a dashboard causes confusion and can overwhelm users. Too little information decreases relevance. Poorly organized information takes too much time for an admin to digest and decreases efficiency. Cloud-based dashboard services offer the ability to configure a variety of dashboards for different users and use cases. Consider the following tips to design an effective dashboard:

- **Five-second rule**—When you're searching for an answer to a frequently asked question, you should be able to identify the information you need on the dashboard within about five seconds. If it takes longer to find what you're searching for, clean up the clutter and organize the information more logically.
- **First things first**—News stories typically start with an eye-catching headline, followed by just enough information to hook you into the story. The next paragraph or two in a news article gives the most newsworthy information. This is followed by additional details that might not be relevant or interesting to all readers. This organization of content is sometimes described as an inverted triangle (Figure 9-6). Similarly, your dashboard tells a constantly changing story of what's happening in your cloud. You want the most critical indicators noticed first, which alert you to key metrics and changes. This should be followed by data showing trends over time, to highlight the big picture of where these key metrics fit. Finally, a good dashboard includes more in-depth details at the bottom where you can follow the data trail and answer more specific questions, if needed.
- **Birds of a feather**—Group related pieces of information near each other. With some dashboard tools, you might be able to label the groups, which makes it easier to understand how they're related to each other. Similarly, make sure that any labels you use are easy to

Figure 9-6 The information with the most impact comes first, followed by more detailed information

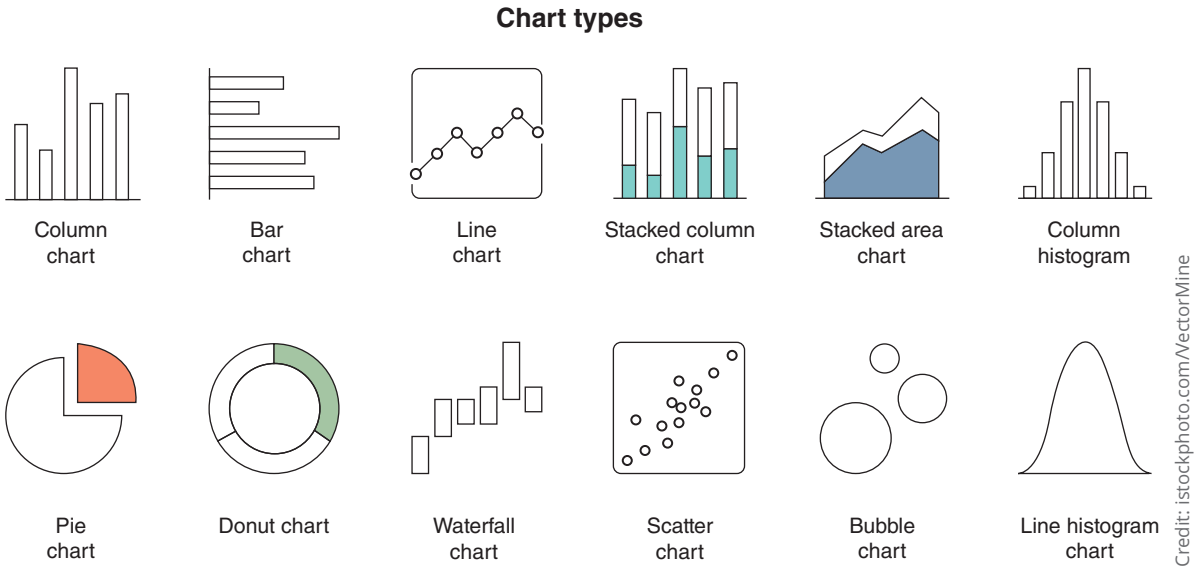


Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

interpret and will intuitively make sense to all the dashboard's users. Don't use such short or cryptic abbreviations that it's difficult to understand what they mean, especially for other members of your team. Be consistent in your naming conventions.

- **Less is more**—Have you ever wondered why phone numbers in the United States consist of seven digits (not counting area code)? Human minds can consume and remember only about five to seven distinct pieces of information at one time. Similarly, most people can only absorb about five to nine KPIs on a dashboard at a time. Even with all the available layers of information, you'll want fewer than 10 visualizations or other widgets on the entire dashboard.
- **A picture is worth a thousand words**—Choose your visualizations carefully. For example, consider whether a line chart, stacked chart, pie chart, or dial will better reveal the meaning and significance behind the data (Figure 9-7). You can use colors to quickly track and analyze trends for one resource in comparison to another, and you can use timeframes (the range of time covered by the visualization, such as 1 hour, 12 hours, or 2 days) and time increments (the time between each data point on the visualization, such as 1 minute, 15 minutes, or 1 hour) to better target the indicators you're searching for. When designing visualizations that incorporate color to reflect meaning, consider what other aspects of the visualization could communicate that same information to someone who is colorblind. For example, the shape of a data point on a chart, such as a square or a circle, can communicate the same information as using different colors.
- **It's what is not there**—Dashboards are made up of more than graphs, text, and boxes. Empty space between these visualizations is also important to assist the human eye in breaking down the information to absorb it. Leave sufficient and consistent white space between visualizations for easy comprehension of the presented information.

Figure 9-7 Choose visualizations that best match the data



Exam Tip | On the CompTIA Cloud+ exam, be prepared to answer questions asking which information is most relevant for a specific dashboard intended for a particular audience, such as a technical admin team or an executive committee. Consider the information most relevant to that audience, and think about how that information could best be presented to address common questions the intended audience will likely have. For example, you might need to decide which kind of chart would best display required data, or which charts—if included at all—should be placed higher or lower on the dashboard.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

How do you choose which data should be included on your dashboard? In the past, IT professionals have been primarily concerned with technical KPIs and thresholds that indicate how well the technology is performing in comparison to what it might be capable of doing. In today's cloud-run businesses, you've got to learn to think about the larger perspective. How do your cloud services fit into the business strategy of your organization? More specifically, how are users experiencing the services provided by your cloud resources? As many experts are now saying, "Slow is the new down." Outages might not be as common in this five or six 9s world as they used to be. But these days, slow services will lose business almost as quickly as outages will.

Dashboards provide a passive reporting of performance data that requires active monitoring from humans to catch and respond to problems. Much of this process can be automated, however. The options for doing this vary depending on your cloud platform. You learn more about automation techniques in Chapter 10, "Cloud Automation."

Note 9-3 | While it's important to limit the amount of data presented on a single dashboard, you'll also want to limit the number of dashboards you need to monitor. Not only does this reduce the number of screens you need to check regularly; it's also less expensive. CSPs often charge for running additional dashboards.

AI in IT | One of the limitations of designing dashboards is that these tools are generally passive interfaces and require the designer to anticipate which data and analytics will be most needed by which audiences. However, today's AI technologies can drive NLP (natural language processing) interactions that allow users to "chat" with their data. AI-powered BI (business intelligence) tools bypass the need for statically designed dashboards or complex SQL (Structured Query Language) queries, providing decision makers with deeper, more intuitive insights from massive pools of business data. Anyone in the company can ask questions without needing specific technical training to use these tools, which enables data-driven decision-making at all levels of job roles. Additionally, these tools can be configured to manage access control so users only receive information they are allowed to access, based on their job roles.

Self-Check

1. A resource's ability to scale automatically is referred to as _____
   - **a.** utilization
   - **b.** availability
   - **c.** performance
   - **d.** elasticity

2. Which resource notifies admins of a breached threshold?
   - **a.** Alarm
   - **b.** Tag
   - **c.** TAP
   - **d.** Metric

3. True or false: It's important to pack as much information as possible into a dashboard.
   - **a.** True, because running extra dashboards costs money.
   - **b.** False, because people can only absorb so much information at a time.
   - **c.** True, because extra space in a dashboard can be distracting.
   - **d.** False, because the most detailed information should be positioned at the top of the dashboard.

☐ Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 9-2: Events, Logs, and Traces

Certification

3.1: Given a scenario, configure appropriate resources to achieve observability.

4.3: Given a scenario, implement identity and access management.

4.6: Given a scenario, monitor suspicious activities to identify common attacks.

5.4: Explain the importance of tools used in DevOps environments.

You've already learned that metrics track information, such as performance and resource utilization. These metrics can be monitored using tools that capture live data, which is reported on dashboards. Other common data sources for monitoring cloud services are events, logs, and traces.

Events

Recall that an event is a detectable change in a cloud service or resource that you're monitoring, usually caused by some activity that prompted the change. For example, when a VM instance changes from a running state to a down state, that change registers as an event. Event collection occurs automatically in the background and, in most cases, causes no additional actions to be taken. Occasionally, however, an event needs more attention. What happens next depends partly on the cloud platform you're using and partly on how you've configured the event monitoring services within that platform.

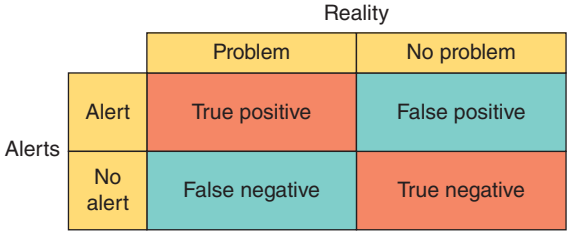
Events can capture almost every kind of activity performed within a cloud deployment, such as user logins, data access or modifications, resource configuration changes, application errors, and API calls. For example, in AWS, EventBridge can be configured to monitor for specific types of events (such as deleting data from an S3 bucket) that can, in turn, trigger an alarm or some other automated response. A difficulty of configuring event monitoring is that you generally must anticipate the kinds of activities that would indicate the problems you're searching for, and then configure monitoring rules that detect these events. Still, event monitoring offers many advantages for increasing observability in your cloud.

An important aspect of event collection and monitoring is **event correlation**. Events are generated from many resources throughout your cloud, and often these events are related. By identifying relationships between events through event correlation, monitoring tools can identify event patterns or anomalies that indicate performance or security problems. For example, IDP (intrusion detection and prevention) systems rely on event correlation processes to identify activities of concern. An IDP system might detect an account that hasn't been used in months suddenly has dozens of failed sign-in attempts. Perhaps the IDP system also detects a successful sign-in attempt to that account, followed by administrative commands being executed from it. The IDP might also associate this activity with an earlier port scan that came from the same IP address now being used to interact with the user account. Out of thousands or even millions of logged events, event correlation connects these dots to reveal an attack in progress.

As this example demonstrates, the more events available to a monitoring system, the more likely the system can identify and connect symptoms of a problem. On the other hand, collecting too many events slows down scanning processes and increases the likelihood of **false positives**, or false indicators of problems. To understand this term, it might help to compare the possible scenarios for each event or alert on a system.

Consider the diagram in Figure 9-8. A true positive is an alert sent for an actual problem. A true negative is when no alert is sent and there is no problem. Either of these situations means the IDP

Figure 9-8 True positives and true negatives accurately reflect reality



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

is functioning well. However, if you get an alert of a problem when there's no problem (a false positive) or you don't get an alert when there is a problem (a false negative), this can result in extra work for no reason or, worse, missed problems. Event collection policies need to be designed so they can determine which events are tracked, what thresholds yield the most accurate alerts, and how these alerts are responded to.

Fine-tuning alert thresholds can help reduce false positives and false negatives. This level of efficiency also helps admin and security teams better prioritize, or triage, the alerts that do come in. **Triage** means to prioritize and organize alerts in a way that ensures the most critical issues are addressed first. The most critical issues typically relate to outages, severe performance problems, or the riskiest security vulnerabilities and threats.

Different teams use different methods of determining these priorities. Some of the most relevant factors when triaging alerts include the following:

- The likelihood the issue will result in a larger problem, such as determining how easy it would be for an attacker to exploit a security vulnerability
- The extent of damage the issue is likely to result in, such as the sensitivity of the resources affected and the scope of potential impact
- The amount of time or other resources that will be required to remediate the issue, such as handling an issue that can be resolved quickly before handling a more concerning issue that will take several days or weeks to resolve

For example, an identified security vulnerability might involve extremely sensitive data that is not encrypted with the best encryption standards, but this vulnerability might still be classified as low priority if the likelihood of an attacker discovering and exploiting the vulnerability is very low. In contrast, a failure of a single node in a database cluster might not appear to be high priority because redundant database servers have already taken over the workload and ensured continued availability of the service. However, the failed node represents a dangerous reduction in redundancy for a critical service, which might result in this alert being rated as requiring urgent attention. Cloud platforms offer services, such as AWS's GuardDuty, that can assist with determining the priority of identified issues, especially when related to security.

Logs

Whereas metrics and events generate data in near real time, **logs** collect streams of data to show historical insights and reveal usage patterns and anomalies over time. Logs collect data about events and then make that data available for retroactive queries and analysis. For example, in AWS, CloudTrail captures information about each API call made—whether by the cloud customer or by interactions between resources—and stores that data in logs, enabling detailed monitoring. This insight supports comprehensive research into activities that might have contributed to problems or behavior indicating a security issue. Three common types of logs include the following:

- **Access or authentication**—A network server logs authentication attempts, both failed and successful, and might also log other access information and related activity.
- **System**—Network systems can log traffic flows; database systems can log changes to data content or design; automated systems can log configuration changes; and security systems can log security events. An OS keeps logs containing information about its drivers, system processes, and errors.
- **Application**—Application logs can collect a variety of event types, depending on how the software developer designed the application. These logs might include periodic information on the state of the application, its performance, or changes to data.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Applying Concepts 9-1: Explore Logs in Windows Event Viewer

Estimated time: 15 minutes
Objective 3.1: Given a scenario, configure appropriate resources to achieve observability.
Resource:

- Windows computer

Description: It's much easier to comprehend how voluminous log files can become when you take some time to examine a few log files yourself. In this activity, you'll find some of the log files your own computer collects as you do your work. Note that this activity can be adjusted for any OS with a little research online to learn where to find that system's logs. The steps here are written for a Windows computer.
Complete the following steps:

1. Right-click **Start** and click **Event Viewer**. What categories of logs are listed in the left navigation pane? How many Critical event types are listed in the Overview and Summary pane for the past 24 hours? What about Error event types?
2. Expand the **Windows Logs** category and click **Application**. How many events have been logged here?
3. Most logged events are benign and do not require attention. Still, some of these events can provide insightful information about your system. Scroll through the list and find one or two you find interesting. What happened that resulted in the event being logged? Which application was affected?
4. Expand the following menus: **Applications and Services Logs** > **Microsoft** > **Windows**. How many log file subcategories would you guess are kept in the Windows subcategory?

Most of these subcategories are likely empty, containing no logs of relevant events to those topics. But some of the information available here can be mined for helpful information if you were troubleshooting an issue or searching for signs of problems. Complete the following steps to explore some of the log entries:

5. Expand the **Dhcp-Client** subcategory and click **Microsoft-Windows-DHCP Client Events/Admin**. Click through some of the logged events for the most recent successful network connection. What network name is reported in these events?
6. In the navigation pane, scroll down, expand the **Windows Defender** category, and click **Operational**. Scroll through these logs for any events labeled with a higher level than Information, such as Warning. Examine the information collected from these events. Is there anything going on with your computer that warrants further investigation? If so, what have you learned so far about the issue?
7. When you're finished, close the Event Viewer window.

Log Costs

As you can imagine, logs collect massive volumes of data in a very short time. One of the biggest expenses related to logs is long-time storage of all this data. Keeping log data for a period allows cloud admins to investigate problems during troubleshooting, conduct forensic analysis during or after a security breach, and generate required reports to prove compliance with standards and regulations. In many cases, companies are required to keep log data for a period of years to meet compliance standards.

Another significant expense related to logging is the indexing process for these logs. **Indexing** organizes and labels data in each log file for easy searching and analysis. This process is performed by an indexing engine and can require significant compute resources. Sometimes the expense of log management can be reduced by making trade-offs in what types and levels of indexing are performed.

Note 9-4 | An important part of the logging process is removing sensitive data from the logs before storing captured traffic. This step is called **log scrubbing**.

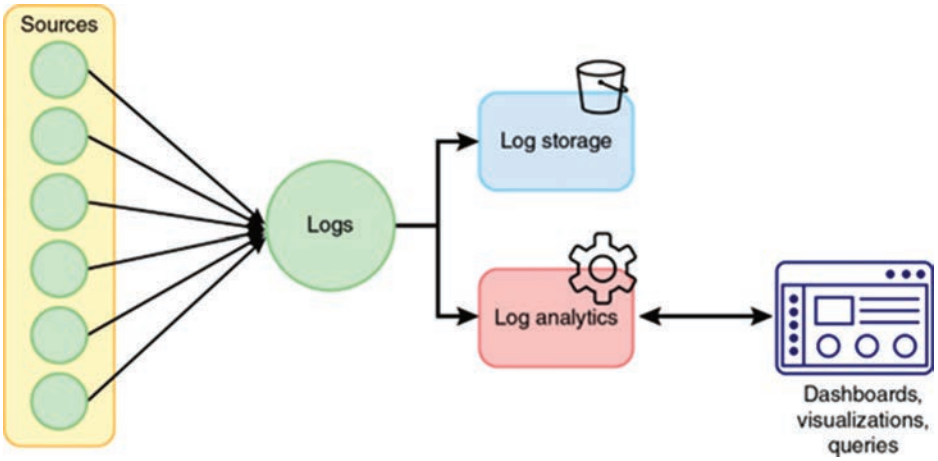
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In the end, logs need to be analyzed by some service. In AWS, the CloudWatch Logs Insights service provides built-in log analytics for AWS logs. Similarly, Azure Monitor Logs offers log analytics services, as does GCP's Log Analytics service. You can also pull in services from other vendors, such as Splunk or Rapid7, both of which offer cloud SIEM products in addition to other cloud monitoring and log analysis solutions.

Log Pipelines

To better understand how logs are collected, combined, and analyzed, it helps to explore the various phases that make up log pipelines. A **log pipeline** is a logical sequence and structure of individual processes that collect, process, enrich, and route log data from many sources to a central location for storage, analysis, and monitoring. At a basic level, the process requires that log files are collected from any monitored sources, these files are combined and sent for storage, and some of the data is extracted for analytics, monitoring, and visualizing in custom dashboards and other reporting tools (Figure 9-9).

Figure 9-9 Log collection, storage, and analysis



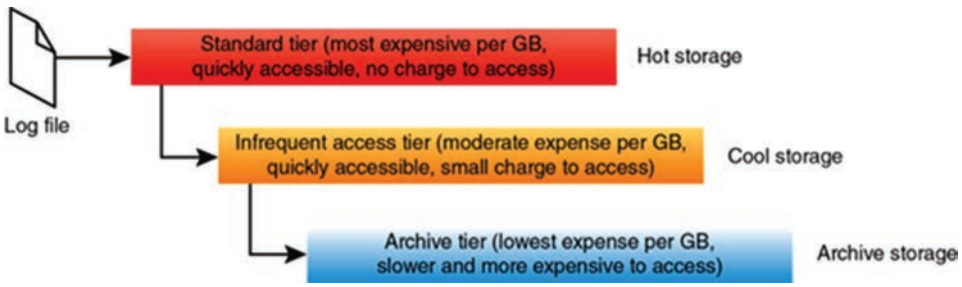
Logs must be handled in a way that tracks each source of the log data over time and makes this data consumable by a log analytics service. To do this work, log pipelines incorporate the following components and processes:

- **Log generation**—Each resource or service must be configured to generate the appropriate logs containing the desired information. Sometimes this is accomplished by turning on a logging service, such as AWS's VPC Flow Logs. Other times, an agent must be installed on the resource (such as with an EC2 instance) to collect the necessary logs and export them to the desired location.
- **Log format**—When enabling logging on a resource or service, the logs should be configured in a way that ensures data is formatted so it's compatible with the target log analytics service. Some formats are more structured with a standardized layout, such as naming each field inline with the data and using a standard separating character, such as a comma. Some formats are unstructured with fewer or no data labels but with a predictable order of data in each line; these logs might not contain standard separators. Other formats are proprietary. In these cases, logs might initially be formatted in JSON (JavaScript Object Notation) and then parsed into the target format. Log transformation is often performed by functions, such as AWS's Lambda, either at the time of collection or later in the pipeline.
- **Log transport**—Log files can be moved between systems with HTTP (Hypertext Transfer Protocol), gRPC (Remote Procedure Call from Google), or other protocols. Sometimes, logs are batched for easier processing, meaning groups of logs are sent together rather than sending one file at a time. This is part of the reason data from logs is usually delayed before it can be received and analyzed. Log transport also involves handling feedback from the target to ensure logs are being received and the target is not overloaded with incoming data.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Log aggregation**—Incoming logs are gathered by a target service in a process called log ingestion or log collection. This is the initial step in a larger process called **log aggregation**, which also might include processes to transform the logs to a standardized structure, filter them, and send them along to other services, such as storage or analysis. Not all logs are pushed to the target service automatically. Sometimes, the aggregation service, such as AWS's Data Firehose, must request—or pull—the logs from their sources. Logs can be transformed to a standardized format at this point. Other changes might include masking sensitive data (such as anonymizing login credentials) or data scrubbing (such as removing duplicates or correcting errors). Log sampling and log filtering can reduce the volume of data contained within each log file to help target more relevant information while reducing the time, space, and network bandwidth needed to handle, transport, and store the logs.
- **Log retention**—Once ingested, logs are retained for a specific period. Log retention might be dictated by compliance standards or other workload needs. Additionally, log files can be shifted between storage tiers to minimize costs while optimizing accessibility (Figure 9-10). For example, perhaps logs should be saved in a more expensive but readily available storage tier for a week or 30 days after collection. Then logs can be migrated to a less expensive storage tier for, say, a year, and finally, they are archived in the cheapest storage tier for the remainder of their retention period (such as 5 years). This storage life cycle varies according to the needs of the system, the types of data held in the logs, and the access patterns for these logs.

Figure 9-10 Storage life cycle for log files



Log Analysis

Analyzing logs is where the real magic happens—where insights are garnered that can alert to problems, guide updates, and influence changes. Logs provide perspective on your users and your systems:

- Logs can provide insights about your customers through clickstream analytics, which is data about the behavior and activities of application or website users. This information might be used to monitor system performance but can also be used to fuel marketing campaigns.
- Effective log evaluation can help with root cause analysis to identify the underlying source of problems and symptoms, thereby reducing your organization's MTTI (mean time to identification) and MTTR (mean time to resolve).

Note 9-5 | MTTR (mean time to repair) and MTTR (mean time to resolve) are two related but slightly different metrics. Repair can be accomplished more quickly and is completed when the problem has been fixed. However, resolution requires additional time and steps, such as making changes to the system to prevent the problem from happening again.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

One of the more common outputs of log analytics is dashboards. Several layers of services are used to make this happen. For example, in AWS, you can use OpenSearch to analyze and query data and use Grafana to organize outputs in customized dashboards. Another option is the **ELK (Elasticsearch, Logstash, and Kibana) stack**, which is a popular combination of services primarily used to explore log data:

- Elasticsearch is an earlier iteration of a search and analytics engine similar to OpenSearch. Where Elasticsearch used to be open source, it is now operated under a less permissive license. In its place, OpenSearch offers similar features with an open-source license.
- Logstash is also open source and performs data ingestion.
- Kibana is a data visualization and exploration tool. Like Elasticsearch, its license is no longer open source. In its place, some organizations prefer other options such as Grafana or OpenSearch Dashboards. However, each of these tools is designed to fill a specific niche in handling particular types of data or workloads. Sometimes one tool is not a good replacement for another.

Audit Trails

Using the information collected in event logs and system logs for security and fault management requires thoughtful data filtering and sorting. After all, you can't assume that all the information in these logs points to a problem, even if it is marked with a warning. For example, you might have typed your password incorrectly while trying to log on to your computer, thus generating a log entry. Still, sometimes seemingly innocuous information turns out to be exactly the data you need to diagnose a problem.

In some cases, you must be able to track every movement of every user. This is the accounting portion of AAA (authentication, authorization, and accounting). By referring to information stored in your logs, you should be able to answer the question, "Who did what activity when and in what way?" When tracking this level of information, the collective data is called an **audit trail**. The data in these logs is consistent and thorough enough to retroactively prove compliance and also to defensibly prove user actions (that is, your logs document user action in a way that is presentable in a court of law). This data is often used in forensics investigations to determine how a particular problem occurred, especially if criminal investigations are involved. Make sure you know exactly what types of actions and other events you must log on your network to meet relevant compliance standards.

Traces

Discussions of monitoring sometimes refer to the three pillars of observability, which are metrics, logs, and traces. Now that you've learned about metrics and logs, let's explore this third pillar, traces.

When you think about a network, you might consider the path of each individual packet as it traverses various devices and network segments. In the cloud, a similar process happens with each request sent between systems, resources, and services. Where metrics and logs can provide insights into individual systems, traces can shed light on the interactions between systems. A piece of information can cross many services in its journey from source to destination. **Tracing** is the process of tracking a request's end-to-end progress across a distributed system, which can involve many services. The outcome of this process is a trace, which is a record of the request's progress through the system. This information can be especially helpful when you need to troubleshoot a problem with any of the components in this system.

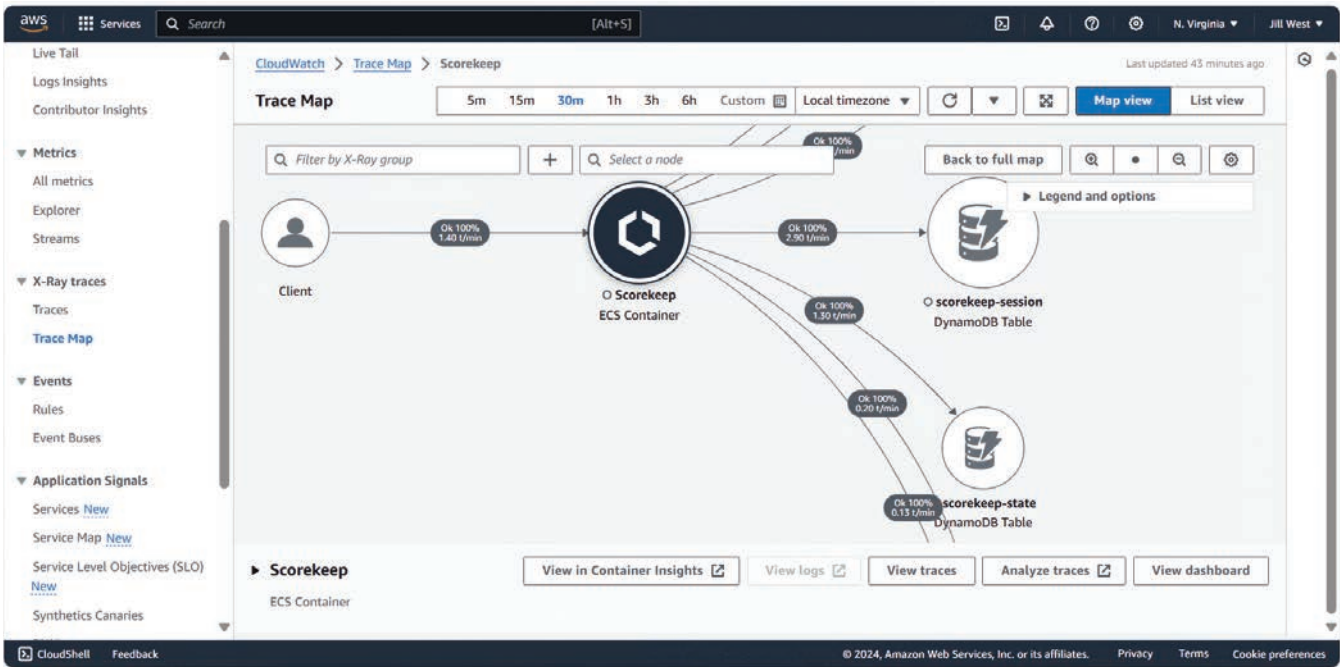
Suppose communication is breaking down as you're attempting to send requests to a Lambda function or logs are not appearing in a storage bucket where you anticipate them to be. In the cloud, you can't track cables from one device to another. But you can track requests. Rather than checking each intervening system, you can use a tracing tool to collect this information for you.

In AWS, each request is assigned a trace ID. Each service that interacts with the request sends the trace ID and information about the interaction to X-Ray, which then collects the information into a service map (Figure 9-11). The service map reveals which services are communicating with each other, which services depend on other services, and where communications are not performing as expected. Each node in the

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

service map can be further evaluated to dig deeper into trace information, evaluate error codes, and explore performance issues. X-Ray can additionally provide helpful information, such as the duration of a function or the parameters passed with a request, to help identify bottlenecks, debug serverless applications, or monitor costs and performance.

Figure 9-11 Sample service map in AWS's X-Ray



Source: Amazon Web Services, Inc.

Cloud-native and platform-native monitoring solutions typically provide the most efficient, effective, and elegant means of processing event, log, and trace information. The next section discusses various monitoring tools built into AWS, Azure, and GCP.

Self-Check

4. Which of the following is an example of an event?
   a. A VM's CPU utilization increases.
   b. A storage bucket is deleted.
   c. A VM experiences slow read rates from a storage volume.
   d. A ping fails.

5. Which process protects private information?
   a. Event correlation
   b. Triage
   c. Root cause analysis
   d. Log scrubbing

6. Which record would best reveal the location of a bottleneck across services?
   a. Trace
   b. Audit trail
   c. Event
   d. Log

☐ Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 9-3: Cloud Monitoring Services

Certification

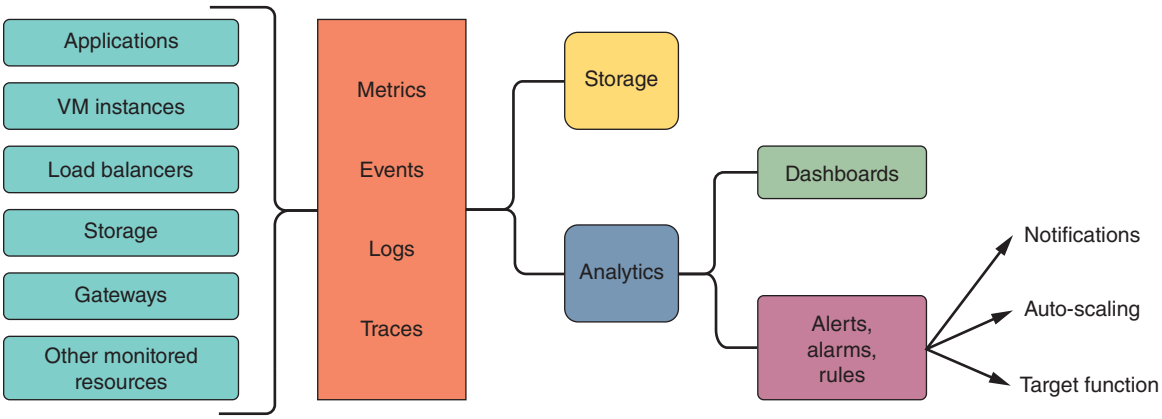
2.5 Given a set of requirements, provision the appropriate cloud resources.

3.1 Given a scenario, configure appropriate resources to achieve observability.

4.6 Given a scenario, monitor suspicious activities to identify common attacks.

Now that you've learned where performance data is collected from—including metrics, events, logs, and traces—let's consider what happens next to the data. Figure 9-12 illustrates how data flows through a typical cloud monitoring system. There are, of course, variations for each platform. This basic illustration can help you understand where the data can come from and then what can be done with it.

Figure 9-12 Cloud monitoring data collection and response



Dashboards provide a passive viewpoint for cloud admins to monitor cloud resources. This information might prompt further research or immediate human intervention. Alerts, alarms, and rules, however, can help automate responses and thereby relieve technicians of some of the time commitment required for cloud monitoring. In this section, you learn more about alerts, alarms, and rules in the AWS, Azure, and GCP environments.

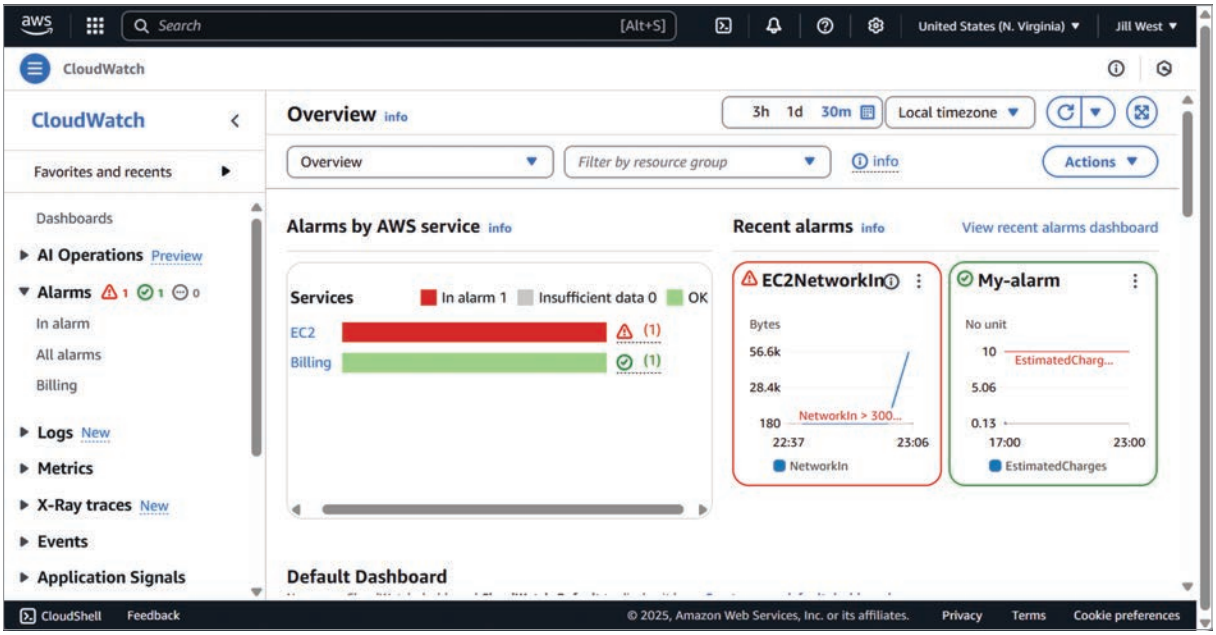
Monitoring in AWS

AWS relies primarily on the CloudWatch service to provide monitoring capabilities. Many of the available features are free and included by default. Metrics about every AWS service you use can be tracked through CloudWatch dashboards. However, you can also configure alarms that send notifications directly to cloud admins via email or text, which uses SMS (short message service). Additionally, notifications can automatically trigger configuration changes and other actions on your behalf.

An AWS alarm monitors a single metric over a specified period to determine threshold compliance. If the metric exceeds a threshold during that time, the alarm is activated and performs whatever response you programmed into it, such as sending a notification to SNS (Simple Notification Service) or enacting an auto-scaling policy. You can also display alarm states on your dashboards (Figure 9-13).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9-13 This dashboard shows one alarm in the alarm state



Source: Amazon Web Services, Inc.

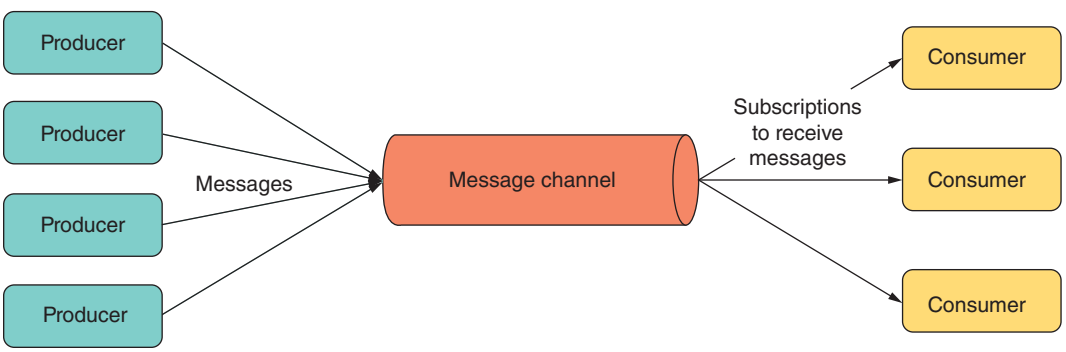
For example, suppose you collect information on an instance's CPU utilization every minute. If that utilization level exceeds a certain threshold for more than, say, five minutes, then the alarm might be configured to send a notification to SNS, which then sends an email to the cloud admin. If the CPU utilization spikes above the threshold for only one or two minutes, however, the alarm is not triggered. It's helpful to first understand the purpose and role of SNS notifications, and then alarm configuration options will make more sense. So let's dig into SNS.

SNS Notifications

An SNS notification can send text or email messages to human users. It can also support an event-driven architecture. Recall that microservices and serverless architectures increase efficiency of applications and encourage evolution of individual components by decoupling, or isolating, services. In this environment, publish/subscribe messaging (pub/sub messaging for short) allows the service components to talk to each other about events that happen within each service while the services themselves continue to function independently of one another.

Pub/sub messaging refers to asynchronous communication between event producers (such as monitored resources) and event consumers (such as a monitoring service). As illustrated in Figure 9-14, the event producer publishes messages, and the event consumer subscribes to message channels. Monitoring is a common use of this technique, but it's not the only one. The decoupling of monolithic applications into microservices, which you learned about in Chapter 3, "Cloud System Design," also relies on pub/sub messaging.

Figure 9-14 Consumers subscribe to event channel to receive messages from producers



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For example, suppose the building blocks of a monolithic (that is, large and complex) application are separated into various, independently run microservices. Maybe you have a database, a web-based user interface, an authentication service, perhaps a payment platform, and a variety of other microservices that, collectively, make the app work. Each of these components runs in a separate environment, and each can be evolved and maintained separately due to their reliance on API-based communication, as you've already learned. Each microservice might be an event producer that broadcasts, or publishes, a message when an event occurs, such as a customer purchasing an item. Another microservice might subscribe to the first service's messaging channel. When the event consumer receives a message meeting certain requirements, it responds with its own action, such as generating a ticket to pull the purchased item from warehouse shelves for shipping to the customer.

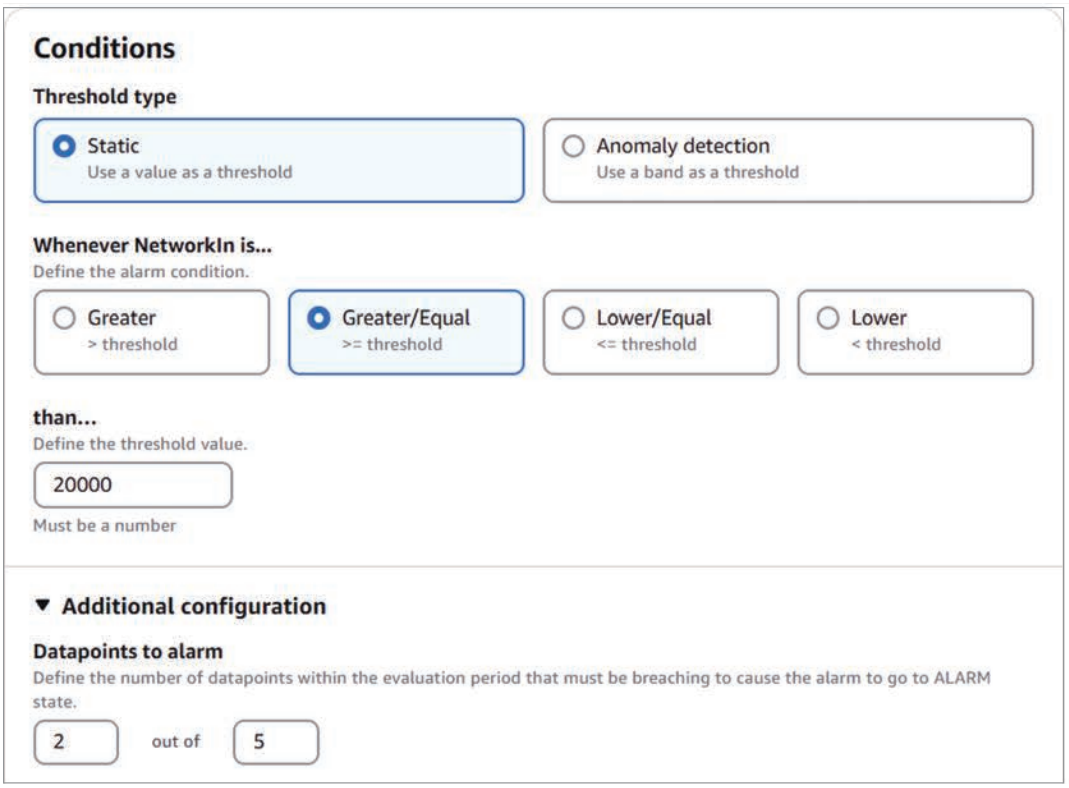
As events occur within each component, communication must be triggered that keeps all related components in sync. The component experiencing the event publishes a message to all subscribers of that particular channel, which is called a "topic" in SNS. Subscribers might include human users but can also include other components of the application or related resources. Each event serves as part of a larger process that emanates throughout the decoupled application. Recall that this arrangement is called event-driven architecture, where an event in one service triggers an event in another service, thus enabling communication throughout the entire architecture. Clearly, SNS is much more than a notification service for humans.

CloudWatch Alarm Configuration

CloudWatch relies on the same SNS system to handle messages in response to alarms. Figure 9-15 shows some of the configuration options for an alarm threshold. Notice that you can select a logical argument such as the following:

- >= (greater than or equal to)
- <= (less than or equal to)
- > (greater than)
- < (less than)

Figure 9-15 Configure a threshold and an action for the alarm to take

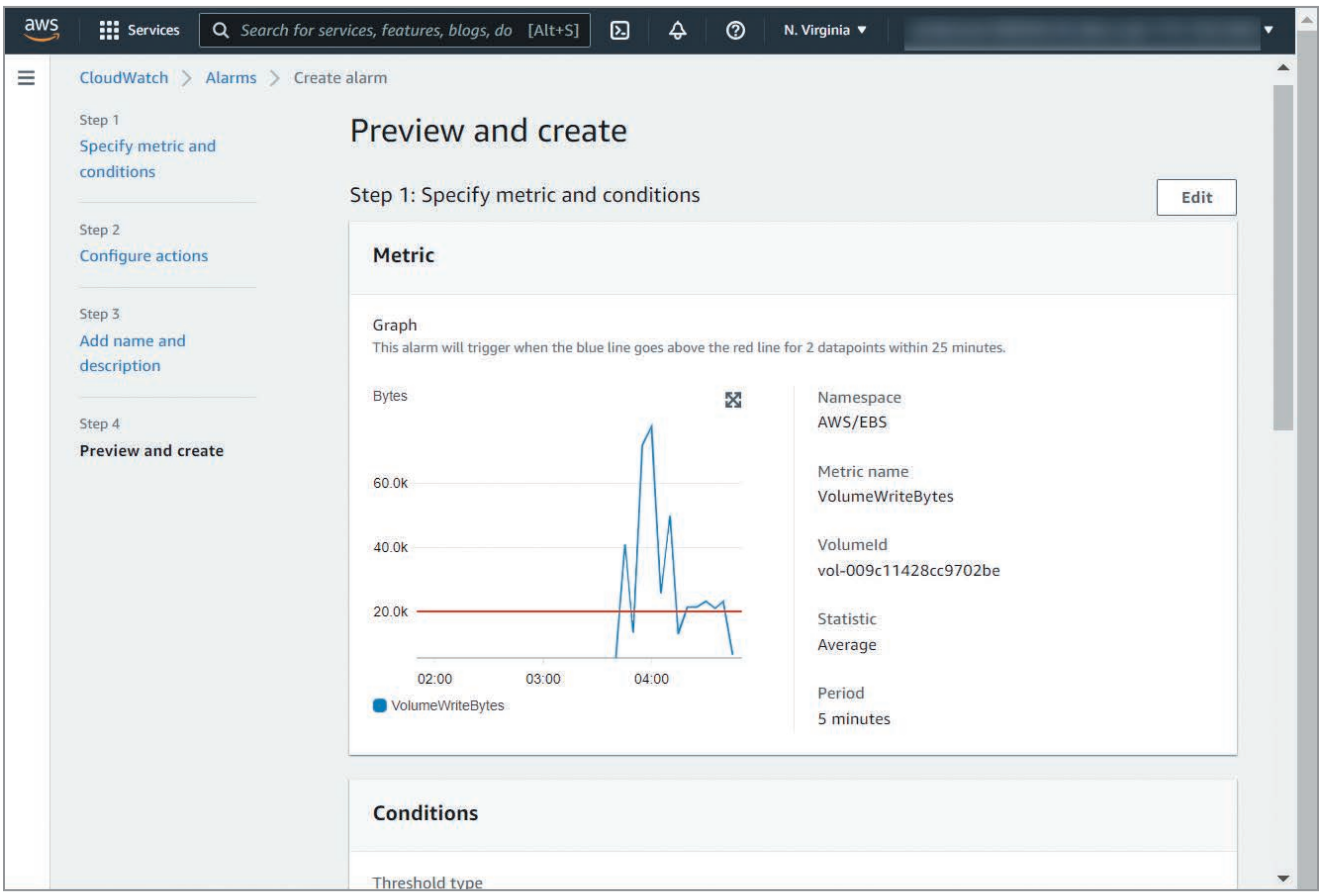


Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9-16 shows where the displayed threshold configuration falls on the metric's existing data. The red line indicates the alarm's threshold. You can further restrict the alarm to a number of occurrences within a specified range of data points. For example, perhaps you only want the alarm to trigger if two of five data points exceed the threshold, but not when just one data point rises above the threshold.

Figure 9-16 The graph's straight red line indicates the metric threshold



Source: Amazon Web Services, Inc.

Notice that you can also set various types of actions. A notification action might send an email or text to the cloud admin. An auto-scaling action, on the other hand, could stop, terminate, reboot, or recover an instance. For example, suppose you have an alarm configured to detect CPU utilization below 10 percent for more than three hours. This alarm is designed to detect instances that are not currently being used. The alarm could trigger a stop instance action, which shuts down the instance until it's needed again.

How do you know what alarm thresholds to configure, that is, at what point should the alarm be triggered? This is really an art form rather than a science. With more experience, you'll get a better feel for what settings and thresholds miss important changes (that is, false negatives) or, conversely, what alarm configurations overwhelm your team with too many false positives. You'll come to understand the nuances of your cloud and integral systems, learning when problems are most likely to occur and what kinds of indicators give you an early warning that can help you avoid problems in the first place.

As you begin configuring alarms, first gather information on target object baselines. A baseline indicates current functioning and configuration of the systems in your network. This includes collecting information on KPIs such as CPU and RAM usage, page load time and error rates, storage and network usage, and patch and application versions. What's normal for each object's metrics? At what point does deviation

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

from those baselines indicate a true problem? How long should that deviation be in play before you want to receive a notification or make an automated change? Additionally, consider the following tips for alarm configuration:

- **Fine-tune your alarm thresholds**—If your acceptable band of values is too narrow, you'll have loads of false positives. If the acceptable band is too wide, you'll likely miss important indicators of problems creeping in. Watch the visualizations and metrics closely when setting thresholds to determine if you're satisfied with the resulting level of reporting. Also, adjust your thresholds as needed for existing and anticipated conditions. For example, if you're updating a server, you can expect increased traffic across connected network interfaces. You don't want this planned traffic to unnecessarily trigger alarms.
- **Configure granular or escalated notifications**—For example, assign to first- or second-tier IT staff the more sensitive alarms that are more likely to result in false positives or the low-priority alarms that require attention but little troubleshooting. Alarms that indicate more complex problems or that are monitoring more critical services can send notifications directly to more experienced staff and admins. The people handling alarms will also need policies that clearly define how issues should be communicated to other staff, if needed, such as when a first-tier employee needs to escalate the alarm to someone more qualified to handle a difficult problem.
- **Choose notification types that resonate with staff**—For example, if you're consistently overwhelmed by emails but your phone never leaves your side, you might prefer text messages. If you tend to forget a text message moments after you read it, emails might work better. That said, never set up an email filter for the notifications coming to your inbox unless those messages are going to a folder that you check very frequently and you receive notifications on your phone when emails arrive to that folder. "Out of sight, out of mind" presents a significant threat to follow-through on alarm notifications.
- **Include backup monitoring for critical services**—Just like with other parts of the infrastructure, avoid relying on a single point of failure for monitoring systems. A single monitoring solution can fail, leaving crucial red flags unnoticed until a pending failure starts costing users productivity time and costing the company money that could be better used for other purposes.
- **Explore monitoring tools and resources from third-party vendors**—Cloud monitoring is a crowded market with several excellent tools available to adaptively monitor resources, cut costs, and optimize services. Several third-party providers post these kinds of solutions to the AWS Marketplace. For example, ParkMyCloud monitors and helps reduce cloud costs by identifying ways to optimize a cloud budget. Datadog, a popular solution that works across nearly any cloud platform and on-prem, collates data from servers, containers, databases, third-party services, and many other sources into a single location—called a single pane of glass—with sophisticated monitoring algorithms.
- **Configure monitoring tools throughout the cloud stack**—For example, APM (application performance management) tools monitor and can, in some cases, help manage applications for maximum performance and reliability. While AWS's CloudWatch provides some APM features, many third-party tools provide more in-depth functionality for monitoring applications. Examples include Stackify, Raygun APM, and New Relic APM.
- **Consider that alarms can't monitor every possible indicator of a problem**—Sometimes problems that crop up are unpredictable and not indicated by the thresholds you are currently monitoring. If you regularly skim through your logs, reports, and dashboards, you might notice a problem that is not triggering an alarm but should (that is, a false negative). When a problem does occur, you can do some forensic research to determine what early indicators you missed and how you might configure alarms to detect those red flags sooner. Also, these performance anomalies can sometimes be captured by third-party tools designed specifically for this purpose, such as Watchdog by Datadog. These tools rely on ML technologies to detect problems you might not think to search for.

Note 9-6 | These alarm configuration tips apply to other cloud platforms as well, although the features might be called something else, such as alerts, rules, or policies.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CloudWatch Events and Rules

Whereas alarms respond to continual metric changes (such as the rise and fall of CPU utilization), event rules respond to events (such as when an EC2 instance shuts down or an object in a storage bucket is deleted). When the specified event occurs, a rule is triggered that can initiate a corresponding action in a different service. This could be a simple response, such as pushing an SNS notification. Or it could be something more complex, such as creating a new VM from a snapshot. Event rules are often configured to trigger Lambda functions. Lambda is a serverless computing service in AWS that can perform various compute functions without requiring the configuration or maintenance of underlying VM instances. Lambda offers a convenient method to insert some basic scripting or code to implement changes or other processes within pipelines and other automations.

Monitoring in Azure

The Azure Monitor service collects four kinds of data on the resources it monitors. Notice the correlation to the three pillars of observability, along with one additional data type:

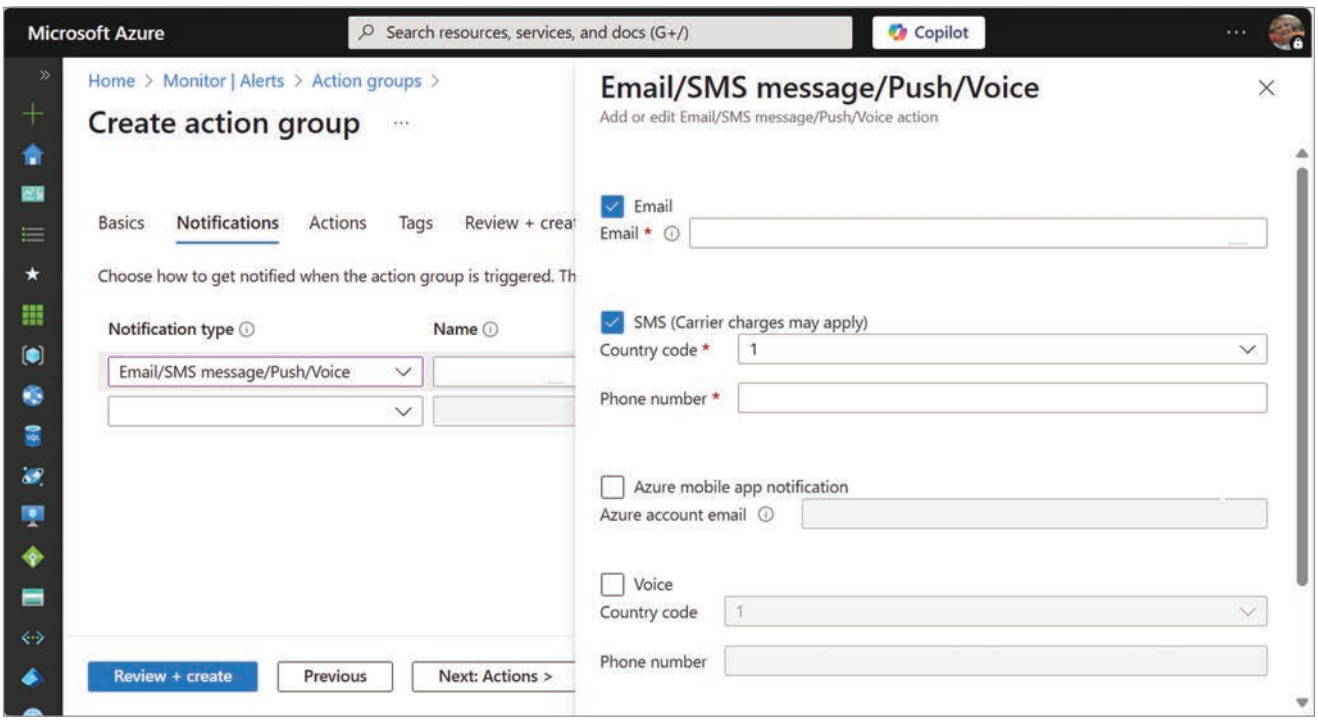
- Metrics consist of lightweight, numerical data collected in real time and reported live in visualizations.
- Logs are more complex data, including both numerical and text data types, describing system events or activities and collected over time for retroactive analysis.
- A trace documents a request's path through services and resources.
- A change is a series of related events that can provide useful insight when troubleshooting issues.

This data is retained at no charge for a few weeks, depending on the type of data and the current pricing scheme. If you need to keep data longer than that, you can send it to a storage account for paid storage.

Alerts can trigger when specific thresholds are breached or when certain events occur. An alert activates some sort of response. For example, an action group can notify cloud technicians. An autoscale response can adjust the number of resources handling a workload. And an AIOps (AI for IT Operations) response can act independently based on data provided to it to automate actions, identify issues, or detect anomalies.

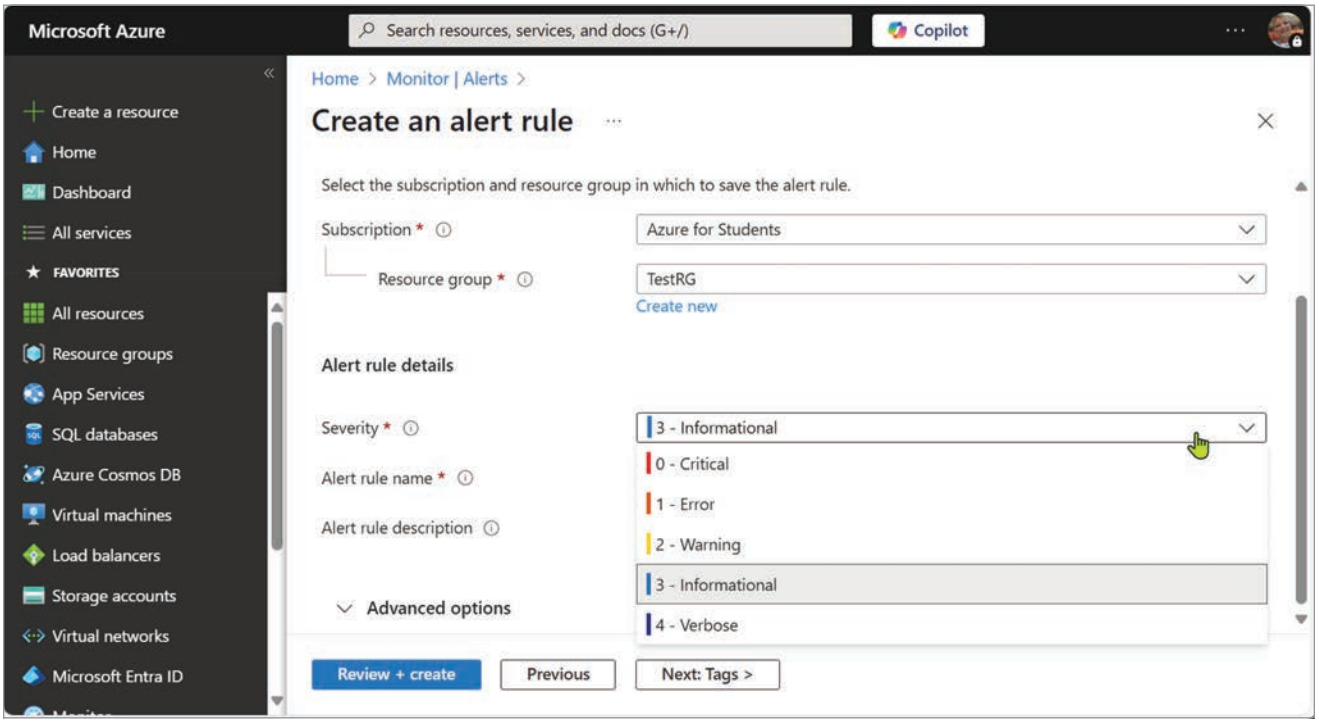
Figure 9-17 shows the configuration blade for creating a new action group and the sub-configuration for an Email/SMS/Push/Voice action type. Figure 9-18 shows available severity levels for a new alert rule.

Figure 9-17 Configure a notification via email, text message, Azure app, or voice message



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9-18 Severity levels help with prioritizing alerts



Monitoring in GCP

GCP offers monitoring options through a variety of services:

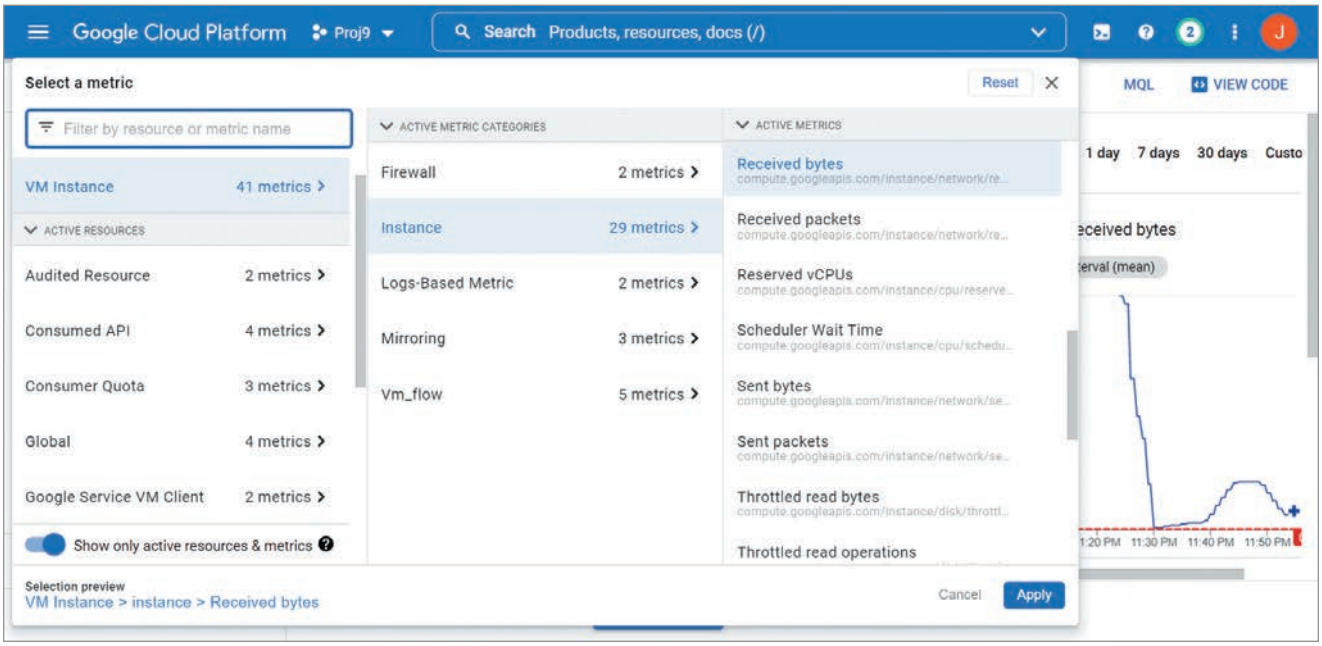
- **Cloud Monitoring**—Collects basic system metrics by default and can also be used to monitor resources in hybrid and multi-cloud environments.
- **Cloud Logging**—Collects logs from across services and can generate metrics based on those logs.
- **Error Reporting**—Analyzes logs and aggregates errors detected in the log data.
- **Cloud Trace**—Tracks requests across applications to support troubleshooting.

As shown in Figure 9-19, you can create an alerting policy on any number of available metrics and, when violated (Figure 9-20), the alert will send a notification. Notification channels, such as email messages or texts via SMS, transmit information about triggered alerts.

Other notification types are available through integrated third-party services, such as Slack, and are configured through the Alerting page in the Monitoring service (Figure 9-21).

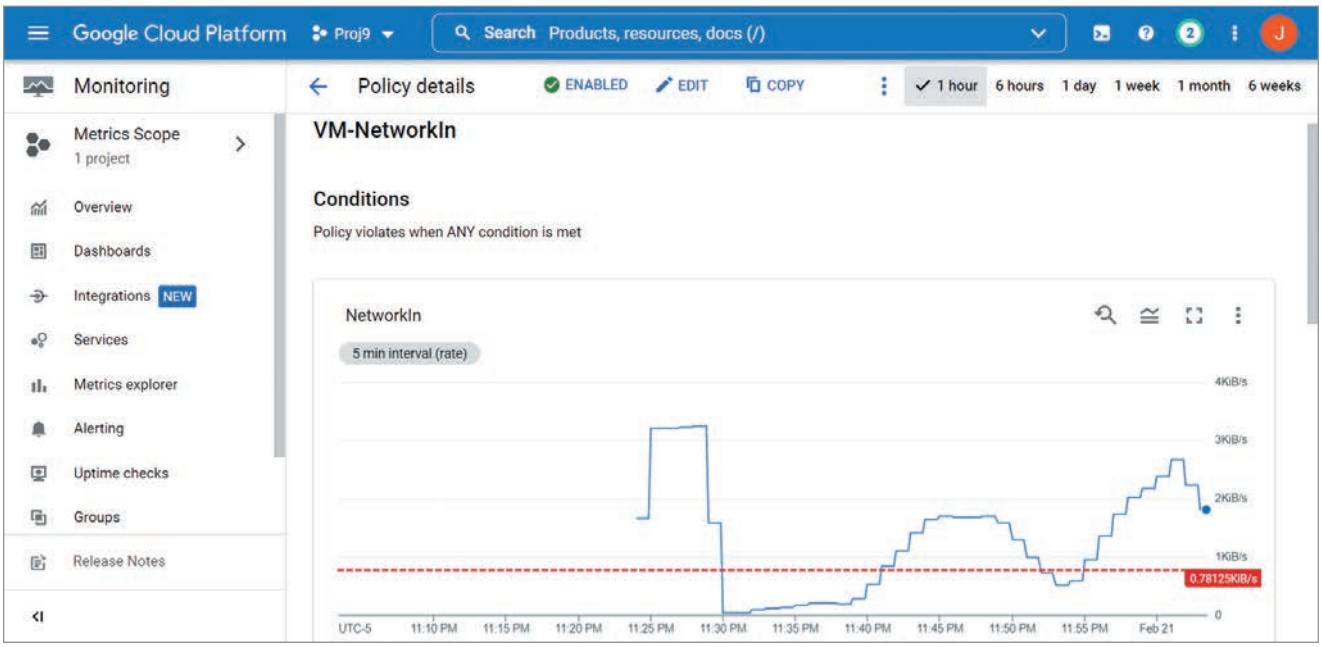
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9-19 VM instance metrics for monitoring



Source: Google LLC

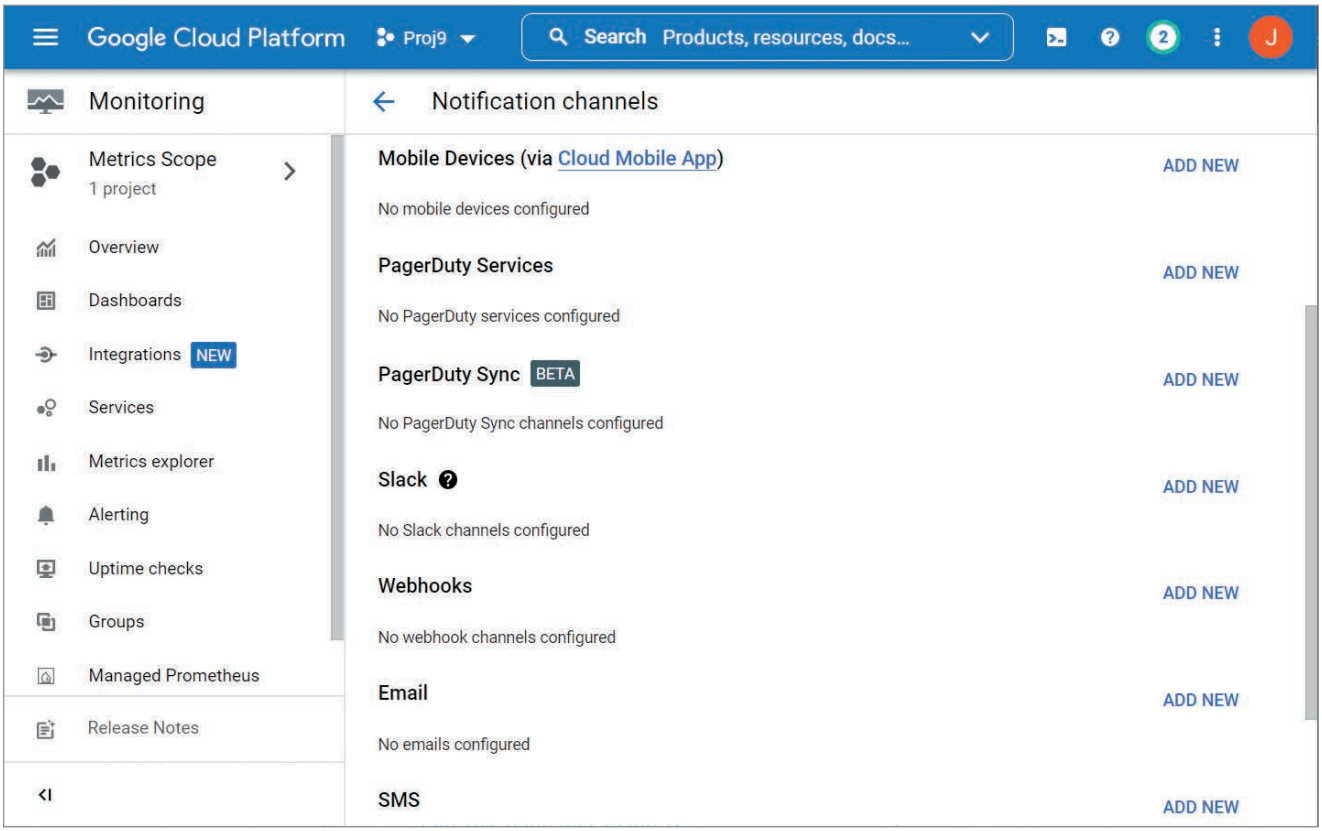
Figure 9-20 This alerting policy fired and sent a notification



Source: Google LLC

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9-21 Add a notification channel



Source: Google LLC

Self-Check

7. An event producer _____ messages.
   a. subscribes to
   b. analyzes
   c. responds to
   d. publishes

8. What kind of application relies on event-driven architecture?
   a. Monolithic
   b. Decoupled
   c. Authentication
   d. Independent

9. What does the symbol >= stand for?
   a. Less than
   b. Greater than
   c. Less than or equal to
   d. Greater than or equal to

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 9-4: Troubleshooting Problems and Performance

Certification

1.8: Summarize cost considerations related to cloud usage.

3.2: Given a scenario, configure appropriate scaling approaches.

3.3: Given a scenario, use appropriate backup and recovery methods.

5.2: Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.

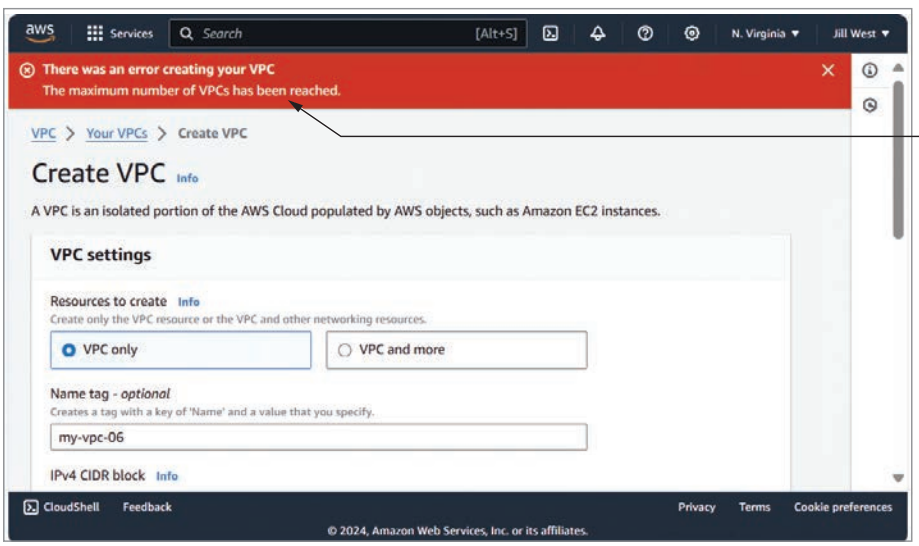
6.1: Given a scenario, troubleshoot deployment issues.

Despite all preparations, problems still occur. Monitoring helps identify symptoms of problems. How can you use this information to find the source of a problem? And how do you respond to the problems that surface?

Quotas and Throttling

One aspect of your cloud deployment that you will need to monitor is your approach toward limits placed on your deployments by your cloud provider. To help manage costs, increase security, and ensure the anticipated performance of underlying infrastructure, most CSPs incorporate service quotas on most of their services and resource types. A **service quota** is the maximum number of resources or the maximum value for a feature that each customer can deploy within a single account. For example, in AWS, you can only create up to five VPCs within each region (Figure 9-22), and you're limited to a certain number of network interfaces per EC2 instance (this number varies by instance type). Some of these service quotas can be changed by special request while others are limited due to the nature of the resource.

Figure 9-22 The VPC service quota cannot be exceeded



Error message when attempting to create a sixth VPC in this region

Source: Amazon Web Services, Inc.

It's important to understand how service quotas can affect your cloud deployment. If you're having trouble with auto-scaling or with creating a new resource, service quotas could be the culprit. In AWS, you can use the Service Quotas service to monitor and manage service quotas across your deployment.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A similar limitation is **API throttling**, which restricts the number of API requests allowed to each customer's APIs. These limits can help ensure service performance and availability. They can also protect cloud customers from expenses caused by misconfigurations or caused by attacks intended to artificially increase costs.

If a high-traffic API is producing error messages, check the API throttling configurations to ensure your system is not attempting to exceed these limits. You might need to request an increase to your overall API quota per region, you might be able to redistribute your allotments across all your APIs in that region, or you might consider moving one or more of your APIs to a different region or a different account. Keep in mind, however, that you cannot count on API throttling as a guaranteed limit—in AWS, for example, API Gateway works to throttle requests on a best-effort basis.

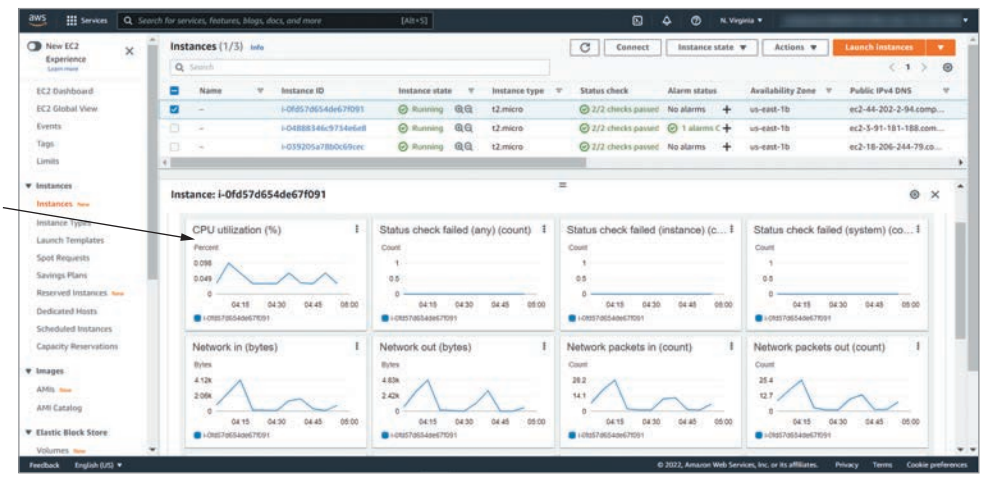
The limits imposed by service quotas and throttling can be a good reason to consider creating multiple accounts for handling various workloads. Each account's service quotas are isolated from other accounts, even if they're owned by the same organization. A multi-account strategy can also increase security by reducing the so-called "blast zone" of a security incident to the resources and services in a single account, and by isolating access controls to help enforce the principle of least privilege. Using different accounts can also increase accounting efficiency for billing to the most relevant department, project, or customer.

Utilization Issues

A significant portion of the monitoring you will do in the cloud relates to how resource utilization directs resource sizing to support performance metrics. The information you collect will inform your work in troubleshooting performance degradation issues. The following list explores some tips for improving performance on various factors:

- **CPU**—CPU resources are monitored by a CPU utilization measurement (Figure 9-23). Very low CPU utilization could indicate you have too many instances running, and you might want to configure the cluster to automatically reduce the number of servers in the cluster. Very high CPU utilization (typically over 85 or 90 percent) indicates the need for scaling out the cluster by adding more instances. Recall that this technique is called rightsizing—resources are increased or decreased according to need to balance performance and cost.

Figure 9-23 CPU utilization monitored in EC2



Source: Amazon Web Services, Inc.

- **GPU**—When using a GPU, the GPU utilization metric can be monitored in much the same way as CPU utilization. This metric might trigger an autoscaling action, or it might trigger an alert to a cloud admin. You can also monitor metrics for a GPU's memory utilization, temperature, or power usage.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Memory**—Memory utilization is another critical metric to monitor clusters for too little or too much demand on the existing instances; memory utilization problems can also indicate misconfigurations, bugs, or other errors. Unlike other metrics, typical cloud monitoring services (such as CloudWatch) don't offer memory monitoring for VM instances because they don't have the necessary visibility into processes the instance is running within its own OS. To monitor memory utilization, the instances themselves must push RAM utilization information to the monitoring service. Some cloud providers, such as AWS, offer prebuilt scripts you can run on each instance where this configuration is needed. Additionally, when running clusters of managed servers such as for containers, memory utilization can be monitored via built-in agents.
- **Storage**—Commonly monitored storage metrics include the following:
  - IOPS (input/output operations per second), which indicates speed or throughput. IOPS might be calculated from DiskReadOps (disk read operations) or DiskWriteOps (disk write operations).
  - Available capacity, which indicates the amount of storage space still available on a storage volume. This information is especially relevant to determining storage costs and checking remaining capacity in cases where storage space must be paid for when it's reserved, even if it's not in use (such as with block storage).

Scaling and Sizing Issues

Even in the public cloud, it's possible to run into limitations of the physical hardware underneath the virtual layer. Many cloud providers, especially for the public cloud, use oversubscription, or overprovisioning, to increase their customer base faster than their physical infrastructure can support. This strategy is common even within a company's own data center where, for example, you might deploy more storage space than you have available on your physical storage devices. You then increase the physical storage only as your need for storage space nears your hardware's current capacity.

Additionally, oversubscription can cause problems for you as a public cloud customer if many customers draw on the CSP's physical infrastructure at the same time beyond typical bandwidth levels. In this case, you might not receive actual performance levels consistent with your realistic expectations. One way to avoid this problem is to ensure you choose a reputable public cloud provider who is known for maintaining sufficient physical infrastructure to meet fluctuating customer demand. Another option is to replicate some of your resources across the cloud provider's physical data centers so that, if one data center is overwhelmed, your services can run on duplicate resources in another data center within the same cloud platform.

One of the most common methods of balancing resource costs with performance is to configure auto-scaling. These configurations, however, can cause problems if scaling doesn't happen responsively to changing demands. The following list describes some common problems you might face with auto-scaling clusters of VM instances:

- **Conflicting policies**—If auto-scaling is not performing as expected, check to confirm you don't have multiple policies being triggered at the same time. For example, if one policy says to scale out but another policy says to scale in on the same trigger, you might get results you didn't want and, therefore, degrade performance. In this case, you can check the related alarm's history or the auto-scaling group's history to determine if two auto-scaling policies are being triggered simultaneously.
- **Inadequate policy**—If an auto-scaling group's minimum or maximum capacity has already been reached, additional instructions to scale in or out, respectively, won't work. If this limitation is preventing auto-scaling, the thresholds might need to be adjusted.
- **Stuck instance**—Even if an instance was automatically marked for creation or termination, occasionally an instance will get stuck in a Pending:Wait or Terminating:Wait state. This can distort auto-scaling group metrics. The instance should eventually time out, but you can push the process from the CLI (command line interface), if needed.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Related to auto-scaling, automatic replication is designed to increase performance or increase availability and redundancy. Similar to auto-scaling, replication processes can snag or fail. The following list examines some common replication issues in relation to auto-scaling processes:

- **Incorrect role or permissions**—To create replicated resources, the service performing the replication must have appropriate permissions through an assigned role. If the service is attempting to replicate while assigned the wrong role or other permissions, the replication will fail.
- **Insufficient access**—If the replica is supposed to be created in a different account and the service does not have permission to create or upload resources into that account, replication will fail.
- **Missing or incorrect tags**—Resources intended to be replicated must be appropriately tagged at the time of creation. If the creation process does not assign tags during creation or is using the wrong tags, replication will fail.

Restoration from Backup

When things do go wrong, you'll need processes for restoring data and functions from your backups. Most of the time, backups rely on redundant resources of some kind, whether that's a redundant web server or redundant copies of data. The options for restoring that data vary in whether the data is deployed to the same hosts or to replacement hosts, as follows:

- An **in-place recovery** restores backup data to the same resources that hosted the original data. Perhaps the original data was damaged in some way or the host application had a problem and needed to be restarted. The backup copy replaces the original data on the same server cluster or database. For this to work, the original hosts must be uncompromised, and the backup data must be transferred to those hosts, which might require system downtime.
- A **parallel recovery**, also called a side-by-side recovery, restores a system using backup data that is hosted in a different location, such as a replacement server cluster or database. This recovery process can be completed more quickly because the backup copy can be accessed with a simple configuration change that redirects traffic to the backup location. However, it's typically more expensive and more complex to implement. This option is well suited to disaster recovery strategies where downtime is undesirable.

Many cloud platforms offer built-in backup services. For example, AWS Backup provides managed backup protection throughout your AWS environment. Backups can be automated, monitored, and encrypted. Azure Backup similarly offers backup management throughout many Azure services. GCP also offers a comprehensive managed backup service called Backup and DR Service in addition to backup features within specific services, such as Backup for GKE (Google Kubernetes Engine) and Persistent Disk snapshots.

Self-Check

10. Which technique adjusts resources to match demand?

a. Throttling
b. Rightsizing
c. Oversubscription
d. Replication

11. Which of the following problems could result in an orphaned resource whose expenses are not properly tracked?

a. Incorrect role
b. Conflicting policies
c. Missing or incorrect tags
d. Stuck instance

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

12. Which of the following statements applies to an in-place recovery?

    a. Replaces damaged data with backup data to the same host.
    b. Is well suited to disaster recovery strategies where downtime is undesirable.
    c. Restores a system using backup data that is hosted in a different location.
    d. Accesses a backup copy of the data with a simple configuration change.

☐ Check your answers at the end of this chapter.

Summary

Section 9-1: Monitoring Resources

- Observability is a systematic design that ensures transparency of activities and resources in a distributed system that enables monitoring and automated responses.
- In a cloud environment, where you might have difficulty taking inventory of all cloud resources and tracking resources throughout their various life cycle phases, tagging can help identify and monitor this expansive, abstract infrastructure.
- Some monitoring processes, such as logs, offer retroactive analysis of past events; other monitoring processes provide intermittent insights into a resource's state or performance; and some monitoring tools support continuous monitoring, which means the tool can provide near-immediate feedback on performance or security metrics in a continuous, unbroken manner.
- A well-structured dashboard can give a cloud technician a quick overview of critical services and draw attention to issues that might indicate problems—too much information in a dashboard causes confusion and overwhelm, and too little information decreases relevance.

Section 9-2: Events, Logs, and Traces

- By identifying relationships between events through event correlation, monitoring tools can identify event patterns or anomalies that indicate performance or security problems.
- Logs collect streams of data to show historical insights and reveal usage patterns and anomalies over time.
- A log pipeline is a logical sequence and structure of individual processes that collect, process, enrich, and route log data from many sources to a central location for storage, analysis, and monitoring.
- Tracing is the process of tracking a request's end-to-end progress across a distributed system, which can involve many services.

Section 9-3: Cloud Monitoring Services

- AWS relies primarily on the CloudWatch service to provide monitoring capabilities.
- Pub/sub messaging refers to asynchronous communication between event producers (such as monitored resources) and event consumers (such as a monitoring service).
- The Azure Monitor service collects four kinds of data on the resources it monitors: metrics, logs, traces, and changes.
- GCP offers monitoring options through the following services: Cloud Monitoring, Cloud Logging, Error Reporting, and Cloud Trace.

Section 9-4: Troubleshooting Problems and Performance

- A service quota is the maximum number of resources or the maximum value for a feature that each customer can deploy within a single account.
- API throttling restricts the number of API requests allowed to each customer's APIs, which can help ensure service performance and availability.
- A significant portion of the monitoring you will do in the cloud relates to how resource utilization directs resource sizing to support performance metrics.
- Common auto-scaling problems include conflicting policies, inadequate policies, stuck instances, incorrect roles or permissions, insufficient access, and missing or incorrect tags.
- An in-place recovery restores backup data to the same resources that hosted the original data, while a parallel recovery restores a system using backup data that is hosted in a different location.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Key Terms

For definitions of key terms, check the Glossary.

agent
API throttling
audit trail
ELK (Elasticsearch, Logstash, and Kibana) stack
event correlation
false positive
indexing
in-place recovery
ITSM (IT service management)
log
log aggregation
log pipeline
log scrubbing
observability
parallel recovery
pub/sub messaging
resource metering
service quota
tag
tracing
triage

Acronyms Checklist

The following acronym is listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use this acronym in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what the acronym stands for and the general concept of the term itself. This acronym is used in context in this chapter.

ELK Elasticsearch, Logstash, and Kibana

Review Questions

1. Which CPU utilization measurement most likely does *not* indicate a problem?
   a. 100 percent
   b. 95 percent
   c. 50 percent
   d. 5 percent
2. Which accounting strategy tracks cloud usage but not cloud costs?
   a. Log monitoring
   b. Chargeback model
   c. Showback model
   d. Event monitoring
3. When you create a tag called Project on all your EC2 instances, what kind of tagging are you using?
   a. Implicit
   b. Inherent
   c. Native
   d. Explicit
4. Which visualization would be the best fit for a KPI whose status fluctuates significantly and quickly?
   a. Dial
   b. Line chart
   c. Stacked chart
   d. Pie chart
5. How can you get deeper insight into your VM server's OS metrics?
   a. Configure logging.
   b. Install an agent.
   c. Customize your dashboard.
   d. Add more notifications.
6. What kind of service can help you identify the location of a bottleneck in your application's performance?
   a. Resource metering
   b. Log retention
   c. Log aggregation
   d. Tracing
7. Which of the following solutions can you use to generate custom dashboards to display data collected in logs?
   a. Grafana
   b. Logstash
   c. Root cause analysis
   d. Lambda
8. What information can help you determine where to set the threshold of an alarm monitoring a VM server?
   a. Audit trail
   b. Resource baseline
   c. Log pipeline
   d. Dashboard

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9. What technique is the best approach to help control the cost of running an application that is exposed for public usage?
   a. Resource metering
   b. API throttling
   c. Service quota
   d. Parallel recovery

10. Which phase of a log pipeline is most concerned with data life cycle configurations for cost optimization?
   a. Log aggregation
   b. Log analysis
   c. Log retention
   d. Log generation

Scenario-Based Questions

Question 9-1

Heidi is investigating a connectivity problem involving AWS EC2 instances running in a particular region. The problem began about three weeks ago and has resulted in repeatedly rebooting or recovering the server instances in this cluster. The servers appear to be caught in a cycle where the system's automated attempts at solving the problem seem to trigger the problem all over again. Heidi wants to go back to the first few error messages collected, which might help her determine what started the problem in the first place. What sort of data does Heidi need to access to examine this information?

a. Metric reports
b. Event logs
c. EC2 metrics
d. EC2 analytics

Question 9-2

Rex is managing a cluster of four web servers with a load balancer on the front end. His company's website has recently started receiving a significant increase in traffic due to a successful advertising campaign, and so he's ready to increase the budget available for running this cluster of web servers. However, Rex needs to decide whether to scale up or scale out. Currently, he has three reserved instances running in the cluster and one on-demand instance. He needs to improve overall performance capacity by about 25 percent. What approach should Rex take to reach his new capacity goal while minimizing the impact on his budget?

a. Add another reserved instance to the cluster.
b. Allocate more infrastructure resources to all four existing servers.
c. Add two more on-demand instances.
d. Allocate more resources to the one existing, on-demand server.

Question 9-3

Kylie works for a retail chain that hosts its own e-commerce site. The site relies on four databases: one that holds the product catalog, one that stores historical order data, a third that processes payments as purchases are made, and finally, a database that stores customer loyalty credits and advertising campaign information. To minimize disaster recovery costs while increasing her company's preparedness, Kylie has decided to invest in an upgraded failover service for one database to reduce RTO and RPO for that data to seconds. Which database should she choose?

a. The historical orders database
b. The product catalog database
c. The customer loyalty database
d. The payment processing database

Hands-On Projects

Note 9-7 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Project 9-1: Monitor an EC2 Instance in AWS

Estimated time: 45 minutes
Objective 3.1: Given a scenario, configure appropriate resources to achieve observability.
Resources:

- AWS account
- Access to an active email account
- Internet access

Description: Generating notifications in response to an alarm is a great way to familiarize yourself with some of the basic monitoring services and options in AWS. In this project, you'll configure an alarm for a Windows VM and then trigger the alarm.

Note 9-8 | At the time of this writing, all the services used in this project offered sufficient free tier limits to complete the project with no fees so long as you follow the instructions as written. To display a current list of free units included with AWS CloudWatch, search online for an Amazon CloudWatch pricing list, and check the "Free tier" portion of the pricing scheme. At the time of this writing, this information could be found at aws.amazon.com/cloudwatch/pricing.

Complete the following steps:

1. In your AWS Management Console, create a VM using the following details:
   a. Choose a free tier eligible version of Windows Server (such as **Microsoft Windows Server 2022 Base**) and the **t3.micro** instance type. Make sure to choose a combination of Windows Server AMI and instance type where both options are free tier eligible.
   b. You can associate a key pair with the VM or not—you will not need to remote into it.
   c. In the Network settings section, add a security group rule that allows **All ICMP - IPv4** traffic from **Anywhere**.
   d. All other configurations can be left at the default setting.
2. Before leaving the EC2 service, make note of the VM's instance ID. You will need this information later in this project.

Before you create an alert, you need an SNS topic, which will define where information about the alert will go. To create an SNS topic, complete the following steps:

3. Go to the **Simple Notification Service**, which is grouped in the Application Integration category. In the navigation pane on the left, click **Topics**. Click **Create topic**.
4. Select the **Standard** topic type. Give the topic a name, such as AdminAttn. The topic name could describe the resource the topic will be attached to (such as EC2), the application it's connected to, or the people or services that will be subscribed to it. Optionally, add a display name, such as your name, the name of your cloud network or resource, or the name of your organization. Click **Create topic**.

You now need to subscribe to this topic. Complete the following steps:

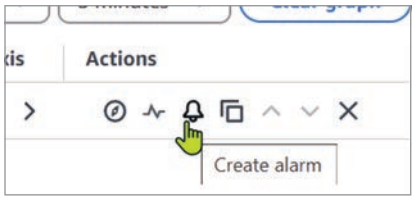
5. On the topic's page, scroll down and click **Create subscription**. For Protocol, choose **Email**. In the Endpoint field, enter your email address. Make sure to provide an email address you have access to so you can confirm the subscription. Click **Create subscription**.
6. When you receive the confirmation email, open it and click **Confirm subscription**.
7. To test your topic and subscription, return to the Topics page, click the topic to select it, and click **Publish message**. Give the message a subject line, such as Test message. Enter some text for the message body. Click **Publish message**. Check your email to confirm the message came through. Did it work? If not, troubleshoot the problem.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Now you're ready to create your CloudWatch alarm. Complete the following steps:

8. Go to the **CloudWatch** service, which is grouped in the Management & Governance category. In the navigation pane's Metrics group, click **All metrics**.
9. On the Browse tab, notice that several tiles are listed, such as Billing and SNS. Because you created a VM at the beginning of this project, the EC2 tile should also be listed. If it's not, check your EC2 service to confirm your instance is fully initialized.
10. In the Metrics pane, click the **EC2** tile. Click **Per-Instance Metrics**. Notice the Metric name column identifies several metrics, such as EBSIOBalance%, CPUUtilization, and NetworkPacketsOut. Click the check box to select the **NetworkIn** metric for the EC2 instance you created at the beginning of this project–confirm you're choosing the correct Instance ID that you saved from Step 2.
11. Click the **Graphed metrics** tab. In the Actions column, click the **Create alarm** icon (Figure 9-24).
12. Change the Period field to **1 minute**. In the Conditions section, select **Greater/Equal** and define the threshold at **15000**. Click **Additional configuration** and notice that the Datapoints to alarm option is set at 1 out of 1. Click **Next**.
13. Click in the "Send a notification to..." field and select the topic you configured in Step 4. In addition to sending a notification, what other actions are available for this alarm?
14. Click **Next**. Give the alarm a name, such as EC2NetworkIn, and click **Next**. Click **Create alarm**.
15. On the All alarms page, what state is the alarm in?
16. Every 30 seconds or so, refresh the list of alarms until the alarm state changes. What state does the alarm change to?

Figure 9-24 Create an alarm in CloudWatch

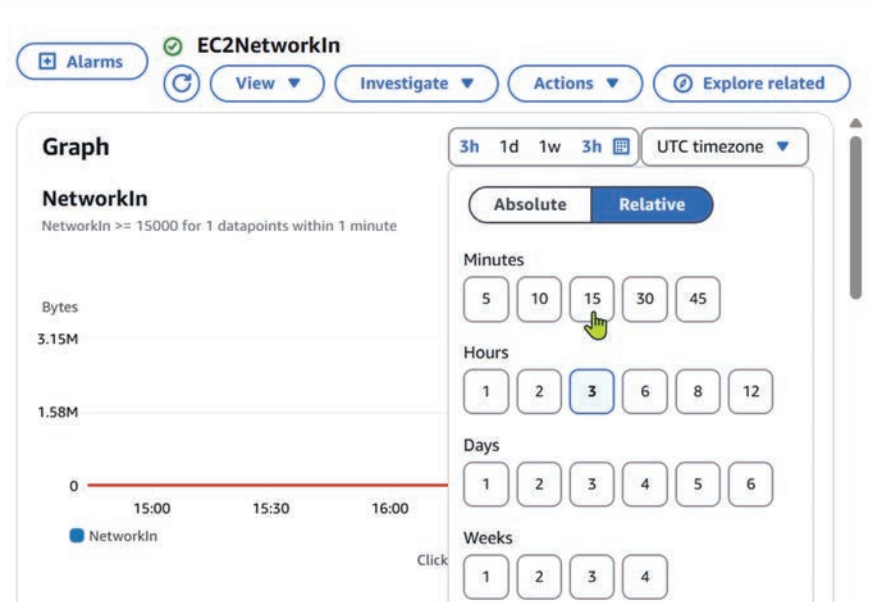


Source: Amazon Web Services, Inc.

Now that you have an alarm and you know your email subscription works, you're ready to trigger the alarm. Complete the following steps:

17. Click to select your alarm. Change the graph's timeframe to 15 minutes (Figure 9-25).

Figure 9-25 Set timeframe to 15 minutes



Source: Amazon Web Services, Inc.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

18. In your browser, open a second tab and go to the EC2 service. Copy the VM's public IP address.
19. On your local computer, open a Terminal window, such as PowerShell or Command Prompt. Enter the following command, using your VM's public IP address:

    ```
    ping <ip-address> -t -l 1500
    ```

    The `-t` switch runs the ping indefinitely until you tell it to stop. The `-l` switch increases the size of the ping packets. Note that the ping might not complete successfully; however, all you need to do is generate the incoming traffic so you can collect metrics on it.
20. While your ping is running in your Terminal window, repeatedly refresh the graphed metric in CloudWatch until the metric shows increased traffic and the alarm is triggered. Note that it might take a minute or two for the alarm's action to take place. What happens when the alarm triggers?
21. **Take a screenshot** showing the outcome of the alarm's action; submit this visual with your responses to this project's questions.
22. Return to the **EC2** service and check the **Monitoring** tab for your VM. If needed, click the **Refresh** icon. Change the timeframe to 15 minutes. What other metrics show recent increased utilization?
23. Stop the ping (press **Ctrl+C** in the PowerShell or Command Prompt window) and wait a few minutes until the traffic level returns to normal. What happens to the alarm?
24. Delete all the resources you created in this project, including the VM, topic, subscription, and alarm. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 9-9 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 9-2: Monitor a VM in Azure

Estimated time: 45 minutes
Objective 3.1: Given a scenario, configure appropriate resources to achieve observability.
Resources:

- Azure account and subscription
- Access to an active email account
- Internet access

Description: Generating notifications in response to an alert is a great way to familiarize yourself with some of the basic monitoring services and options in Azure. In this project, you'll configure an alert rule for a Windows VM and then trigger the alert.

Note 9-10 | At the time of this writing, all the services used in this project offered sufficient free limits to complete this project with no fees so long as you follow the instructions as written. To check a current list of free units included with Azure Monitor, search online for an Azure Monitor pricing list, and check the "Free Units Included" portion of the pricing schemes. At the time of this writing, this information could be found at azure.microsoft.com/en-us/pricing/details/monitor.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Complete the following steps:

1. In your Azure portal, create a new VM in a new resource group. Use the **Windows Server** image and the **B1s** VM size.
2. While the VM is being created, go to the **Monitor** service. In Monitor's navigation menu, click **Alerts**. Before you can configure an alert, you need an action group. Action groups are similar in concept to AWS's topics and determine where alert notifications are sent. At the top of the Alerts page, click **Create** and choose **Action group**.
3. Choose the resource group you created in Step 1. Give the action group a name and display name, such as AdminAttn, and click **Next: Notifications**.
4. Select the **Email/SMS message/Push/Voice** notification type. The Email/SMS message/Push/Voice blade opens. Select the **Email** check box and enter your email address in the field. Scroll down and click **OK**. Give the notification a name, such as MyEmail, and click **Next: Actions**.
5. While you will not configure actions for this action group, take a moment to explore your options. Which actions could this action group trigger?
6. Click **Review + create** and then click **Create**.
7. Click the **Action groups** tab, click your action group, and then click **Test**. Select the **Metric alert - Static threshold** sample type and click Test. Did it work? If not, troubleshoot the problem.
8. When the test is complete and successful, click Done to close the Test blade. By this time, your VM should be deployed. Return to the **Virtual machines** service and click your new VM. On the Overview page, click the **Monitoring** tab to display the various metrics running by default for your VM. What visualizations are listed by default?
9. In the VM's navigation menu, scroll down, click **Monitoring**, and then click **Alerts**. Click **Create** and click **Alert rule**. In the list of signals, what popular signal types are included?
10. Click **See all signals**, search for **deallocate**, and then click **Deallocate Virtual Machine (Virtual Machines)**. Click **Apply** and then click **Next: Actions**.
11. Select the action group you created earlier in this project, and then click **Select**. Click **Next: Details**.
12. Make sure your resource group that you created in Step 1 is selected, and give the alert rule a name, such as VMStop. Click **Review + create** and then click **Create**.
13. Click **Alert rules**. What severity level is listed for your alert rule?
14. Return to the VM's Overview page, click **Stop**, and then click **Yes**. Return to the **Monitor** service and examine the **Alerts** blade. Note that you might need to wait a few minutes for the alert's response to take place. What happens when the alert fires?
15. **Take a screenshot** showing the outcome of the alert's action; submit this visual with your responses to this project's questions.
16. Delete all the resources created in this project by deleting the resource group from Step 1. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 9-11 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Project 9-3: Monitor a VM in GCP

Estimated time: 45 minutes
Objective 3.1: Given a scenario, configure appropriate resources to achieve observability.
Resources:

- GCP account
- Access to an active email account
- Internet access

Description: Generating notifications in response to an alert is a great way to familiarize yourself with some of the basic monitoring services and options in GCP. In this project, you'll configure an alert policy for an Ubuntu VM and then trigger the alert.

Note 9-12 | At the time of this writing, all the services used in this project offered sufficient free limits to complete the project with no fees so long as you follow the instructions as written. To check a current list of free allotments included with GCP Monitoring service, search online for a GCP Monitor pricing list, and check the "Free allotment per month" portion of the pricing schemes. At the time of this writing, this information could be found at cloud.google.com/stackdriver/pricing.

Complete the following steps:

1. In your GCP console, create a new project. Switch to the new project and create a VM instance using the **e2-micro** machine type and an **Ubuntu** boot disk. If you need to avoid accruing charges, be sure to select one of the free regions. (Search online for a current list of regions that support free instances in GCP.)
2. After the VM is deployed, click it in the list, and then click the **Observability** tab. Refresh the metrics every few minutes until the initial data appears. Which metrics are monitored here?
3. In the navigation menu, scroll to the bottom and click **Monitor VM Instances**. (Note that you can hold down the Ctrl key on your keyboard while clicking the link to open the Observability Monitoring service in a different tab in your browser. You will need to view both Monitoring and Compute Engine services multiple times in this project.)
4. In the Monitoring navigation pane, click **Dashboards**. Notice you have several pre-existing dashboards to choose from and the option to create a dashboard, which would help you organize and monitor your alerts and other resources.
5. In the Monitoring navigation pane, click **Alerting**, and then click **Create policy**.
6. Click **Select a metric** and click the **VM Instance metrics** category. In the **Instance** subcategory, click **Received bytes**, and then click **Apply**.
7. Change the rolling window to **1 min** and click **Next**.
8. Configure the following settings:
   - Alert trigger: **Any time series violates**
   - Threshold position: **Above threshold**
   - Threshold: **800**
9. Click **Next**. Click the **Notification Channels** dropdown list, and then click **Manage Notification Channels**. On the new blade, scroll down and add your email address and a display name for the channel. Click **Save**. What are three other notification channels in the list that seem interesting to you?
10. Close the notification channels blade. Click the **Notification Channels** dropdown list again and select your email channel. Click **OK**.
11. Scroll to the bottom of the page and give the alert a name, such as VM-NetworkIn. Click **Next**. Review the policy and click **Create policy**.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

12. Return to Compute Engine, go to your VM's Details page, and copy its external IP address.
13. On your local computer, open a Terminal window, such as PowerShell or Command Prompt. Enter the following command, using the VM's public IP address:

    `ping <ip-address> -t -l 1500`

    The **`-t`** switch runs the ping indefinitely until you tell it to stop. The **`-l`** switch increases the size of the ping packets. Note that the ping might not complete successfully; however, all you need to do is generate the incoming traffic so you can collect metrics on it.
14. While your ping is running, return to the alerting policy's page in Monitoring, and observe it until the metric shows increased traffic and the alert is triggered. It could take several minutes for the alert to register. What happens when the alert triggers?
15. **Take a screenshot** showing the outcome of the alert's action; submit this visual with your responses to this project's questions.
16. Return to your instance's page, and then click the **Observability** tab. What metrics show increased utilization?
17. Stop the ping (press **Ctrl+C** in the PowerShell or Command Prompt window).
18. Delete the project you created in Step 1. Confirm the project is shut down and scheduled for deletion. Make sure to properly sign out of your account.

Note 9-13 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Challenge Projects

Note 9-14 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 9-1: Research Billing Alarms and Alerts

Estimated time: 30 minutes (+15 minutes for group work, if assigned)
Objective 3.1: Given a scenario, configure appropriate resources to achieve observability.
Group work: This project includes enhancements when assigned as a group project.
Resources:

- Public cloud account (such as AWS, Azure, or GCP)
- A GenAI tool, such as ChatGPT, Microsoft Copilot Chat, Google Gemini, or Perplexity
- Internet access

Description: In the Hands-On Projects for this chapter, you practiced creating an alarm or alert on a VM instance and then triggering that alarm or alert by pushing the VM instance beyond the defined threshold. In this project, you will figure out how to create an alarm or alert that will notify you if your cloud expenses exceed a comfortable threshold. Note that

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

students using educational access to their cloud platform (such as AWS Academy) might not be able to perform all the steps, since no charges are tracked to begin with. However, these students may still complete the research steps and click through some of the related services in their cloud platform to examine the configuration and feature options.

Complete the following steps:

1. Choose a public cloud platform, such as AWS, Azure, or GCP. Research online how to configure a budget alarm or alert that will notify you if your estimated or actual monthly expenses exceed a certain threshold, such as $10. You can also consult with a GenAI tool, such as ChatGPT or Perplexity, but make sure to validate any of the information it gives you. Which cloud platform did you choose? Which service do you need to use in that platform to configure a billing alarm or alert?
2. As you research, what special requirements must be taken into consideration? For example, do you need to use a specific region? Are there different types of billing alarms or alerts you can choose from?
3. How much will it cost to configure this alarm or alert? How much will it cost to continue using the alarm or alert over time? For example, will you be charged each time the alarm or alert sends a notification?
4. Using a GenAI tool, such as ChatGPT or Perplexity, ask about tips for determining the threshold for a billing alarm or alert. What are three tips you learned that you think would be most helpful in setting this threshold?

Note 9-15 | Sometimes AI tools don't at first give you the kind of output you had in mind. You might need to ask again for what you want using different terms, providing more information, or even giving examples. To get the best quality output from an AI chat tool, give specific and detailed instructions for what you want, provide context with relevant background information, instruct the AI tool to ask questions for any missing information, and phrase your request differently or add more information if you need the AI tool to try again.

5. **For group assignments:** Complete the following steps:
   a. Ask your GenAI tool for a practice scenario where you need to determine a billing alarm or alert's threshold. Make sure to instruct the AI tool not to give you the threshold amount–that you want to practice determining this amount yourself.
   b. As a group, collect the threshold tips you each came up with in Step 4 and work together to determine a threshold for the AI-generated scenario. Note that the provided scenario might not include all the information you need. Ask the AI tool for additional information if necessary, or determine what information is required but cannot be provided. For example, the scenario might state there is a strict budget for an upcoming sales event without telling you the budget. In this case, you would identify this piece of information as being needed before you can determine the threshold.
   c. If you're able to determine a threshold with the given information, check with the AI tool to determine the appropriateness of your proposed threshold and request feedback on your suggestion. What determination did you make, and what response did you get from the AI tool?
   d. If you're not able to determine a threshold with the given information, what missing information would you require in order to set a threshold? Within the context of the scenario, who would you need to ask in order to obtain that information?

If you have a standard cloud account with your selected platform, complete the following steps to set up your billing alarm or alert. If you're using an educational account (such as AWS Academy), experiment with how far you can progress toward setting up this alarm or alert before you run into account limitations.

6. Follow the steps you found in your research. If you're able to complete the configuration, **take a screenshot** of the configured alarm or alert; submit this visual with your responses to this project's questions. If you're not able to complete the configuration due to account limitations, **take a screenshot** of the error message you receive; submit this visual with your responses to this project's questions.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7. You might choose to leave your alert or alarm in place in your cloud account to provide ongoing monitoring. If so, confirm your earlier research regarding any costs that might be incurred over time and ensure you understand what the expected costs might be. If it's helpful, use the cloud platform's pricing calculator tool to help you determine this information.
8. If you prefer to delete the alert or alarm, you may do so now. Check through your account to confirm you permanently deleted or terminated any resources created in this project that you prefer not to keep. Make sure to properly sign out of your account.

Note 9-16 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Solutions to Self-Check Questions

Section 9-1: Monitoring Resources

1. A resource's ability to scale automatically is referred to as _____.

   Answer: d. elasticity

   Explanation: Elasticity refers to a cloud resource's ability to scale up or down automatically without requiring intervention from a cloud admin. Utilization refers to how much a resource's capacity is being used. Availability is measured by uptime versus downtime, connectivity to other services, and the health and performance of a system's various components. Performance is determined by latency measurements and relative health of cloud resources while minimizing disruptive incidents.

2. Which resource notifies admins of a breached threshold?

   Answer: a. Alarm

   Explanation: Alerts and **alarms** can inform staff when a threshold is breached or when a concerning event occurs. Tags are key-value pairs you create to track resources with whatever organization you need. A TAP (test access point) is a device installed on a network to capture traffic flowing between devices. Metrics help indicate a resource's or system's health, performance, and cost but they don't, on their own, notify admins when a threshold has been breached.

3. True or false: It's important to pack as much information as possible into a dashboard.

   Answer: b. False, because people can only absorb so much information at a time.

   Explanation: Human minds can consume and remember only about seven distinct pieces of information at one time. Similarly, **most people can only absorb about five to nine KPIs on a dashboard at a time**. While these KPIs can be designed in a way to present a lot of information efficiently, too much information is overwhelming and would be missed. While it is true that running extra dashboards costs more money, this does not change the fact that too much information in one dashboard is overwhelming. Extra space in a dashboard can help mitigate this issue by assisting the mind in grouping the information according to relevance. The most detailed information should be positioned at the bottom of the dashboard after the big-picture perspective has drawn attention to the most important information first.

Continues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 9-2: Events, Logs, and Traces

4. Which of the following is an example of an event?

Answer: b. A storage bucket is deleted.

Explanation: An event is a detectable change in a cloud service or resource, such as **deleting a storage bucket**. A VM's CPU utilization or read rates from a storage volume change constantly and can be measured using metrics. Unless it has been specifically enabled in a specialized monitoring tool or script, failed pings are not recorded as events.

5. Which process protects private information?

Answer: d. Log scrubbing

Explanation: Log scrubbing removes sensitive data from logs before storing captured traffic. Event correlation evaluates many events across systems to detect patterns and anomalies. Triage means to prioritize and organize alerts so the most critical issues are addressed first. Root cause analysis is used to identify the underlying source of problems and symptoms.

6. Which record would best reveal the location of a bottleneck across services?

Answer: a. Trace

Explanation: A **trace** is a record of a request's progress across a distributed system involving many services, which can quickly identify the location of a bottleneck in service communications. An audit trail is the collection of data that thoroughly documents activities in a system to prove compliance, not performance. An event records information about a single activity in a single service. Events can be collected into logs, but analysis of many logs across services might still not detect the location of a communications bottleneck.

Section 9-3: Cloud Monitoring Services

7. An event producer _____ messages.

Answer: d. publishes

Explanation: An event producer **publishes** messages when events occur. Event consumers subscribe to messaging channels. After analyzing the data from these messages, the consumer might also respond to them with automated actions.

8. What kind of application relies on event-driven architecture?

Answer: b. Decoupled

Explanation: Each event in an event-driven architecture serves as part of a larger process that emanates throughout a **decoupled** application's microservices. A monolithic application is a large and complex, self-contained system with deeply integrated components. An authentication service might be one of the microservices within an event-driven architecture, but it does not necessarily depend on events. Each of the microservices within an event-driven architecture can run independently.

9. What does the symbol > = stand for?

Answer: d. Greater than or equal to

Explanation: The symbol > = combines the logical argument > ("greater than") with the logical argument = ("equal to"), and it collectively means "**greater than or equal to**." The symbol < means "less than" and the combination of < = means "less than or equal to."

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 9-4: Troubleshooting Problems and Performance

10. Which technique adjusts resources to match demand?

Answer: b. Rightsizing

Explanation: Rightsizing means resources are increased or decreased according to need to balance performance and cost. Throttling restricts the number of requests allowed within a given period. Oversubscription is a strategy to offer more capacity than what underlying infrastructure can actually handle at any given time. Replication is one way of mitigating performance issues caused by oversubscription.

11. Which of the following problems could result in an orphaned resource whose expenses are not properly tracked?

Answer: c. Missing or incorrect tags

Explanation: Resources must be appropriately tagged to allow for proper tracking of expenses. An **untagged or improperly tagged resource** could be left running without being properly identified, tracked, secured, or updated. An incorrect role assignment can prevent a service from replicating resources for auto-scaling. Conflicting policies might prevent appropriate auto-scaling if these policies are triggered by the same circumstances. A stuck instance can distort auto-scaling metrics.

12. Which of the following statements applies to an in-place recovery?

Answer: a. Replaces damaged data with backup data to the same host

Explanation: An in-place recovery **restores backup data to the same resources that hosted the original data**. A parallel recovery, also called a side-by-side recovery, restores a system using backup data that is hosted in a different location. This recovery process can be completed more quickly because the backup copy can be accessed with a simple configuration change that redirects traffic to the backup location. This option is well suited to disaster recovery strategies where downtime is undesirable.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Chapter 10

Cloud Automation

Chapter 10 Objectives

After reading this chapter and completing the exercises, you should be able to:

1. Explain how code supports automation in the cloud.
2. Describe how pipelines automate code integration and testing.
3. Identify cloud maintenance tasks that are commonly automated.
4. Troubleshoot automation issues.

Chapter 10 Outline

Study the chapter's outline to help you organize the content in your mind as you read.

Section 10-1: Automation
- Why Automate?
- IaC (Infrastructure as Code)
- CaC (Configuration as Code)
- Comparing Tools
- Cloud CLIs

Section 10-2: Automating Code Pipelines
- CI/CD Pipelines
- Scripting Logic

Section 10-3: Cloud Maintenance
- Types of Updates
- Patch and Update Methodologies
- Decommissioning

Section 10-4: Troubleshooting Automation Issues
- Common Automation Issues
- Breakdowns in Workflow

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 10 Scenario "I passed!" you announce as you walk in the door and throw a thumbs-up and a cheesy grin at the security camera.

"What's going on?" Kendra asks as she peeks cautiously out of her office door.

"All that research we've been doing for the cloud . . .," you begin. "I thought I might put my knowledge to the test, literally, and take the CompTIA Cloud+ exam. So I did. And I passed!"

"Wow, that's awesome!" Kendra exclaims.

"Way to go!" Malik agrees as he offers a double high-five.

"Seriously, that shows a lot of initiative," Kendra adds. "Goes to show how much we've all learned over these past few months."

"Now if only our cloud would show that much initiative," you jokingly suggest, "and build itself."

"Ha! That would be a smart cloud," Malik laughs.

"Actually, it might not be so far-fetched," suggests Kendra. "I've been reading up on some pretty amazing tools for automating cloud deployment."

"That still takes a lot of work up front," you say as the wheels in your head start turning. "But if we design our cloud with the intention of building in a lot of automation, it wouldn't take nearly so much work to keep it running smoothly later."

"All hail our resident cloud expert," Malik jokes. "You actually sounded like you knew what you were talking about just then."

"Well, I had to study up on automation for the exam. I've only touched the surface, though," you add. "Automation is a whole field of expertise all on its own."

"This is true," Kendra agrees.

"Which part?" asks Malik as he realizes she's been staring into space and not really listening.

"Oh! Building in automation, I mean," Kendra says as she starts thinking aloud. "You know, that's really the piece I've been missing with our cloud migration plans, that part about how to keep the cloud running smoothly without needing so much micromanaging from us."

"Right—we can set it up to almost run on its own," you add. "You can't totally ignore it, of course, but you don't have to get bogged down in so much of the tedious, nitpicky stuff . . . if you set it up right."

"You know," Malik muses, "I think this is kind of like your management style, Kendra."

"How's that?" Kendra asks with a quizzical expression.

"The way you set us up to do our own research, to come up with our own ideas, to keep learning," explains Malik. "You don't tell us what we have to know. You ask questions, we go find our own answers, and then come back and share what we learned."

"This is true," you add. "I really appreciate how you share your curiosity instead of dictating what we must do and how. It makes me want to know even more and gives me confidence to find my knowledge gaps so I can keep learning."

"Aw, now I'm blushing," grins Kendra. "Okay, well, enough of that," she says as she recovers her composure. "How can we incorporate some of these automation tools into our cloud?"

"Can I ask the questions this time?" Malik pleads excitedly.

"Go for it," Kendra says.

Malik pauses for a moment to think and then starts scribbling on the whiteboard in Kendra's office:

- How does automation work in the cloud?
- How does automated coding work in the cloud?
- What tasks can automation tools perform?
- What can go wrong with automation?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 10-1: Automation

Certification

1.6 Compare and contrast containerization concepts.

1.10 Compare and contrast methods for optimizing workloads using cloud resources.

2.4 Given a scenario, use code to deploy and configure cloud resources.

5.1 Explain source control concepts.

5.2 Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.

5.4 Explain the importance of tools used in DevOps environments.

Have you ever watched a movie or cartoon where a genius inventor is woken up in the morning by an alarm clock, a robot lifts him out of bed, and robot arms reach in and start combing his hair and brushing his teeth; meanwhile, the coffee pot starts dripping automatically, and a can of dog food is cut open and dumped in the dog's bowl? Audiences have long been fascinated with the idea of automation, and moviemakers play on this interest with often humorous results. The opening scene to the classic movie *Back to the Future* illustrates a perfect example of one inventor's shenanigans gone hilariously wrong. Also, search online for images of Rube Goldberg machines (rubegoldberg.org), which make simple tasks complicated in the interest of automation.

Thankfully, cloud automation is more productive, more streamlined, and generally less chaotic. Automation refers to the use of software-based solutions to perform a process or procedure with minimal human intervention. In the physical world, examples include cars that can park themselves, digital assistants like Siri or Alexa that can start your coffee pot in the morning, drones that can deliver packages, or robots that deliver meals on a college campus (Figure 10-1). In the cloud, examples include an automated process for creating, cloning, or shutting down a VM instance or server cluster, configuring and deploying a virtual network, or even deploying an extensive and complex cloud environment. The idea with cloud automation is to reduce the demand on human staff's time while increasing accuracy and repeatability of common tasks.

Figure 10-1 Automated drone delivery of meals and snacks



iStockphoto.com/MattGush

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Think about the steps required to create a new VM. You choose the boot source, size (sometimes referred to as shape), network connections, security configurations, and storage allocations. The first time you do this, you must think through the options and implications of your decisions. However, the next time you need a VM with the same configuration, you don't have to invest so much thought. If this is a configuration you use often, you eventually reach a point where creating the new VM consists of a series of mind-numbing clicks where you implement the same decisions you made the first time, over and over again. The more boring or repetitive the task, the more likely you'll make a mistake. The thing is, humans tend to be particularly bad at doing the same thing exactly the same way over and over. Computers, however, are very good at doing exactly what you tell them to do and doing it exactly the same way every time. When you find yourself doing the same thing repeatedly, this is often an ideal candidate for automation.

Why Automate?

Automation in the cloud accomplishes several important goals, such as the following:

- **Faster deployment, adjustments, and corrections**—As you've already learned, monitoring tools can be programmed to respond automatically to problems. Automated tools can also deploy or resize resources quickly and exactly when needed.
- **Better control**—Standardized processes are easier to define, monitor, and replicate. When you know a task is performed the same way every single time, you can better predict the effects of those changes and fine-tune the process as needed.
- **Lower costs**—With fewer errors and less demand on staff's time, your team can focus on higher-level, strategic activities that add value to your organization while paying less for cloud management and resources.
- **More secure**—It might seem counterintuitive to think of security increasing when admins hand off some control to the computer. However, a well-orchestrated system is much less likely to suffer from vulnerabilities due to misconfigurations and other human errors.

Modern automation tools designed for the cloud offer flexible, robust features that can help your organization take full advantage of the cost-saving and performance benefits the cloud has to offer. Some of the maintenance and security tasks that these tools can perform include the following:

- Create snapshots and backups.
- Clone, resize, or remove resources.
- Apply patches and updates.
- Generate internal API communications.
- Restart or shut down VM instances.
- Enable or disable alerts.
- Detect signs of an attack.
- Quarantine compromised resources.
- Manage security rules.
- Migrate traffic to redundant resources.

IaC (Infrastructure as Code)

How does all this work? How can you set up a cloud system to make good decisions and perform complex maintenance, provisioning, or configuration tasks reliably without human intervention?

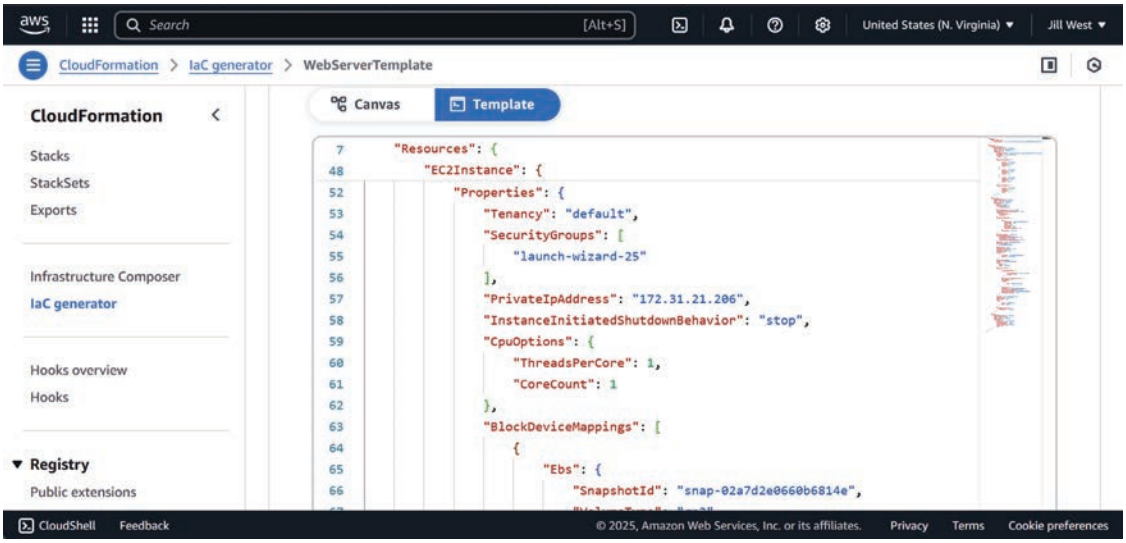
To understand the basic principles of automation, it helps to first understand the concept of a runbook. A **runbook** is a collection of physical or digital documents that outline the precise steps to complete

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

procedures or operations in a specific IT environment. For example, an admin might document the steps to provision a new server, update a running VM instance, or recover from a database backup. Considering what you've learned in this course, you could write your own runbook for how to create a virtual network with an Internet-accessible subnet in AWS, Azure, or GCP. A good runbook provides sufficient detail so other people could follow the steps and obtain the exact same results each time. Good runbook management keeps the documentation current and well organized, as well as accessible by the right people. This documentation might be stored in some kind of CMS (content management system) or knowledge base.

A runbook is typically designed to be read and followed by humans. However, consider what you could do with a runbook that is written in a computer-readable language, such as a programming language or scripting language. With sufficient detail, you could create runbooks that could spin up a VM, create or replicate a database, or configure a virtual network with multiple servers. This application of the runbook concept in an automated environment is called **IaC (infrastructure as code)**, where the steps needed to configure or change a cloud environment are written in code and are run collectively at one time rather than requiring the human admin to complete one step of the deployment at a time in a GUI (graphical user interface). For example, Figure 10-2 shows the code needed to deploy an EC2 instance in AWS using CloudFormation.

Figure 10-2 Cloud resources represented in code



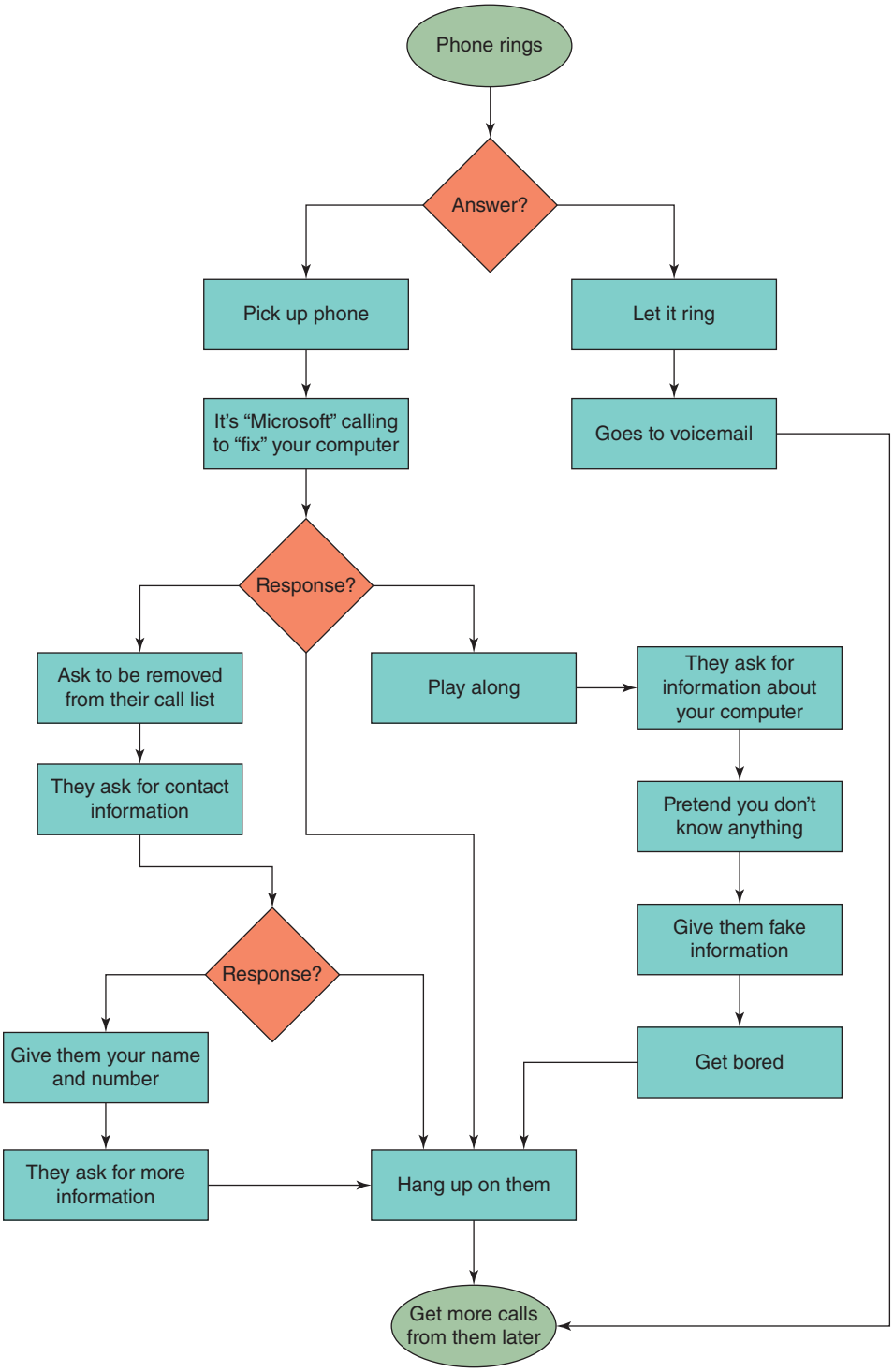
Source: Amazon Web Services, Inc.

Now imagine that you've created several runbooks, all designed to work together to create multiple nodes and all the related security, networking, and OS configurations. In cooperation, these runbooks could be used to deploy a large and complex testing environment complete with network services, security groups, VMs running various OSs, and predictable access keys and passwords as desired. Collections of automated processes like this in a carefully choreographed manner is the essence of orchestration. Where automation performs a single task (such as create a VM), orchestration designs a series of tasks in a workflow (such as creating an entire testing environment). A **workflow** defines a series of steps, tasks, or operations that can be automated, scheduled, and monitored.

Figure 10-3 illustrates an amusing example of a workflow for the simple process of how to handle a spam caller. Notice the various symbols used. The diagram begins with some kind of input presented in an oval, change occurs through decisions (diamonds) and processes (rectangles), and then the result of the workflow is the output.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 10-3 A simple workflow for handling spam calls



The sequencing of multiple tasks, phases, decision points, or deployments in orchestration is a key part of designing effective IaC scripts or templates. A **script** is a document defining a simple, static set of instructions for a series of tasks. A **template** is a document or collection of documents defining a specific outcome of automated actions. For example, you might have a script that deploys a VM or updates a server. A template

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

might define configurations for all the resources needed for a specific solution, including multiple cloud services, such as databases, containers, storage buckets, or VM instances. Some automation tools require scripts while others work with templates.

IaC Benefits

Documenting and designing infrastructure in a code format presents many benefits. For example, IaC and related automation and orchestration tools enable a more consistent environment throughout the process of developing, testing, deploying, and patching an application or service. During an app's development phase—or later, as changes are made—developers need testing environments that are as nearly identical to the app's final production environment as possible. IaC can ensure that all these environments are properly synced and updated.

Think about how quickly applications are changed and updated in a DevOps-based organization. New releases might occur every day, even many times a day. By maintaining consistency between the various environments representing stages of an app's development, IaC helps minimize problems as the app is transitioned from one environment to the next.

IaC also assists in troubleshooting efforts or security investigations. In a large, active organization—or even in a small company—it could be nearly impossible to remember who made what changes if everything is done manually, where a staff person gives direct reconfiguration commands to the cloud deployment. What was the state of a cloud environment when a particular network connection went down or when a problematic intrusion occurred? Making manual configuration changes typically provides no records of what changes were made when and by whom.

In contrast, by first defining a cloud environment in a text file, changes to that environment can be tracked by examining historical versions of the file. Each time a change is executed in an environment, the previous state of the environment is still documented in the earlier version of the file. This is especially useful when you experience a problem with the new configuration and must roll back to a previous configuration. This reversion can be accomplished by deploying the previous environment's configuration file. Similarly, if a system experiences an outage, configurations can be recovered and the environment quickly redeployed, a strategy known as DRaC (disaster recovery as code).

Because IaC provides documentation of each automated environmental configuration along the way, troubleshooting or security teams can investigate historical changes made in the cloud environment to determine when a problem occurred and why. In essence, IaC provides a "paper trail" (digital, of course). This same IaC documentation also helps in meeting regulatory standards to prove historical compliance, which is essential during audits or court-ordered investigations. And IaC processes increase consistency across an organization's staff members, especially when the organization experiences staff turnover.

Versions of the code in a script or template used to deploy and maintain resources can be kept for historical analysis as changes are made to the deployment and for post-deployment validation. Older versions of the deployment's configurations can be reviewed as needed, changes to current configurations are thoroughly documented, and future changes can be evaluated and tested before deployment. Versioning is a foundational concept in managing, monitoring, and auditing an automated environment.

IaC Tools

Automation tools might be built into the cloud platform or provided by third-party vendor applications that interact with resources through APIs. Many of these tools are primarily managed through a CLI (command line interface) and rely on a variety of programming languages, such as Python, JSON (JavaScript Object Notation), and YAML (Yet Another Markup Language or YAML Ain't Markup Language, rhymes with "camel"). Some of these tools offer a web GUI or cloud portal, especially if you're using the paid version of the application. However, by understanding the underlying file format and programming languages, you can also write scripts or custom programming to design automated processes unique to your environment.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exam Tip | Notice there are two common spellings for the YAML acronym: Yet Another Markup Language, which was the original spelling, and YAML Ain't Markup Language, which is the current spelling.

YAML is not considered a markup language. YAML is used to represent data structures, which makes it a data serialization language. For example, YAML can be used to define configurations for a stack of resources or steps in an automated pipeline. It's a way to structure data for efficient processing, such as when deploying cloud resources with an automated tool.

In contrast, a markup language is used to enhance a document with information that a computer can use to better interpret and display the included content. For example, HTML (Hypertext Markup Language) is a markup language that can identify what text to bold or to change font colors on a webpage.

Note that, although the CompTIA Cloud+ objectives list YAML using its original spelling, YAML is not a markup language. However, you will likely encounter many instances where YAML is discussed as a markup language, especially because some other markup languages, such as XML (Extensible Markup Language), can be—and often are—used as a data serialization language. With this much crossover between how languages are used compared to how they were designed, the terminology for describing these languages can be murky, resulting in people using these terms in ways that are not always accurate.

No single tool can meet all your automation needs. Some automation tasks address infrastructure provisioning or server configurations while other automation workflows target management of applications or IAM accounts. A tool that attempts to address all these areas typically doesn't handle any of them well. As a result, you need to know about the wide range of tools available to wisely choose which combination of tools will best serve your situation.

To begin, it's important to distinguish between two major categories of automation tools: provisioning and configuration management. You learned about some common automated tasks in the previous chapter where you read about auto-scaling features. Auto-scaling is a type of provisioning task. **Provisioning tools** are used to deploy new resources. However, creating the resource is only the beginning. Automated processes might also need to configure, update, and reconfigure these resources over time to ensure they meet a certain standard of settings, software installations, and security configurations or to adapt to changing factors. This latter process is performed by **CM (configuration management) tools**. While many automation tools can do some or both these processes (provisioning and CM), most tools are specialized for one or the other. Whereas provisioning tools deploy new cloud resources, CM tools reconfigure existing resources to match a particular configuration template.

One of the most popular provisioning tools is Terraform. **Terraform** was developed by HashiCorp and provides cloud-agnostic infrastructure provisioning. The open-source version is free to the public while the enterprise version offers additional features, such as GUI, collaboration, and governance capabilities that work well in an enterprise environment where many teams rely on the same tools. To use Terraform, you first create a plan using the Terraform language, which is declarative in nature. This means you describe the outcome rather than the steps to get there. Next, you apply the plan to deploy the resources to your cloud or multi-cloud. Terraform works with AWS, Azure, GCP, and more than 150 other service providers.

Note 10-1 | Platform-specific provisioning tools, which could be used instead of Terraform, include AWS's CloudFormation, Azure's Resource Manager, and GCP's Deployment Manager.

CaC (Configuration as Code)

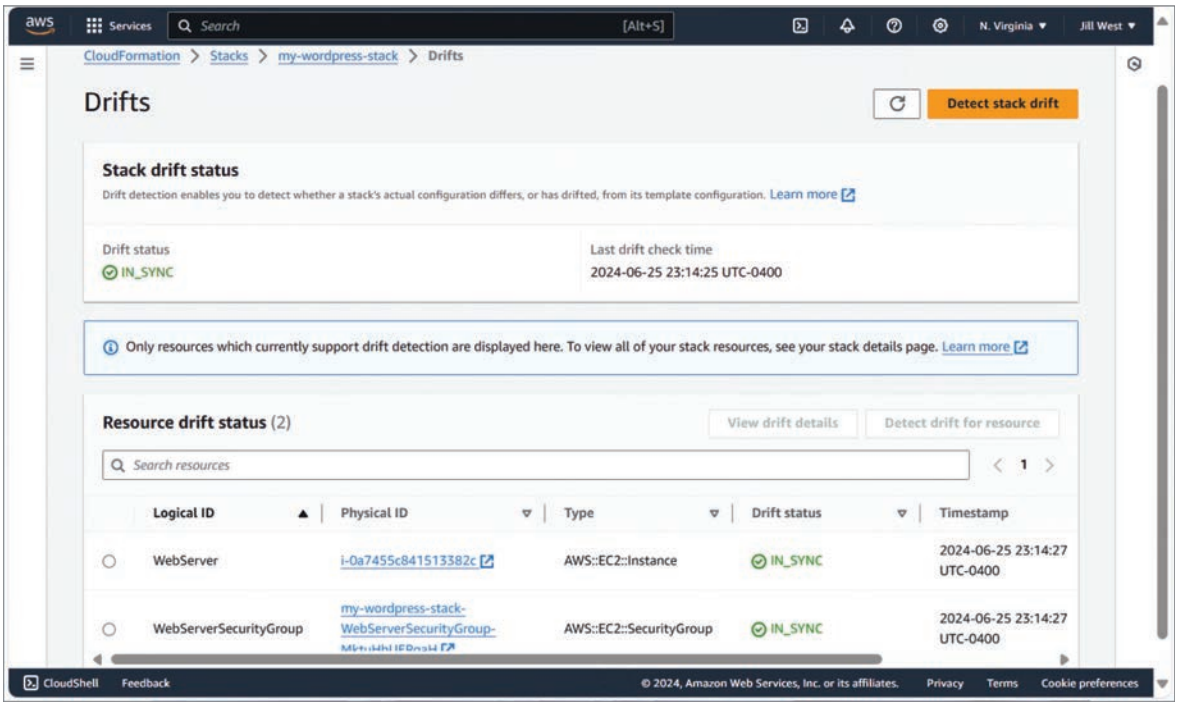
Once an environment is deployed, you are faced with the challenge of maintaining the configuration of all the components in that environment. Suppose you deploy a production environment for a business-critical application. One of your admins is trying to troubleshoot a problem with a performance bottleneck

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and manually changes the configuration of a network interface. The configuration change causes a critical server to lose access to its database, resulting in the entire application failing. This kind of problem falls into the realm of configuration management. CM tools work to monitor and enforce the configuration of each managed resource. When this process is performed using code to monitor configurations, it's called **CaC (configuration as code)**. CaC can help prevent configuration drift, which is the changes to configurations that cause them to exceed the limitations of compliance requirements.

Resource configurations are often defined in configuration files, which can be stored in a repository for access and monitoring. For other resources, their configuration information can be gathered by scanning the item or interacting with it. A CaC tool, such as AWS Config, can collect and monitor this information for unauthorized changes, which is a process called drift detection. Figure 10-4 shows the drift status of a CloudFormation stack in AWS. When attempting to change the security group for the Web Server EC2 instance, the change was automatically reverted to the security group defined in the CloudFormation stack. Some tools, such as in this example, can prevent any configuration changes from being made to monitored resources in the first place. The CaC tool can also automate the deployment of configuration changes and track changes over time, which makes it easier to revert to an older configuration if something goes wrong with a change.

Figure 10-4 Web server forced back into conformity with desired configuration



Source: Amazon Web Services, Inc.

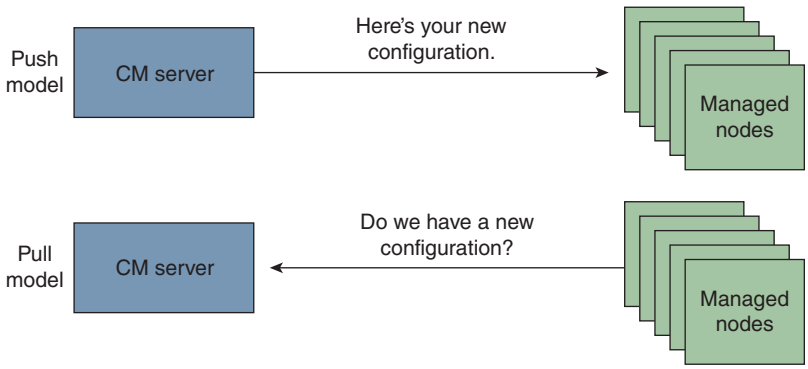
Ansible is an open-source (but not free) automation tool and supports the automation of managing infrastructure configuration. It's sponsored by Red Hat and relies on YAML. Ansible requires no agent installation on managed nodes and, instead, connects to these nodes via SSH (Secure Shell) by default. For this reason, Ansible is generally considered easier to set up than other third-party options. To automate changes, you create a text inventory file that contains definitions of variables for managed resources.

Note 10-2 | Other cloud-agnostic CM tools include Chef and Puppet. Cloud platforms often offer their own configuration management services, such as AWS Config, Azure Automation State Configuration, and GCP's Infra Manager.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Comparing Tools

Figure 10-5 Push model versus pull model



Provisioning and CM automation tools use a variety of methods to manage automated resources. Consider the following variations:

- **Push model versus pull model**—In a push model, the server pushes out configuration changes to managed resources (Figure 10-5). In a pull model, managed resources check in with the server to retrieve configuration details. The push model works well when managed resources are easily identified and located, and they're consistently reachable through established network connections. For more transient resources, such as auto-scaled VMs, the pull model is typically more effective. Puppet and Chef use the pull model, whereas Ansible uses more of a push configuration.
- **Immutable versus idempotent**—Immutable environments or resources aren't changed after they're deployed. They're deployed in the state they're intended to be used. If you need to make changes, you destroy the old resources and replace them with new resources that meet your desired configurations. Idempotent, however, means that you can run the configuration and it will reconfigure resources without destroying them. Ansible and Chef rely on idempotent techniques whereas Terraform, CloudFormation, and AWS's auto-scaling groups take an immutable approach to resources.
- **Imperative versus declarative**—Imperative tools require that you give them the exact steps you want taken to complete a task. Declarative tools, however, require less hand-holding. You tell the tool the outcome you want, and the tool figures out how to accomplish that goal. Chef and Ansible take an imperative approach whereas Terraform, Puppet, and CloudFormation offer a declarative style.

Cloud CLIs

One way to start getting more comfortable with using code to deploy resources is to practice deployments using a CLI. When you're first learning how to use and troubleshoot resources in the cloud, you typically start out using GUI consoles and dashboards. As you become more comfortable with cloud features and options, you might find yourself drawn to the ease and convenience of working from the CLI where you can also use code and scripts to control your cloud resources. AWS, Azure, and GCP, as well as many other CSPs, offer one or more CLI tools. Using a CLI is one way you can create scripts for deploying cloud resources, which gives you the option to automate this and many other tasks. Let's take a quick tour of some of these CLI tools.

AWS CLI

The AWS CLI is a tool you install on your computer so you can use your local computer's CLI to interact with AWS services. From your computer's terminal, such as Windows PowerShell, you can perform similar tasks to those you would do in the AWS Management Console through your browser. The commands interact with AWS's public APIs using account credentials stored on your computer. The AWS CLI can be installed on Windows, Linux, macOS, and Unix computers.

To install the AWS CLI, you download an installation package, which includes all the dependencies needed to install and run the AWS CLI on your computer. Make sure to source the package directly from AWS using their instructions for your OS. And ensure that you choose the most recent, stable AWS CLI version. If you have formerly installed an older version of AWS CLI, you will likely need to uninstall that version first.

After installation, you configure the AWS CLI with security credentials, default output format, and default AWS region. This can be accomplished easily using the `configure` command.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The AWS CLI's command syntax consists of a base call to the AWS program (written as `aws`), a command that identifies the AWS service, a subcommand that identifies the operation, and any options or parameters, as shown here:

```
aws <command> <subcommand> [options and parameters]
```

For example, a commonly used command is `ec2`, which interacts with the EC2 service in AWS. If you want to get help on this command, enter `aws ec2 help`. If you want to list your existing instances, enter `aws ec2 describe-instances`. You could also add parameters to this command in the form of filters. The following example includes a filter that limits output to instances with the instance type t3.micro:

```
aws ec2 describe-instances --filters "Name=instance-type, Values=t3.micro"
```

Table 10-1 lists some common commands used in the AWS CLI.

Table 10-1 Common AWS CLI commands

Command	Description
`aws --version`	Gives the installed version of AWS CLI, which is a good way to confirm the installation was successful
`aws help`	Gives information about AWS CLI options and top-level commands
`aws ec2 help`	Gives information about EC2 commands
`aws ec2 describe-instances help`	Gives verbose information about the describe-instances operation
`aws ec2 describe-instances`	Lists all EC2 instances

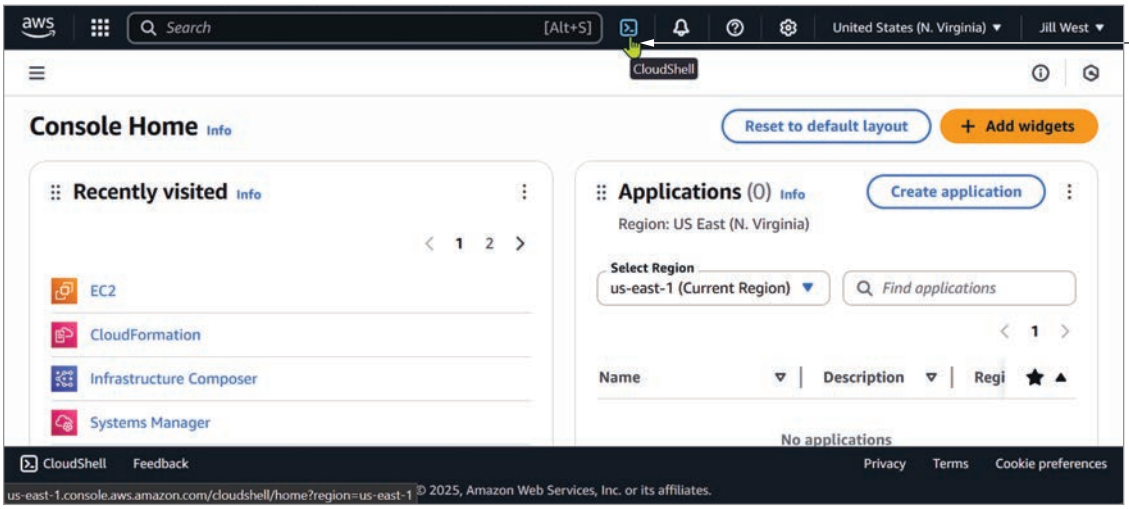
`aws ec2 run-instances`
`--image-id ami-xxxxxxxx`
`--count 1`
`--instance-type t3.micro`
`--key-name MyKeyPair`
`--security-group-ids sg-xxxxxxxx`
`--subnet-id subnet-xxxxxxxx`
	Creates a t3.micro instance using a specific AMI, key pair, security group, and subnet

`aws ec2 terminate-instances` `--instance-ids i-xxxxxxxx`	Terminates an instance identified by its Instance ID

AWS also offers a similar tool within the console called CloudShell (Figure 10-6). The CloudShell service opens a CLI pane in your browser where you can use the same AWS CLI commands without having to install any software on your local computer. In a project at the end of this chapter, you practice using AWS CloudShell to create and terminate an EC2 instance.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

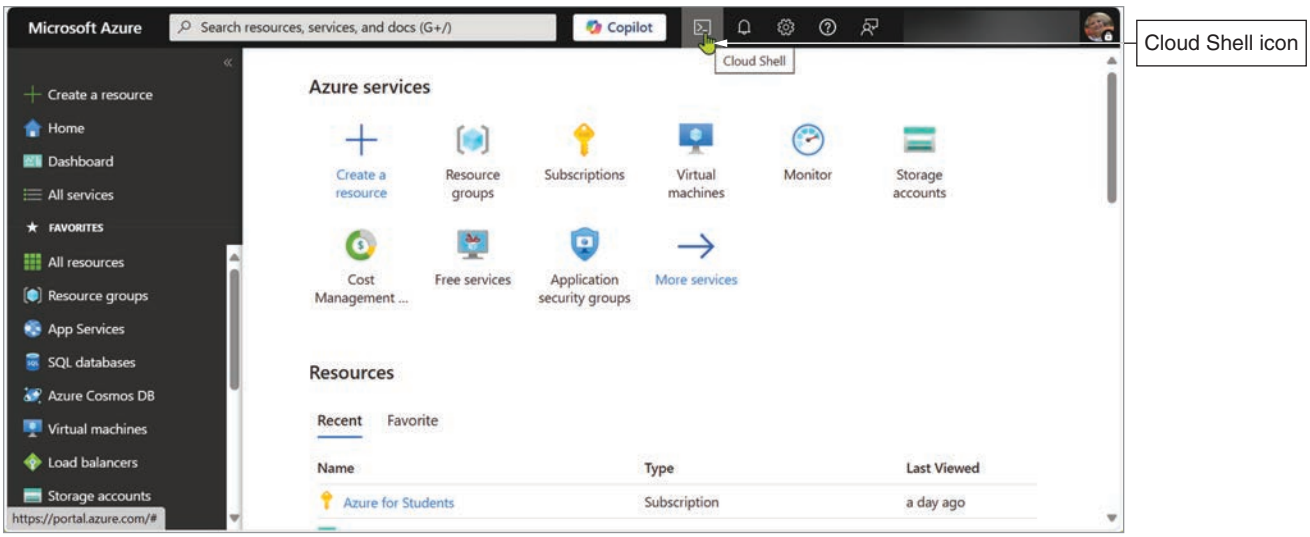
Figure 10-6 Access AWS CloudShell in a browser



Azure CLI

The Azure CLI can run in the Azure Cloud Shell in your browser, or you can download and install the Azure CLI for use through your computer's terminal, such as PowerShell in Windows. Launch Cloud Shell from the portal (Figure 10-7) and then choose a subscription to associate the session with. To install Azure CLI, download the installation package for your OS and complete the installation. This is a much more straightforward process in Windows for Command Prompt or PowerShell than in other OSs. If you're using a Linux system, macOS system, or the WSL (Windows Subsystem for Linux), you'll need to check the Azure site for specific instructions. After installing Azure CLI, you open your computer's terminal, open an Azure session with the `az login` command, and sign in using your Azure credentials.

Figure 10-7 Access Azure Cloud Shell in a browser



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Similar to the AWS CLI, you start all Azure CLI commands with the base call `az`. Check the installed version with the `az version` command. Table 10-2 lists a few additional, common Azure commands.

Table 10-2 Common Azure CLI commands

Command	Description
`az version`	Gives the installed version of Azure CLI, which is a good way to confirm the installation was successful
`az --help`	Gives information about Azure CLI options and top-level commands
`az interactive`	Enables interactive mode, which displays help information as you work, provides auto-completion, and gives examples
`az find vm`	Uses AI to find popular commands containing the search term "vm"
`az vm --help`	Gives information about the `vm` command
`az vm list`	Lists details of all VMs associated with the current subscription
`az vm create -n NewVM -g MyRG --image UbuntuLTS`	Creates a VM using a specific name, resource group, and image
`az vm delete -g MyRG -n NewVM --yes`	Deletes a specific VM without prompting for confirmation

Note 10-3 | An older version of the Azure CLI, now called the classic CLI, used the base call `azure`.

GCP CLI

The Google Cloud CLI includes several command-line tools, depending on the tasks you're performing. The following list describes some of the most important tools:

- **gcloud**—Among many other capabilities, gcloud is used to create, start, and manage VM instances, networks, firewall rules, disk storage, and more.
- **gsutil**—This tool is designed specifically to manage Google Cloud Storage.
- **bq**—This CLI tool is used for managing Google BigQuery, which is a data analytics service that operates in conjunction with Google Storage.

To use any of these CLI tools, you install the free Google Cloud CLI installer from cloud.google.com/sdk/. Alternatively, you can access Cloud Shell in the GCP console (Figure 10-8). Cloud Shell is a Debian-based VM with preinstalled utilities such as gcloud and is accessed directly through your browser.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 10-8 Activate Google Cloud Shell

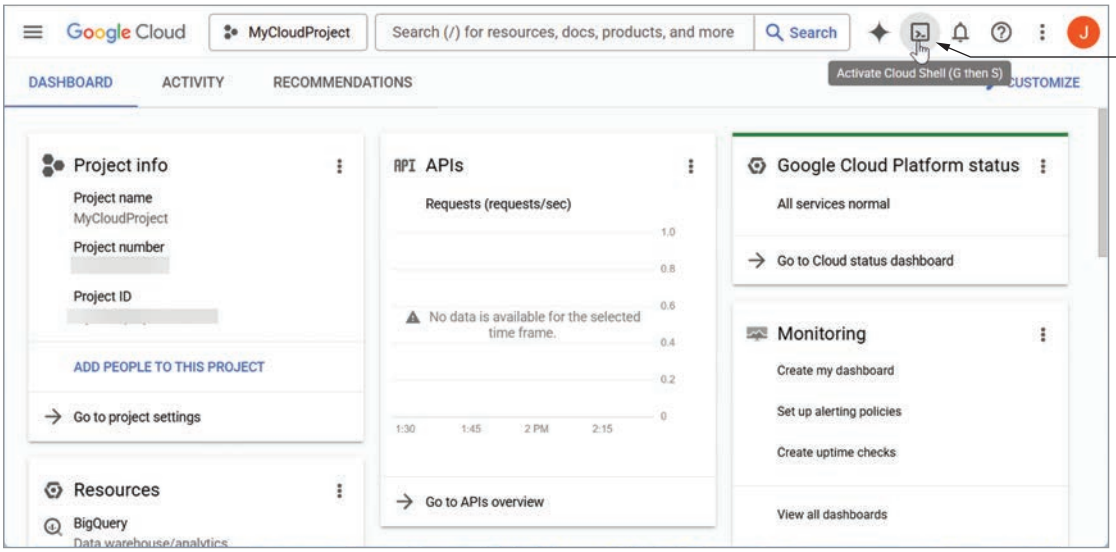


Table 10-3 lists a few simple commands you can use in gcloud.

Table 10-3 Common gcloud commands

Command	Description
`gcloud version`	Displays the current gcloud version and any installed components
`gcloud help`	Gives information about gcloud commands and shows the major groupings of commands and their syntax
`gcloud help compute`	Gives information about the compute group of gcloud commands and shows the major subgroupings of these commands and their syntax
`gcloud compute instances list`	Lists all compute instances
`gcloud compute instances create my-instance --zone=us-east4-b`	Creates the VM instance named "my-instance" in one of the N. Virginia zones
`gcloud compute instances start my-instance`	Starts the VM instance named "my-instance"
`gcloud compute instances delete my-instance`	Deletes the VM instance named "my-instance"

Self-Check

1. Automation increases _____.
   a. costs
   b. response time
   c. security
   d. errors

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. What IaC feature would enable you to deploy an earlier state of a web server's configuration?
   a. Immutable
   b. Versioning
   c. Provisioning
   d. Runbook

3. What text initiates communication with an installed CLI application?
   a. Parameter
   b. Base call
   c. Subcommand
   d. Filter

☐ Check your answers at the end of this chapter.

Section 10-2: Automating Code Pipelines

Certification

2.4 Given a scenario, use code to deploy and configure cloud resources.

5.2 Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.

Often, a primary responsibility of cloud technicians is to support software developers with the infrastructure they need to develop, test, and deploy code. This is typically managed through a code pipeline that consists of stages of development, such as building, testing, and deploying. But keep in mind that cloud infrastructure is usually managed through code as well. IaC code can also be automated through a code pipeline using most of the same tools that software developers use. In this section, you'll learn more about how these pipelines can be automated for faster and more efficient deployment.

CI/CD Pipelines

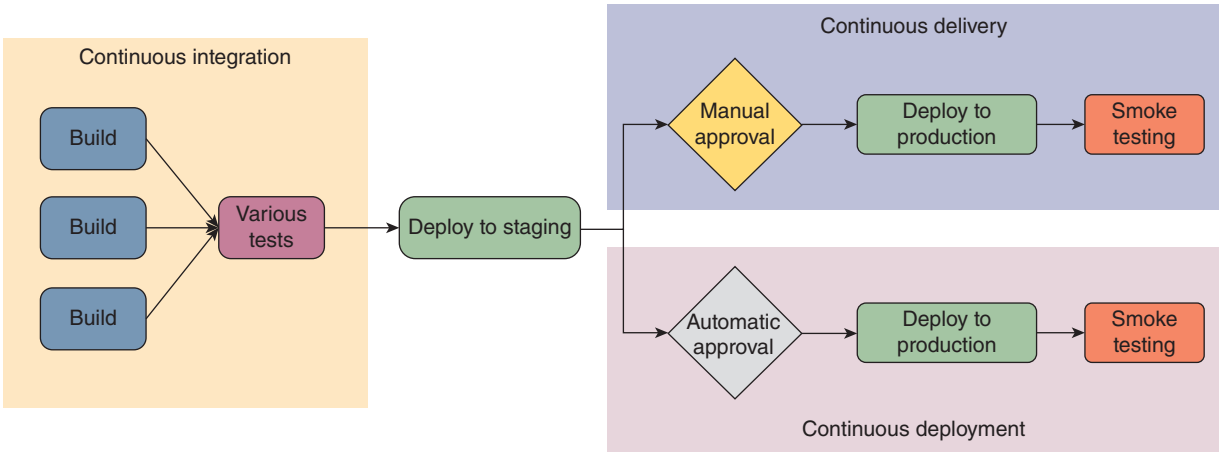
An important part of automation is the ability to deliver frequent and reliable code changes to meet needs that shift quickly and often. This iterative cycling through development phases for smaller, faster, and more adaptive improvements is referred to collectively as CI/CD (continuous integration and continuous delivery or continuous deployment). CI/CD techniques rely heavily on automation to integrate code changes, test these changes, and deploy code changes to production.

More specifically, **CI (continuous integration)** refers to the frequent merging of code changes into the central repository where several tests are performed automatically. Recall that each new version of a project is called a commit. Testing types might include unit testing, which tests individual components such as libraries or files; functional testing, which ensures the code works as expected; and many other types of testing. This frequent testing allows developers to catch issues sooner and fix them more easily.

CD (continuous delivery) pauses the process after testing to wait for manual approval from a human and then delivers the code changes to the next environment in the pipeline, such as a staging environment or the production environment. **CD (continuous deployment)** skips the manual approval step and automatically deploys the code to the next environment as soon as testing is successfully completed (Figure 10-9). After deployment, end-to-end testing can also be automated to confirm the changes are successfully implemented without breaking anything.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

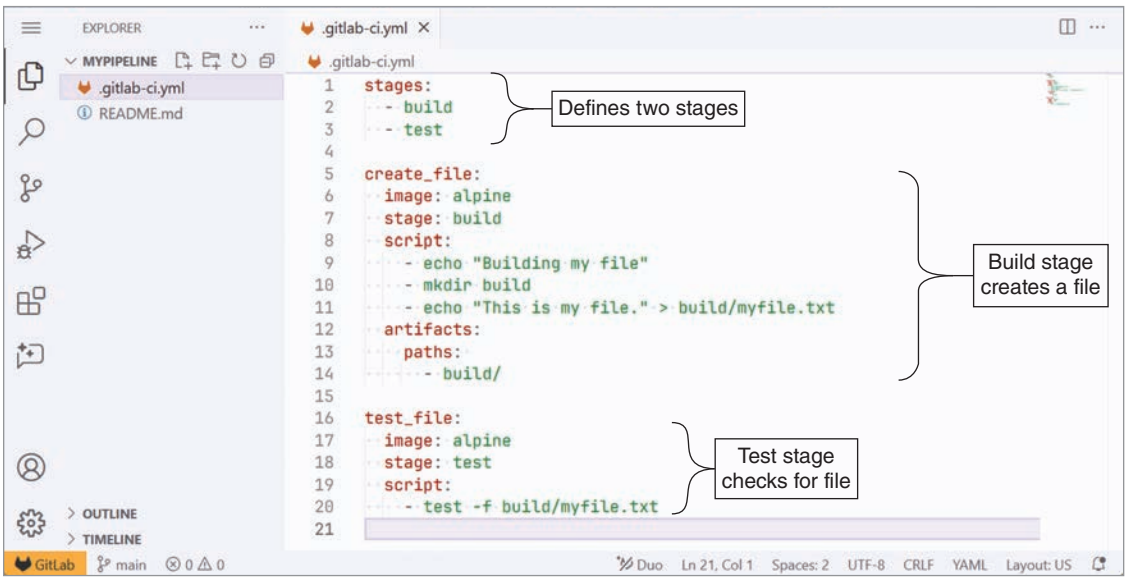
Figure 10-9 Continuous delivery versus continuous deployment



Stages and Artifacts

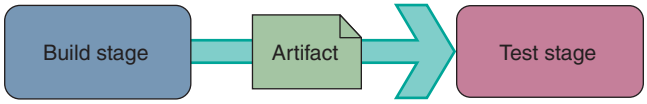
Each stage within a pipeline is defined in the automation code. Figure 10-10 shows YAML code where two stages are defined in a pipeline: build and test. However, as each stage completes, the underlying infrastructure (such as a container) that ran the code is terminated, leaving no connection to the next stage. To connect the flow from stage to stage, you need some kind of file that continues to exist outside the temporary container that created it. In the example shown in Figure 10-10, the build stage creates a file called myfile.txt, which is stored in a directory called build. The test stage then checks that the file exists at the defined location. The file is an artifact that provides continuity from one stage to the next.

Figure 10-10 Two stages in a pipeline



An **artifact** is a single file or a collection of data and files produced by one resource in one stage and consumed by another resource in another stage (Figure 10-11). For example, a container image or VM image can be an artifact in a system pipeline. Exported data might be an artifact in a data pipeline. Source code might be an artifact in a software development pipeline. Especially in the cloud, artifacts are often posted to an artifact repository for centralized access. Artifact repositories can be public or private.

Figure 10-11 Artifacts connect stages

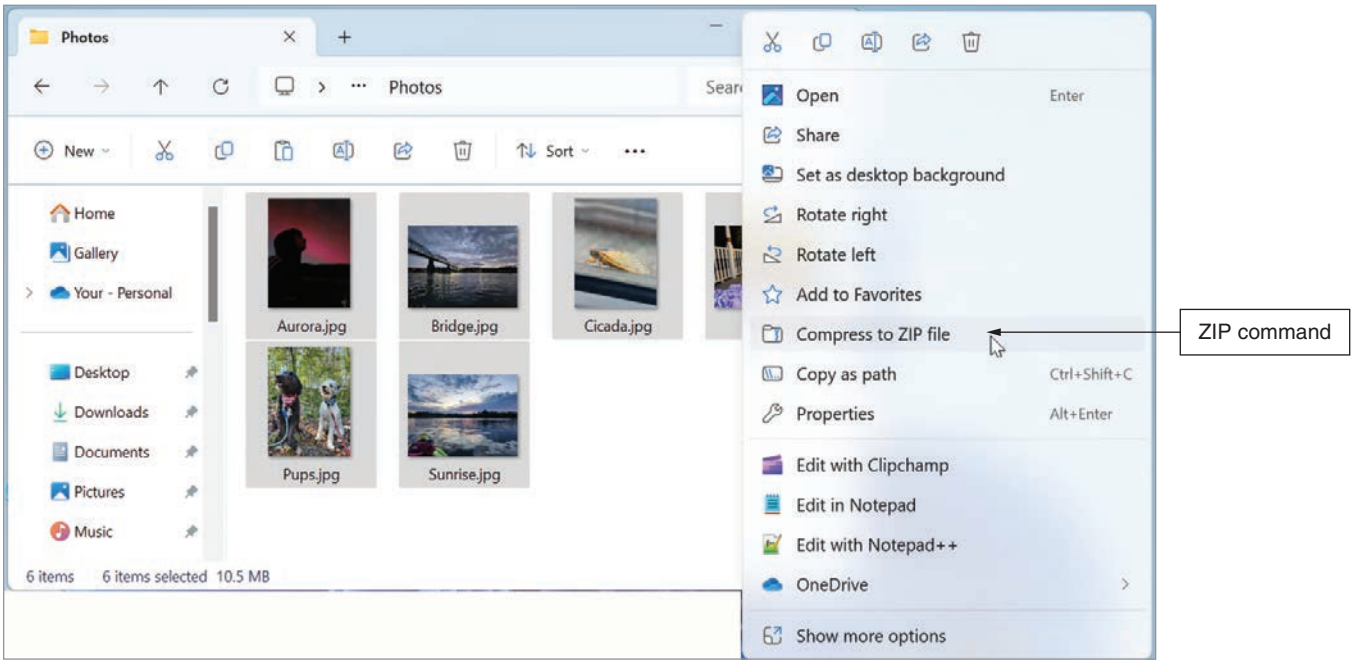


Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A simple artifact might be a basic file. However, most artifacts are more complex. Files that make up an artifact can be collected in a package. Common package types used in software pipelines (and listed in the CompTIA Cloud+ objectives) include the following:

- **Flat file**—You might have experience with exporting data from a data source and being given the option to create a CSV (comma separated values) file. The resulting file is a **flat file** that contains plaintext data in a list of some sort. Data fields and records are separated by a symbol such as a comma, period, space, or hard return. This is a common way to collect data into a single file to transfer that data from one place to another.
- **ZIP**—You're probably already familiar with ZIP (Figure 10-12), which compresses each file individually and then packages them together in a single **ZIP file** (which is usually called a ZIP folder in Windows). The ZIP file contains a directory that allows access to individual files within the ZIP file.

Figure 10-12 Compress files to a ZIP file in Windows



- **Tar**—A **tar (tape archive) file**, also called a tarball, is also a collection of files but they are not compressed. The files are bundled into a single file for easy transport. If the tar file must be compressed, this is performed by another tool, such as gzip, which compresses the entire file as a single entity (as opposed to compressing each included file individually). A tar file is created using the `tar` command, and `tar` is not natively supported by Windows versions before Windows 11. Where ZIP allows easy access to individual files within the ZIP file, tar does not.
- **RPM**—An **RPM (Red Hat Package Manager) artifact** is a special kind of package used to install or update software on Red Hat–based Linux distributions, such as Fedora and CentOS.
- **DEB**—A **DEB (Debian) artifact** is similar to an RPM package but for Debian–based Linux distributions, such as Ubuntu, Kali, and Linux Mint.

Exam Tip | Make sure you understand the differences between the listed package types. On the CompTIA Cloud+ exam, you might be asked to choose the best package type for a particular scenario.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Scripting Logic

Pipelines can perform complex logic as defined in a script. The ability to script pipeline activities allows you to automate much of the administrative work of cloud management. Often, you can use boilerplate code—that is, standard code blocks that perform typical actions or create typical configurations—and piece together the blocks of code that you need to get the results you want. To do this, you need to understand some basic concepts:

- A **variable** is a logical container for information that is stored in code for later use. Typically, a variable is defined by a name and a value, such as `x = 1` or `greeting = 'Hello, world'`. The name for each of these variables is `x` and `greeting`, respectively. The values are `1` and `'Hello, world'`, respectively.
- Notice that one variable example is a number and that the other is text. Variables can be any of several **data types**, which are categories of data that determine what can be done with that data. For example, you can perform mathematical operations on numbers, such as adding or subtracting, but you can't do so with text. Common data types include integers (that is, whole numbers), floating-point (that is, numbers with one or more decimal places), string (that is, text), Boolean (that is, binary values such as true or false), date (such as a date showing the year, month, and day), and array (that is, a list of objects). Each coding language has its own list of available data types and their labels.
- **Operators** are the symbols used to define operations on variables. Common operators you're probably familiar with include + (addition), − (subtraction), * (multiplication), and / (division). Other common operators include == (equal), `!=` (not equal), & (AND), | (OR), and ** (exponent).
- A **function** is a block, or portion, of the code that runs when it is activated. This block of code might receive incoming data, called parameters, and produce outgoing data, called results.
- A **conditional statement** is a statement that must be evaluated to determine the next step. For example, `if a = b` is an if statement. Evaluating the truth or falsity of that statement will indicate what to do next. Consider the following example:

```
x = 5

y = 10

if x > y:

   print("x is greater than y.")

else:

   print("x is not greater than y.")
```

In this statement, the values of x and y could change depending on previous actions. Once those values are set, the code compares the values with the if statement. If the first evaluation is true (that is, if x is greater than y), then the code will print (that is, provide as output) the phrase "`x is greater than y.`" If the first evaluation is not true (that is, if x is equal to y or if x is less than y), then the code will print the phrase "`x is not greater than y.`" Conditional statements can be used to create complex scripting logic that allows a wide variety of actions, responses, and variations from your code.

Note 10-4 | With arithmetic operators, such as addition and subtraction, the usual order of operations applies in code just as it does in math class. You might have learned about this in grade school: PEMDAS (Please Excuse My Dear Aunt Sally). This acronym refers to lower-order arithmetic operations and the order in which they should be performed: anything in parentheses is calculated first, then exponents, then multiplication and division (performed together, left to right in the statement), and finally addition and subtraction (performed together, left to right in the statement).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As you prepare for the CompTIA Cloud+ certification exam, you should read through some samples of code to practice interpreting the logic. Common coding and scripting languages used in cloud environments include Python, Go (Golang), Java, YAML, JSON, SQL (Structured Query Language, pronounced "see-quel"), and JavaScript. A good site for introductory coding concepts and practice exercises is w3schools.com.

Exam Tip | For the exam, most likely you will not be expected to write code from scratch or know detailed syntax rules for any coding language. However, you should be prepared for questions that require you to interpret basic blocks of code, such as Python and YAML. You might need to determine the effects of the code, identify allowed permissions, anticipate next steps, or even find problems with the code.

Applying Concepts 10-1: Working with Python Code

Estimated time: 20 minutes
Objective 2.4: Given a scenario, use code to deploy and configure cloud resources.
Resources:

- A GenAI tool, such as Perplexity, Microsoft Copilot Chat, or Google Gemini
- Internet access

Description: Python is a programming language that can be used to define and direct processes. It is often used in the cloud for a wide variety of tasks, including serverless applications, script automation, and cloud monitoring. This project gives you a basic introduction to the concepts involved in using Python. Respond to the following questions:

1. Using your choice of GenAI tool, ask the tool to give you three metaphors that compare interpreted coding languages and compiled coding languages. Which of these metaphors do you believe does the best job of explaining the differences between these types of languages? Why did you choose this one?
2. Ask the AI tool to explain what an IDE (integrated development environment) is and to give three examples of how IDEs are used in cloud automation. Which of these examples do you think is most relevant to what you're learning in this chapter? What is the connection?
3. In your browser, go to **python.org**. Within the documentation, find the Beginners Guide Overview. At the time of this writing, the document is located at **wiki.python.org/moin/BeginnersGuide/Overview**. Use this information to respond to the following questions:
   a. What other languages are comparable to Python?
   b. Which OSs can run Python?
   c. What website can be used to check Python code?

Use your selected GenAI tool to respond to the following questions:

4. A common test command that programmers practice when first learning how to code is to output the text line "Hello world." What code is needed to perform this task in Python?
5. Tell the AI tool to generate Python code that will fetch current weather data from an API (like OpenWeatherMap), and then display temperature, conditions, and humidity. Also prompt the AI tool to include instructions to you on what to do with the code to implement it. What instructions did the AI tool give you to use the code?
6. Examine the code to understand the basic workflow. What is the text the script would use to request the desired location? What error message would the script output if that location's weather is not available from the programmed data source?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

4. What resource defines the phases of an automated development process?

   a. Pipeline
   b. Commit
   c. Repository
   d. Artifact

5. Which process requires no human intervention so long as the pipeline does not encounter any problems?

   a. Continuous delivery
   b. Continuous integration
   c. Central repository
   d. Continuous deployment

6. Which of these items is subject to PEMDAS?

   a. Operators
   b. Data types
   c. Variables
   d. Functions

☐ Check your answers at the end of this chapter.

Section 10-3: Cloud Maintenance

Certification

3.4 Given a scenario, manage the life cycle of cloud resources.

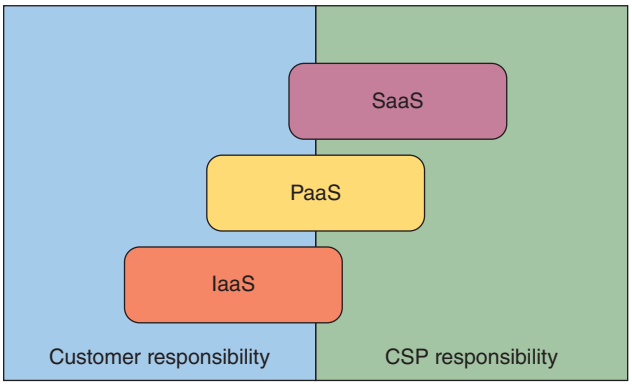
Standard cloud maintenance tasks and maintenance schedules receive much of the focus of automation capabilities. Almost any action that is repeated in basically the same way every time can be automated. In many cases, you can also include orchestration when appropriate. To give you an idea of what's possible, maintenance automation and orchestration tasks could include any of the following activities or many other options: clearing or archiving logs, compressing drives or storage, removing orphaned resources, reclaiming resources as needed, removing stale DNS entries, removing outdated rules from a firewall or security group, removing inactive accounts, or maintaining ACLs for target objects.

Because maintenance activities can impact the performance of affected resources, much of this work must be completed within a maintenance window to minimize disruption to cloud services. Timeliness is also critical when responding to issues related to security or disaster recovery. This reality creates a demand on IT staff that can be time-consuming or even impossible to handle manually.

The responsibility for maintenance tasks will vary by cloud service model, just as the responsibility for security varies. Recall the shared responsibility model you learned about earlier that illustrated the division of security responsibilities. Figure 10-13 illustrates a similar division of labor when it comes to maintenance responsibilities, which shifts according to SaaS, PaaS, or IaaS service models.

Regardless of who is responsible for which processes, automating these tasks means that problems are addressed sooner, patches are applied more consistently across the server fleet, IT staff are freed to focus on more complex tasks directly beneficial to business endeavors, and productivity increases for everyone relying on these cloud services.

Figure 10-13 The distribution of maintenance responsibility varies by cloud service model



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

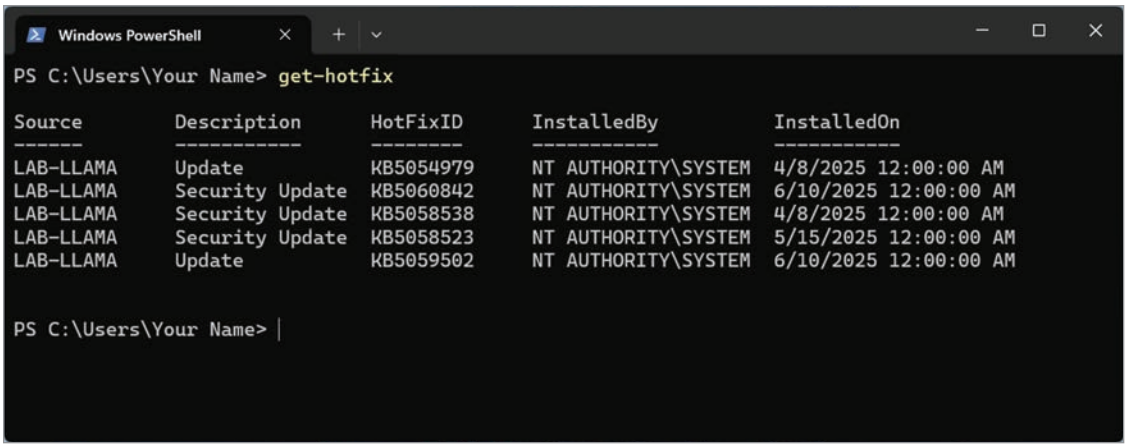
Types of Updates

One of the more complex maintenance processes is applying updates to cloud resources. This is because the update process can disrupt normal business operations as resources are changed, power cycled, or replaced with new deployments. Cloud elements that might require updates in a public or private cloud include virtual machines and server clusters, networking components (such as firewalls or load balancers), storage components, applications, and agents installed on VMs.

Not all updates are created equal. Some updates apply small fixes to minor bugs. Other, more critical updates patch security loopholes and vulnerabilities to protect resources against known threats. The terms used to describe various types of updates are not standardized in the industry and can be used in any number of ways for different purposes. For example, Microsoft releases regular and small updates that might be grouped together over time in a larger update called a build. Updates that address specific security issues are called security updates; updates that address other high-priority issues are called critical updates. Yet, Microsoft updates of all kinds are sometimes referred to as hotfixes.

Figure 10-14 shows output from the `Get-Hotfix` cmdlet (pronounced "command-let," which is a command run in Windows PowerShell). The figure shows a history of applied updates and security updates in Windows. It's important to familiarize yourself with the update terminology used by your organization's vendors so you know which updates are the most critical.

Figure 10-14 Recent Windows hotfixes



Keeping in mind that these terms have variable meanings when used by different vendors, the following list gives a general description of each concept:

- **Hotfix**—For many vendors, a hotfix is typically a minor update that addresses a specific issue, receives less testing, might not require restarting the system, or might be intended only for temporary use. It might not even be released to the public if it's addressing an issue for one or a handful of customers experiencing a unique issue in a given environment. For example, a driver required for a specialized piece of hardware might interfere with normal functioning of an application. A hotfix—either to the driver or to the application—could restore most or all functionality until the next scheduled update, which should include a permanent correction of the problem. One challenge with hotfixes is ensuring that all servers in a cluster receive the same hotfixes. This is especially important in a failover cluster, where a backup server must be able to take over a failed server's workload at any time, and so both servers must be identically configured.
- **Patch**—The terms patch and update often refer to the same thing, or a patch might be a smaller update, more like a hotfix. A patch is released to the public to fix bugs or security loopholes and add new features or enhancements.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Version update**—This is a major update that justifies a new version number. Sometimes the number increases to a new integer, such as going from version 3 to version 4, and sometimes the vendor only adds a decimal place, such as going to version 3.1.
- **Rollback**—The opposite of an update is a rollback. When an update causes a problem, a rollback will revert the system to its state from before the update was applied. Rollbacks are sometimes applied automatically when the system detects an error and, other times, the user initiates the rollback.

Exam Tip | Security professionals have taken the concept of a patch and applied it to vulnerability mitigation. When a new vulnerability is discovered, a collection of rules can be applied to the resource, software, or network to temporarily provide protection from related threats. This is called a **virtual patch**. Virtual patching is often performed by or through a WAF (web application firewall) or an IPS (intrusion prevention system). It's also called external patching or just-in-time patching. The virtual patch can provide protection to the vulnerable system until the vendor's security patch can be obtained and applied.

Patch and Update Methodologies

Patch processes can be difficult to design and implement. Consider that different patch methodologies are required for different kinds of resources. This might be obvious at first, but consider that even different OSs on VMs must be updated from different sources with different settings and steps. You might also need to apply different patch rules and restrictions to servers hosting a particular application or service. This kind of granular control should be built into your update or patch management solution.

Patch or update management across cloud platforms that offer this feature consists of some standard operating procedures, including the following:

- **Installing agents**—Many patch management solutions require that you install an agent that can monitor and control managed resources. In major cloud platforms, such as AWS and others, this agent is often preinstalled on a VM, depending on which image you use to create the VM.
- **Scanning**—Resources (such as servers, applications, or storage systems) must be scanned for missing patches or other known vulnerabilities. This is especially important for open-source software where patches aren't automatically pushed to existing installations. Scans should be conducted at all layers of the stack for which you're responsible: network, operating system, and application.
- **Testing**—You'll want to test any new patches to ensure they won't create new problems in your particular environment. Consider resource dependencies for each element to be patched and how those dependencies will be affected by the update. You'll also need to consider patch dependencies, which means that some patches might need to be installed before other patches. These dependencies sometimes necessitate a specific order of operations.
- **Choosing update targets**—You can patch resources in groups to better control which resources get which updates and when. AWS calls these patch groups. For example, a web server cluster might be able to handle current updates with no problem, while current updates might "break" a server hosting a legacy application. The tags feature you learned about previously works well for this purpose. You might tag VMs according to their environment (such as testing, staging, or production) or desired patch window (such as "Sunday night").
- **Considering backups . . . and more backups**—You might schedule a backup of data to occur just before updates are applied, which is recommended by most OS and software makers. To back up services, you might stagger maintenance windows for different groups of servers so one group can be updated while another group handles the active workload. This way, only a few servers at a time are exposed to the inherent unpredictability of applying updates.

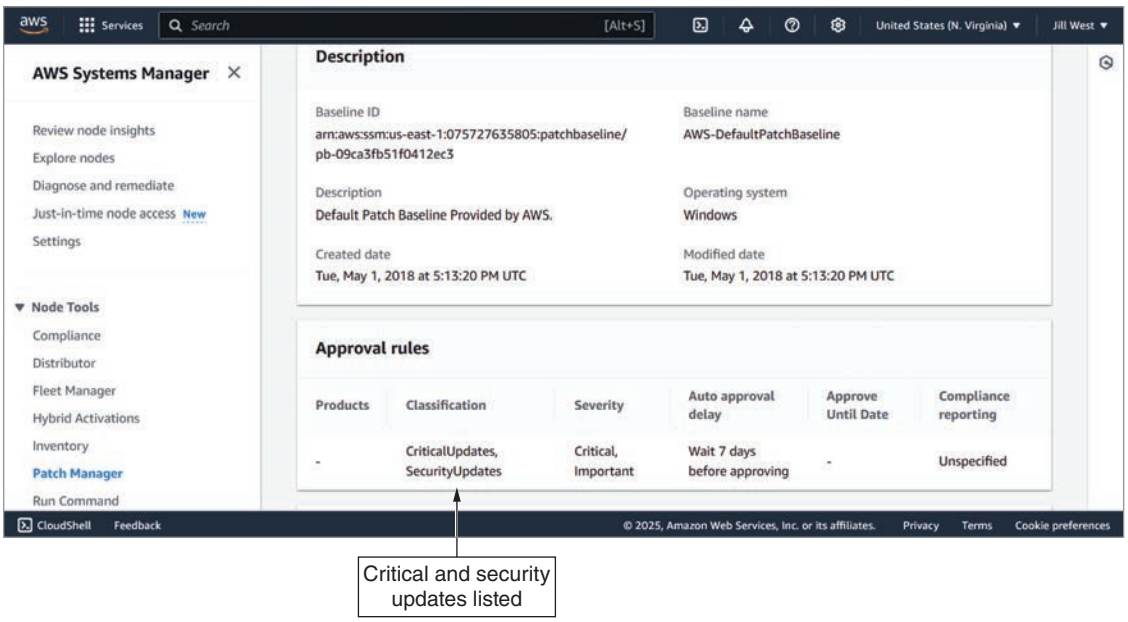
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Scheduling maintenance windows**—Applying patches can interfere with performance; therefore, maintenance windows must be carefully selected. Other factors affecting maintenance window schedules include the following:
  - When planning a maintenance window, you might want to spread scheduled updates across multiple maintenance windows. For example, you could update inactive, failover servers first as a test run before updating active servers.
  - Often, you can also choose what percentage of servers should be running update processes at any given time and at what point before the maintenance window closes no more update processes should be initiated.
  - Determine how many errors may occur before the update process should be halted and an admin notified.
  - You'll also need to consider how critical pending updates might be—patching a critical security vulnerability will need to be completed sooner and more quickly, even if that means taking a harder hit on performance.
- **Monitoring compliance and completion rates**—The patch management solution provides a dashboard or some other reporting feature to show which patches were successfully applied and which machines are still out of compliance.

Patching in AWS

AWS's Patch Manager (a feature of the AWS Systems Manager service) can be used to automatically scan managed instances for patch compliance and, if desired, apply updates to these instances. You can use patch baselines to manage approved or rejected patches. Figure 10-15 shows some of the configurations for the default Windows patch baseline. With this predefined baseline, critical and security updates will be automatically applied to managed Windows EC2 instances seven days after updates are released. You can also set maintenance windows to automate patch processes. Patch Manager is integrated with AWS IAM and CloudWatch for increased automation and notification capabilities as well as auditing abilities.

Figure 10-15 Approval rules for Windows patches



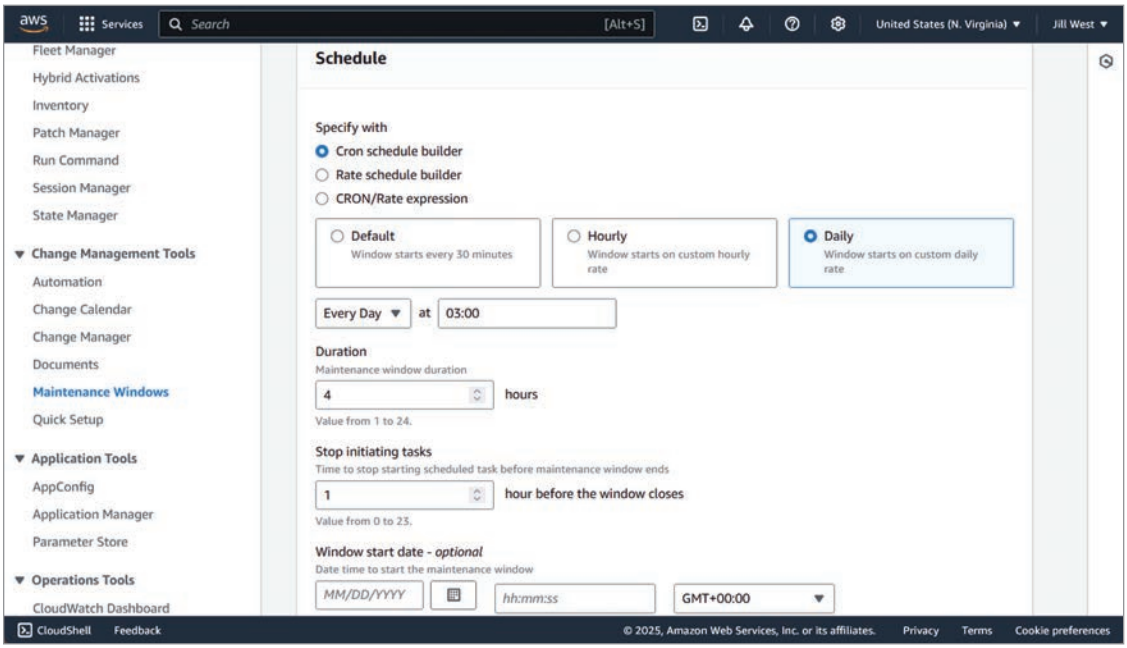
Source: Amazon Web Services, Inc.

To function, Patch Manager relies on an agent preinstalled on many of the more popular AMIs (Amazon Machine Images), such as Windows Server, Amazon Linux, and Ubuntu Server. When scheduling automated update tasks, know that the agent itself needs to be regularly updated as well. This task can also be automated.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Automated patch tasks occur within restrictions defined by a maintenance window. Figure 10-16 shows some of the available settings for a maintenance window in AWS. You can schedule frequent windows, such as every 30 minutes, or schedule windows further apart, such as every Sunday. You can also decide at what point within the maintenance window new tasks will no longer be initiated. This is helpful to prevent time-consuming updates or other tasks from starting moments before the window closes.

Figure 10-16 Configure maintenance window start frequency and duration

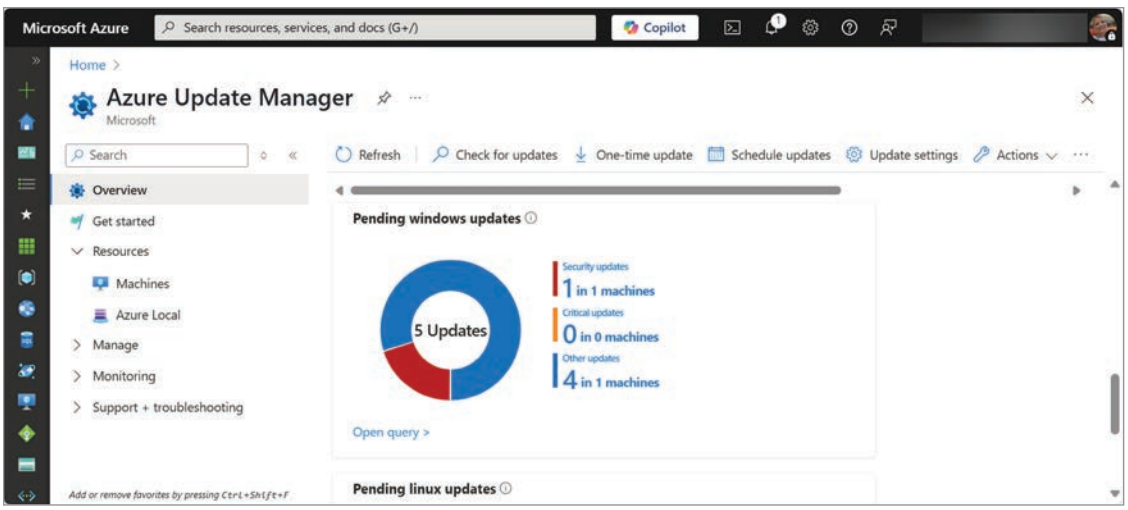


Source: Amazon Web Services, Inc.

Patching in Azure

Update management in Azure is provided through the Azure Update Manager service. Similar to AWS, Update Manager requires an agent to be installed on each target machine. Figure 10-17 shows the analysis of pending Windows updates for a managed VM that is missing a security update.

Figure 10-17 An automated update is pending



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Patching in GCP

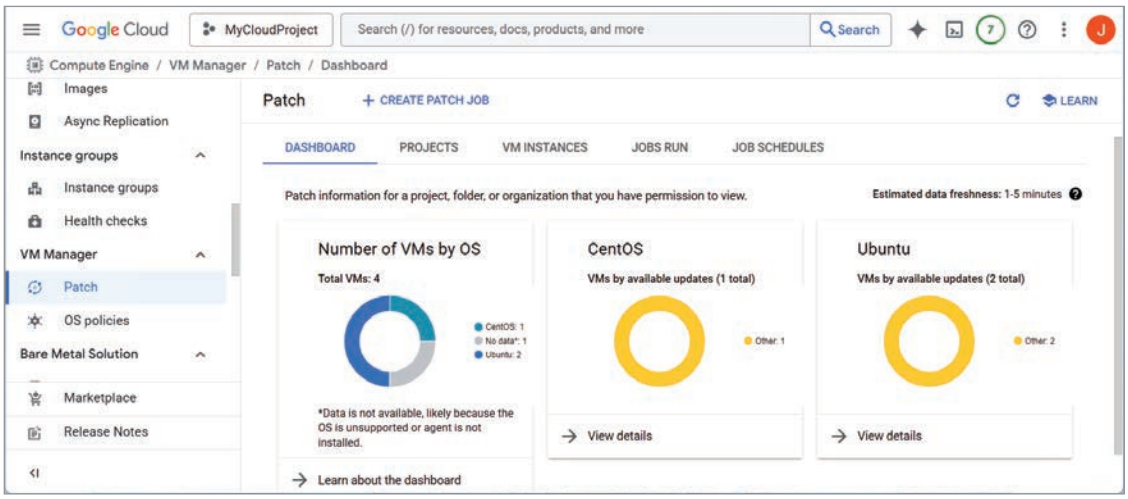
In 2021, GCP released a new suite of tools designed to manage updates for VM fleets. The service is called VM Manager, and it consists of three components, as follows:

- **Patch**—Applies patches to VM OSs on demand or by schedule and reports activities and patch compliance for the entire VM fleet.
- **OS inventory management**—Presents information on installed OSs.
- **OS policies**—Installs, removes, or updates software.

When creating a project, you can decide if VM Manager should be activated by default on all VMs or manually on selected VMs. Most VM images in Compute Engine already contain the OS Config agent, and this agent must be activated on each VM to enable communications with VM Manager. You can install the agent on other images as well. The agent uses built-in OS utilities to retrieve files for an update, apply the update, adjust the VM's state, or collect logs on the VM.

Figure 10-18 shows the Patch dashboard with three VMs needing updates.

Figure 10-18 OS patch management in VM Manager



Source: Google LLC

Decommissioning

Deprecating a product at the end of its life cycle can consist of many phases. For example, consider the difference between EOL (end of life) and EOS (end of support):

- **EOL (end of life)**—The date past which the product is no longer sold and service subscriptions can no longer be renewed. Support might still be available, especially for security issues; however, continuing support might cost more. This can give customers time to consider their options for updating or replacing it with a new product.
- **EOS (end of support)**—The date past which support is no longer provided for the product. Customers are advised to cease all use of the product once support has ended, as security issues could surface that will no longer be addressed by the manufacturer.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exam Tip | Unfortunately, the EOS acronym also sometimes refers to "end of sale," which is similar to EOL. Sometimes these terms—"end of life" and "end of sale"—are used synonymously. If there is a differentiation, typically the difference refers to a phase versus a date. Whereas EOL might refer to an entire life cycle phase of a product (including announcements for coming dates, such as end of sale and end of support), EOS might refer specifically to the last date a product can be purchased through a vendor's or manufacturer's sales pipeline.

Read vendor documentation and announcements carefully to know how they're using various acronyms and specifically what information is provided in their announcements. Be sure to clarify how each acronym is used in a particular scenario. As you're studying for the CompTIA Cloud+ exam, notice that the objectives refer to "end of support," not "end of sale." For this reason, if you get an exam question referring to EOS and the question does not provide additional guidance, then the acronym most likely refers to end of support.

Self-Check

7. What factor primarily determines the division of responsibility for maintenance tasks?

a. Maintenance window
b. Service model
c. Criticality of update
d. Source of update files

8. What is the opposite of an update?

a. Virtual patch
b. Hotfix
c. Build
d. Rollback

Check your answers at the end of this chapter.

Section 10-4: Troubleshooting Automation Issues

Certification

6.1 Given a scenario, troubleshoot deployment issues.

When deploying resources through IaC, troubleshooting problems becomes more efficient, and the problems themselves are more easily traced. For example, consider how to troubleshoot a server configuration issue. If the server worked correctly previously, you can examine the IaC script to determine what change to the server's configuration caused the problems. If it has never worked correctly, examining the script responsible for deploying the server might reveal a misconfiguration or a poorly defined environmental variable. Fixing the script can result in repeatable deployments free of that particular problem.

Common Automation Issues

Other problems you might encounter with automation processes include the following:

- **Deprecated features**—As newer technologies emerge, older features are deprecated. Vendors regularly publish notifications on pending changes to features or deprecation of older features.
- **API version incompatibility**—To pressure customers to upgrade their systems, vendors sometimes remove compatibility of their REST APIs with older client systems. As the customer, again, you can rely on published information to anticipate coming changes and adjust IaC scripts and templates.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- **Outdated component definitions**—Automated deployments often refer to software components, such as a library, framework, or OS. If one of these components is outdated, meaning it's no longer maintained or supported, this could leave a serious vulnerability or result in incompatibilities with other systems. Make sure to regularly update and patch all software components of cloud deployment automations.
- **Patch failure**—When applying updates to adapt to feature and API changes, sometimes the patch itself fails. For example, the resource receiving the patch might not have enough available storage space or conflicting patch processes might be running at the same time. Typically, systems are designed to revert to a previous configuration if an automated patch process fails and to also send an alert to admins.
- **Change management failure**—The purpose of change management is to incorporate accountability for changes made to a system. When automation plays a role in changes made, sometimes the automated changes fall outside the purview of change management oversight, and sometimes change management approval doesn't reach far enough to anticipate automatic adjustments. Solving this problem requires a big-picture approach to the implementation of cloud automation so that proper accountability continues to apply while allowing automation tools to do the job they're designed to do.

AI in IT | Building code for IaC workflows is susceptible to problems from syntax errors, missing symbols, illogical dependencies, or drift management challenges. Similar to how AI can provide assistance to software developers as they write code, AI tools can also help prevent errors in IaC scripts and templates. For example, SCA (static code analysis) and SAST (static application security testing) can analyze code without executing it to identify bugs, vulnerabilities, or other security risks. Compliance checks can help ensure that all compliance requirements are met, even as these requirements continue to change or when applied requirements must be adapted to different geopolitical locations. Automated remediation can suggest or apply fixes for detected issues. Continuous monitoring can ensure the ongoing health and stability of code-defined infrastructure.

Breakdowns in Workflow

Unfortunately, automation is not a failproof system. Things can still go wrong, especially during the initial setup phase. Automated systems rely on predictable configurations to find resources for monitoring or actions to trigger. Successful workflows also rely on well-planned processes and user training. Consider the following obstacles to the establishment of successful automation workflows:

- Don't rely on users' fascination with new technologies to ensure that users follow the correct steps to initiate automated processes. Adequate training and follow-through are essential. For example, a sophisticated trouble ticketing system isn't much help if employees continue to email or text IT staff when there's a problem instead of filing a trouble ticket through the system. In cases such as this, gentle and consistent reminders, along with assistance in learning to use the system and the eventual rejection of all manual requests, can help users become familiar with a new system. Additionally, ensuring the new system works well with quick responses from IT staff can demonstrate to users how the new system will help enhance productivity.
- Incomplete automated workflows apply automation for part of the process but not all of it. For example, an automated approval process for purchase orders that relies on manual transfer to your financial software is not true automation. You'll experience bottlenecks in the workflow and continuing issues with errors. Design workflows to complete a process from start to finish.
- Automating a workflow that is already not working well will only make it worse. Fix the broken process first, then consider how to automate it.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- No workflow will stay the same forever. Things change, and the workflow must be adapted for new variables. Review your automations, workflows, and system output regularly for signs of problems. Sources of these problems might include any of the following:
  - Account mismatch issues, such as updated credentials
  - Server name, IP address, or location changes, which might be caused by a change management failure
  - Version or feature mismatch
  - Job validation issues, where the automation system is unable to validate the workflow
  - Batch job scheduling issues, where a task dependency is not scheduled to occur at the right point in the process
  - Automation tool incompatibility, where tools are unable to adequately communicate with each other

A well-tuned automation system will save you time and save your company money by increasing efficiency and reducing risk and loss from human error. Training yourself on how to better automate and orchestrate your cloud processes will also help ensure your employability as cloud computing becomes the new status quo.

Exam Tip | Studying for and earning your CompTIA Cloud+ certification is an entry point into your career as a cloud professional. You've laid a solid foundation in your ability to learn and work with the cloud. Yet there is still much more to learn, with the related technologies changing and maturing every day. No one is an expert in all areas of the cloud, so don't let what you don't know intimidate you. Keep asking good questions, researching new developments, and challenging yourself to explore and experiment with cloud services and platforms.

Self-Check

9. An improperly documented server name change is a type of _____.

a. patch failure
b. deprecated feature
c. API version incompatibility
d. change management failure

10. What is caused by manual portions of a workflow?

a. Account mismatch
b. Bottleneck
c. Job validation
d. Task dependency

Check your answers at the end of this chapter.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

Section 10-1: Automation

- Automation in the cloud provides faster deployments and responses to problems, better control, lower costs, and increased security.
- With IaC, the steps needed to configure or change a cloud environment are written in code and are run collectively at one time rather than requiring the human admin to complete one step of the deployment at a time in a GUI.
- CaC can help prevent configuration drift, which is the changes to configurations that cause them to exceed the limitations of compliance requirements.
- Available characteristics of provisioning and CM tools include push model versus pull model, immutable versus idempotent, and imperative versus declarative.
- Using a CLI is one way to create scripts for deploying cloud resources, which provides the option to automate many tasks.

Section 10-2: Automating Code Pipelines

- CI/CD techniques rely heavily on automation to integrate code changes, test these changes, and deploy code changes to production.
- Common coding and scripting languages used in cloud environments include Python, Go, Java, YAML, JSON, SQL, and JavaScript.

Section 10-3: Cloud Maintenance

- Standard cloud maintenance tasks and maintenance schedules receive much of the focus of automation capabilities.
- Some updates apply small fixes to minor bugs, while other, more critical updates patch security loopholes and vulnerabilities to protect resources against known threats.
- Granular control of patch processes should be built into an update or patch management solution.
- Deprecating a product at the end of its life cycle can consist of many phases, such as EOL and EOS.

Section 10-4: Troubleshooting Automation Issues

- Common automation issues include deprecated features, API version incompatibility, outdated component definitions, patch failures, and change management failures.
- Automated systems rely on predictable configurations to find resources for monitoring or actions to trigger.

Key Terms

For definitions of key terms, check the Glossary.

Ansible
artifact
CaC (configuration as code)
CD (continuous delivery)
CD (continuous deployment)
CI (continuous integration)
CM (configuration management) tool
conditional statement
data type
DEB (Debian) artifact
EOL (end of life)
EOS (end of support)
flat file
function
IaC (infrastructure as code)
operator
provisioning tool
RPM (Red Hat Package Manager) artifact
runbook
script
tar (tape archive) file
template
Terraform
variable
virtual patch
workflow
ZIP file

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Acronyms Checklist

The following acronyms are listed in the CompTIA Cloud+ objectives and could appear on the Cloud+ exam. This means that exam questions might use any of these acronyms in context so that you must know the meaning of the acronym in order to answer the question correctly. Make sure you're familiar with what each acronym stands for and the general concept of the term itself. All these acronyms are used in context in this chapter.

CaC configuration as code
CI/CD continuous integration/continuous deployment
IaC infrastructure as code
RPM Red Hat Package Manager

Review Questions

1. Which kind of tool relies most on disposable resources?
   a. Declarative
   b. Imperative
   c. Idempotent
   d. Immutable
2. When testing a new automation script, the deployment of a web server repeatedly fails with a "Resource missing" error. What characteristic of the script is most likely responsible for the problem?
   a. Security
   b. Testing
   c. Patching
   d. Sequencing
3. Which technique specifically requires frequent code commits?
   a. Continuous deployment
   b. Continuous delivery
   c. Continuous integration
   d. Continuous improvement
4. Which of the following tools requires information on your desired outcome rather than detailed instructions on how to deploy a group of resources? **(Select two.)**
   a. Ansible
   b. Terraform
   c. CloudFormation
   d. Cloud Shell
5. What kind of resource informs the code in one pipeline stage of the outcome from code in a previous pipeline stage?
   a. Artifact
   b. Repository
   c. Build
   d. Test
6. Considering the following code block, what is the line x=5?

```
x = 5
y = 10
if x > y:
   print("x is greater than y.")
else:
   print("x is not greater than y.")
```

   a. Operator
   b. Variable
   c. Conditional statement
   d. String
7. What component in code is used to branch the logic in a script?
   a. Conditional statement
   b. Variable
   c. Operator
   d. Data type
8. Which artifact type allows for accessing individual files within the package without extracting the entire package to do so?
   a. Tar
   b. DEB
   c. ZIP
   d. RPM
9. Which of the following can be used to detail IaC resource configurations for automated deployments? **(Select two.)**
   a. YAML
   b. JSON
   c. XML
   d. HTML

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10. Which of the following lines contains a Boolean data type?

```
1 is_raining = True
2 if is_raining:
3    print("Take an umbrella.")
4 else:
5    print("No umbrella needed.")
```

a. Line 1
b. Line 2
c. Line 3
d. Line 4

Scenario-Based Questions

Question 10-1

Terry is preparing a lab activity for his students where they'll be practicing patching techniques on a variety of operating system configurations. Terry needs to deploy several VMs in the lab, and he'll need to be able to make multiple changes to the VM configurations during the course of the activity. Which of the following tools would be most useful for automated deployment in this situation?

a. Ansible
b. CloudFormation
c. Python
d. Terraform

Question 10-2

Mateo is troubleshooting a recurring issue with one of his applications running on a private cloud in his company's on-prem network. It seems the OS patches are not being applied properly, with frequent errors showing the VM is not in compliance with his patch management policies. Mateo is searching for indications of other patches that need to be applied to fix this problem. Which of the following components most likely needs to be updated?

a. The underlying hypervisor
b. The local database
c. The running application
d. The local agent on the VM

Question 10-3

Dominique and her team are ready to deploy updates to a core business application running in their organization's cloud. They've spun up a second, identical environment to their production environment, applied the updates, and completed testing. Because of the nature of the updates, however, the team has decided there's a moderate risk of problems during the deployment even after all the prep work they've done. Dominique has asked for suggestions on how to minimize the impact of these anticipated issues should they occur. Which of the following strategies would best ensure the deployment goes smoothly with minimal impact on users while giving Dominique's team the time and opportunity they need to resolve issues that might arise?

a. Create a third environment for additional testing.
b. Update a few servers in the original environment that is handling live traffic.
c. Transfer a small percentage of live traffic to the updated environment.
d. Create a third environment that is given only the portion of updates least likely to cause problems.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hands-On Projects

Note 10-5 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Project 10-1: Using Code in AWS

Estimated time: 20 minutes
Objective 2.4: Given a scenario, use code to deploy and configure cloud resources.
Resources:

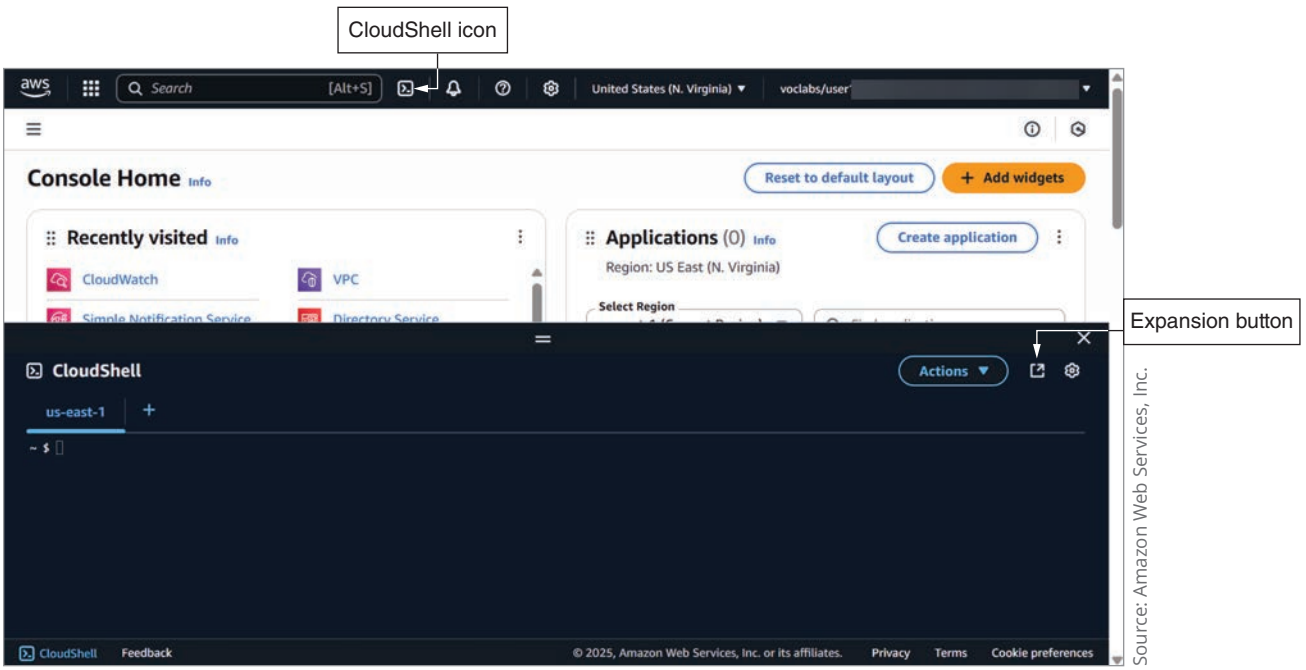
- AWS account
- Internet access

Description: CloudShell is an AWS service that allows you to use a CLI from within the AWS Management Console to interact with resources and services. Practicing with CloudShell will give you some idea of how IaC and scripting work. In this project, you'll enable CloudShell and practice some commands, including creating an EC2 instance.

Complete the following steps to enable and practice using the CloudShell environment in your AWS console:

1. In the lower-left corner of your console, click **CloudShell**. AWS will initialize the CloudShell environment for your account in a new pane at the bottom of your console. Click the expansion button to open CloudShell in a new browser tab (Figure 10-19). Read the information in the pop-up box, and then click **Close**.

Figure 10-19 Open CloudShell in a new tab



Source: Amazon Web Services, Inc.

2. Enter **`aws help`**. What command gives a list of available help topics?
3. Press the **down arrow** or the **Spacebar** to continue scrolling through the help output. Type **q** when you're ready to return to the shell prompt.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. Refer to Table 10-1 earlier in the chapter for some common AWS commands. Request information on the EC2 service. What command did you enter?
5. Again type **q** to return to the shell prompt.

Now you're ready to create an EC2 instance using CloudShell. Complete the following steps:

6. Use CloudShell to deploy an EC2 instance. You'll need to collect the following information to create your command. Some of this information will be available within your AWS account—you can gather this information from the Console Home tab you already have open. Some of this information might be available with an online search.
   - Start with the command text:

     ```
     aws ec2 run-instances --image-id ami-xxxxxxxx --count 1 --instance-type
     t3.micro --key-name MyKeyPair
     ```

   - Replace the AMI text in the sample command with the AMI ID for a free-tier Windows Server system. Where did you find this information?
   - Replace the key name with the name of a key pair in your account.
   - What command did you run that successfully deployed an EC2 instance?

Note 10-6 | It might feel inconvenient to find the required information and type a long command the first time you run a command like this, such as finding an AMI ID or an instance type. However, once you collect this information, you can repeat this command many times much faster than working through the console over and over. Once your command is created, you can repeat this activity as many times as needed, knowing that every resource will receive exactly the same configuration information.

7. Type **q** to quit from the output about your new instance. Enter the command **`aws ec2 describe-instances`** to confirm your instance is listed. What is your instance's instance ID?
8. Type **q** to quit from the output.
9. Now you're ready to delete the resources created in this project. In CloudShell, enter the following command:

   `aws ec2 terminate-instances --instance-id <your-instance-id>`

10. Use a CloudShell command to show that your instance is being shut down. **Take a screenshot** of the command you used and the portion of the output showing a shutdown is in progress; submit this visual with your responses to this project's questions.
11. Check through your account to confirm that all related resources have been deleted. Make sure to properly sign out of your account.

Note 10-7 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 10-2: Using Code in Azure

Estimated time: 30 minutes
Objective 2.4: Given a scenario, use code to deploy and configure cloud resources.
Resources:

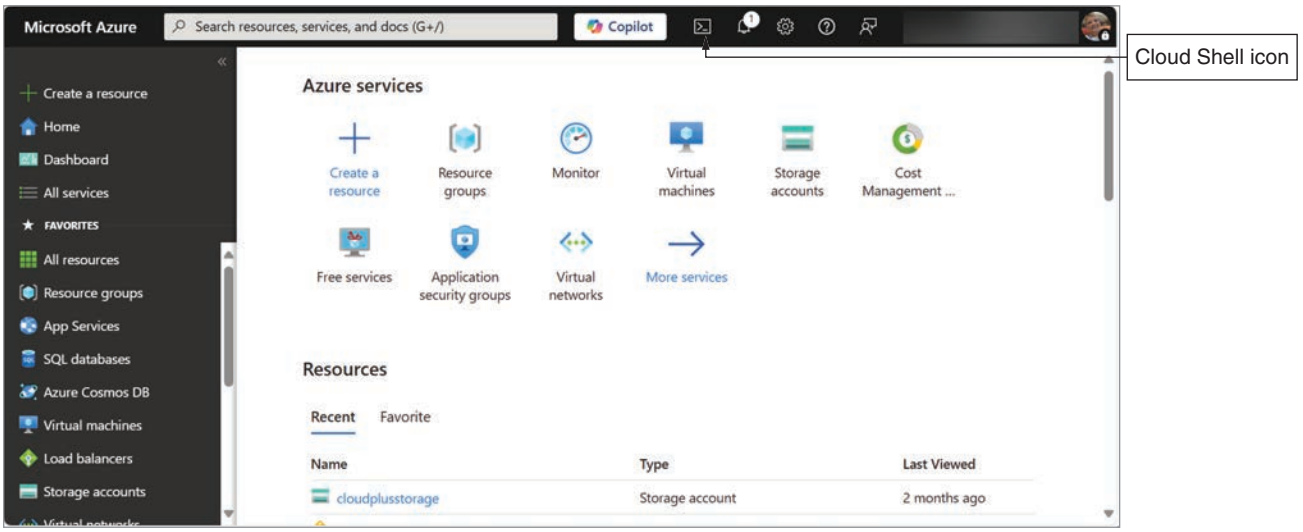
- Azure account and subscription
- Internet access

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Description: Cloud Shell is an Azure service that allows you to use a CLI from within the Azure portal to interact with resources and services. Practicing with Cloud Shell will give you some idea of how IaC scripting works. In this project, you'll enable Cloud Shell and practice some commands, including creating a VM. Complete the following steps:

1. In your Azure portal, click the **Cloud Shell** icon at the top of the screen (Figure 10-20), and select the **Bash** shell. Bash is a Linux-native CLI environment.

Figure 10-20 Open Cloud Shell



2. Azure Cloud Shell needs a storage account to retain files between sessions; however, this feature will not be required for this project. Leave the default selection at "No storage account required." Choose your subscription, and then click **Apply**.
3. When the terminal connects, enter the command **`az version`**. Notice that Azure also gives version information for its components, such as AI examples and SSH. What version of Azure CLI is your Cloud Shell using?
4. If you are using an Azure for Students subscription, enter the command **`az account set -subscription 'Azure for Students'`** to avoid potential problems later in this project. If you're using a standard subscription, this step is not required.
5. Enter **`az --help`**. Near the top of the list, what command allows you to manage your Azure subscription information?
6. Refer to Table 10-2 earlier in the chapter for some common Azure commands. Request some commonly used commands containing the text "vm" using the **`find`** command. What command did you enter?

Before creating other resources, you need to create a new resource group. It's important to place this resource group in a region that supports the type of automation used in this project. For simplicity's sake, you'll use West US 2. Complete the following step:

7. Use Cloud Shell to create a resource group. Enter the following command, using your own name for your resource group:

   ```
   az group create -l westus2 -n MyNewRG
   ```

8. Confirm the resource group was created. How did you do this?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Now you're ready to create a VM using Cloud Shell. Complete the following steps:

9. You'll need the following information to create your command:
   - Start with the command text:

     ```
     az vm create -n NewVM -g MyNewRG --image Win2019Datacenter --size
     Standard_B1s --public-ip-sku Standard --admin-username thisisme
     --admin-password MyP@ssword123
     ```

   - Replace the resource group name with your resource group you created in Step 6.
   - If desired, you can replace the provided text with other text, such as the VM's name or the admin password.
   - What command did you run that successfully deployed a VM instance?

Note 10-8 | It might feel inconvenient to find the required information and type a long command the first time you run a command like this, such as finding an image name or instance size. However, once you collect this information, you can repeat this command many times much faster than working through the console over and over. Once your command is created, you can repeat this activity as many times as needed, knowing that every resource will receive exactly the same configuration information.

10. Enter the command **`az vm list-ip-addresses`**, which will confirm the VM was created and provide networking information for the VM. **Take a screenshot** of the output showing your VM exists; submit this visual with your responses to this project's questions.
11. Now you're ready to delete the resources created in this project. In Cloud Shell, enter the following command, replacing the appropriate text with your resource group name and your VM's name:

    ```
    az vm delete -g MyNewRG -n NewVM --yes
    ```

12. Use Cloud Shell to show that your VM has been deleted.
13. Use Cloud Shell to delete the resource group you created in Step 7, and then use Cloud Shell to confirm this resource group was deleted. Note that a residual resource group, NetworkWatcherRG, should remain in your account and not be changed or deleted. What commands did you use?
14. Check through your account to confirm that all other resources have been deleted. Make sure to properly sign out of your account.

Note 10-9 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Project 10-3: Using Code in GCP

Estimated time: 20 minutes
Objective 2.4: Given a scenario, use code to deploy and configure cloud resources.
Resources:

- GCP account
- Internet access

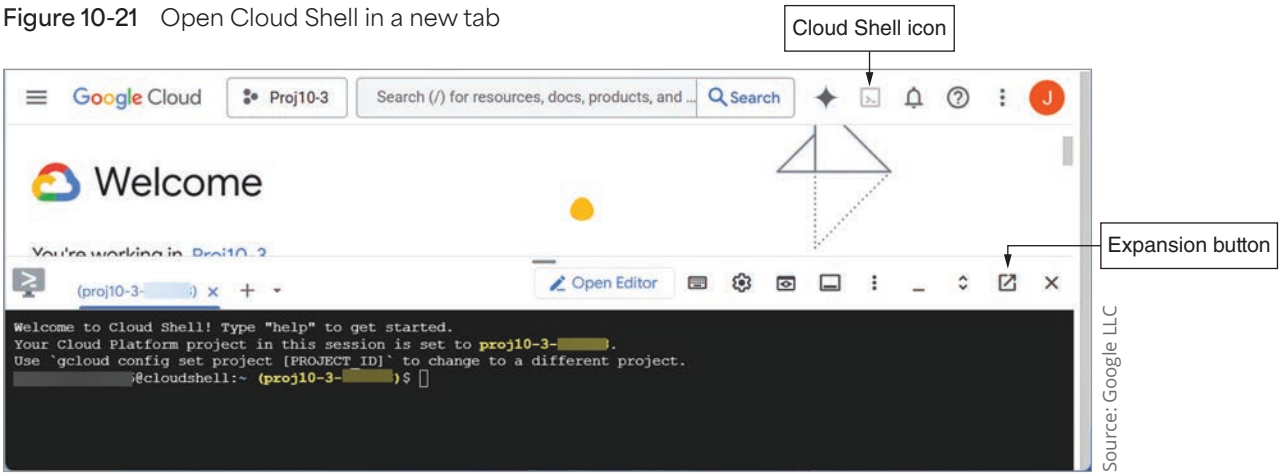
Description: Google Cloud Shell is a GCP service that allows you to use a CLI from within the GCP console to interact with resources and services. Practicing with Cloud Shell will give you some idea of how IaC scripting works. In this project, you'll enable Cloud Shell and practice some commands, including creating a Compute Engine instance. You'll then use the console to configure automated patching on your instance.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Complete the following steps:

1. In your GCP console, create a new project, and switch to the new project.
2. At the top of your GCP console, click **Activate Cloud Shell**. When the Cloud Shell pane opens, if necessary, close the information box by clicking **Got it!** In the Authorize Cloud Shell box, click **Authorize**. Click the expansion button to open Cloud Shell in a new browser tab (Figure 10-21).

Figure 10-21 Open Cloud Shell in a new tab



3. Enter the command **`gcloud version`**. Notice you're given version numbers for several components. What is your version of Google Cloud SDK? What is your version of gsutil?
4. Enter the command **`help`**. How long will the VM hosting your session persist after closing the session?
5. Enter the command **`gcloud help`**. Press the **Spacebar** to keep moving forward. When you're ready, type **q** to quit.
6. Refer to Table 10-3 earlier in the chapter for some common gcloud commands. Request information on any running instances. Authorize use of your credentials or enable APIs, if needed. What command did you enter?

Now you're ready to create a new VM instance using Cloud Shell. Complete the following steps:

7. Use Cloud Shell to deploy a Compute Engine instance. You'll need the following information to create your command:
   - Start with the command text:

     ```
     gcloud compute instances create my-instance --zone=us-east1-b --image-family=ubuntu-2004-lts --image-project=ubuntu-os-cloud --machine-type=e2-micro
     ```

   - If desired, you can replace the instance name with a different name.
   - What command did you run that successfully deployed a VM instance?

Note 10-10 | It might feel inconvenient to find the required information and type a long command the first time you run a command like this, such as finding an image name or machine type. However, once you collect this information, you can repeat this command many times much faster than working through the console over and over. Once your command is created, you can repeat this activity as many times as needed, knowing that every resource will receive exactly the same configuration information.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8. Use a command in Cloud Shell to confirm the VM was created. **Take a screenshot** of the output showing your VM exists; submit this visual with your responses to this project's questions.
9. Now you're ready to delete the resources created in this project. In Cloud Shell, enter the following command, replacing the appropriate text with your VM's name and zone:

```
gcloud compute instances delete my-instance --zone=us-east1-b
```

10. Use Cloud Shell to show that your VM has been deleted. What command did you use?
11. Use Cloud Shell to delete the project you created in Step 1. What command did you use?
12. Confirm the project is shut down and scheduled for deletion. Make sure to properly sign out of your account.

Note 10-11 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Challenge Projects

Note 10-12 | Websites, applications, public cloud platforms, and related account options change often. While the instructions given in these projects were accurate at the time of writing, you might need to adjust the steps or options according to later changes.

Note to Instructors and Students: A rubric is provided for evaluating student performance on these projects. Please refer to Appendix D.

Challenge Project 10-1: Create a Cloud Deployment

Estimated time: 90 minutes (+20 minutes for group work, if assigned)
Objective 2.5: Given a set of requirements, provision the appropriate cloud resources.
Group work: This project includes enhancements when assigned as a group project.
Resources:

- Public cloud account (such as AWS, Azure, or GCP)
- Drawing app, such as Visio or app.diagrams.net
- Internet access

Description: Throughout this course, you've learned to deploy compute, networking, security, and storage resources. In this final Challenge Project, you will combine your skills from previous chapters to create a comprehensive cloud deployment. Complete the following steps:

1. Choose a public cloud platform. Ensure you have MFA enabled on your root account, if possible. (Some educational accounts might not be able to complete this step.) **Take a screenshot** showing where MFA is enabled; submit this visual with your responses to this project's questions.
2. Create a storage container or bucket that contains an index.html file for a static website. Use the sample website text from Challenge Project 8-1, "Host a Static Website," as a starting point. However, make significant changes to the code to ensure your website is unique and reflects some of the information you have learned in this course. For example, you should change the name of the webpage and insert text identifying yourself. You might also include a statement or two about what you have learned, what more you would like to learn, or how this webpage fits into your cloud deployment for this Challenge Project.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Open the webpage in a separate tab and **take a screenshot** of your webpage; submit this visual with your responses to this project's questions.

3. Create a VPC, VNet, or subnet with Internet access. Refer to Chapter 4, "Cloud Networking," for guidance on how to provide access to the Internet for resources within a VPC, VNet, or subnet. For example, in AWS, you need an Internet gateway and a related route. Create a second VPC, VNet, or subnet that does *not* have access to the Internet. Make sure the IP address spaces for these two networks do not overlap. Create a peering connection between them. Refer to Challenge Project 5-1, "Configure Cloud Peering," for guidance on creating a peering connection. **Take a screenshot** showing the peering connection; submit this visual with your responses to this project's questions.
4. Deploy a VM in each VPC, VNet, or subnet so that one VM has access to the Internet and the other one does not. Refer to Chapter 2, "Cloud Compute," for a review of how to deploy VM instances. Make sure one instance is deployed in your public VPC, VNet, or subnet; place the other instance in your private VPC, VNet, or subnet. Use a drawing app, such as Visio or app.diagrams.net, to **draw a network diagram** showing the VPCs/VNets/subnets and their resources, with logical connections, CIDR ranges, gateways, firewall rules, and relative position of each resource. This diagram does not need to be complex but should give another person an accurate understanding of the relationships between your deployed resources. Submit this visual with your responses to this project's questions.
5. Make sure both VMs can send and receive pings. Recall this means you will need to allow ICMP into each VPC, VNet, or subnet. Refer to the projects in Chapter 6, "Securing Cloud Resources," to review how to make these changes. For each VM, ping a location on the Internet (such as Google's DNS server at 8.8.8.8) to prove one VM has access to the Internet and the other one does not. **Take screenshots** showing the ping output for each VM; submit these visuals with your responses to this project's questions.
6. Apply an alert or alarm to one of the VMs. Refer to projects in Chapter 9, "Observability and Monitoring," for guidance on how to do this. **Take a screenshot** showing this configuration; submit this visual with your responses to this project's questions. The alarm does *not* need to be in an Alarm state; it just needs to exist and be configured to monitor the VM in some way.
7. Use your cloud platform's pricing tools to calculate the cost of your deployment for one year. Refer to Chapter 3, "Cloud System Design," for reminders on how to use a cloud platform's pricing calculator. **Take a screenshot** of your results; submit this visual with your responses to this project's questions.
8. Fully document your entire deployment. If you used all free tier resources, or if you are using an AWS Academy account, your instructor might want you to preserve your deployment for now. You can also document your deployment in a way that you could display your work in a portfolio, in a wiki, or during a job interview. You might use screenshots along with your network diagram drawings. You might also export configuration files, snapshots, or image files for one or more of your resources for later use. This portion of the project is not graded, so you can freely think through how best to collect and organize this information for your purposes. The information will be relevant for you in your job applications and in your future cloud work. One of the best ways to impress employers during a job interview is to show evidence like this of your practical skills.
9. **For group assignments**: Practice presenting your full cloud deployment to one or more group members as you would during a job interview. Assume your audience knows nothing of your deployment, and give a sufficient overview using diagrams and demonstrations so your audience can understand the resources and how they are related. Highlight the skills you acquired to design this deployment. Be prepared to answer questions. When you are part of the audience for another group member's presentation, ask at least one question to clarify anything you don't understand about their deployment. What questions were you asked during your presentation, and how did you respond to these questions? What questions do you think you should prepare for if you were to present this deployment during a job interview?
10. When you're finished with the deployment, delete all resources. Check through your account to confirm no resources are continuing to drain credits or accrue charges. Make sure to properly sign out of your account.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note 10-13 | Depending on the status of your account and the selections you made during this project, the resources you created can deplete your credits or accrue charges. Double-check to make sure you've terminated all resources you created in this project.

Solutions to Self-Check Questions

Section 10-1: Automation

1. Automation increases _____.

   Answer: c. security

   Explanation: A well-orchestrated system is much less likely to suffer from vulnerabilities due to misconfigurations and other human error, thus increasing **security**. Automation can lower costs by reducing errors, resizing quickly, and responding automatically to problems rather than relying on human intervention.

2. What IaC feature would enable you to deploy an earlier state of a web server's configuration?

   Answer: b. Versioning

   Explanation: Versioning tracks changes along the way so historical versions of a web server's configuration file can be referenced or redeployed. Immutable environments or resources aren't changed after they're deployed. Provisioning is a process that deploys new resources. A runbook is a collection of physical or digital documents that outline the precise steps to complete procedures or operations in a specific IT environment.

3. What text initiates communication with an installed CLI application?

   Answer: b. Base call

   Explanation: For most of the common cloud CLIs, the command syntax begins with a **base call** to the cloud program, such as `aws` for the AWS CLI and `az` for the Azure CLI. A subcommand identifies the operation to be performed. Parameters and filters can be added to further define information needed to execute the command.

Section 10-2: Automating Code Pipelines

4. What resource defines the phases of an automated development process?

   Answer: a. Pipeline

   Explanation: A **pipeline** is a theoretical model of system development phases that supports automation of a system from one phase to the next. Each set of changes collectively results in a new version of a project, which is called a commit. A repository is a collection point for resources that need to be accessed by users, systems, or other resources. An artifact is a single file or a collection of data and files produced by one resource in one step and consumed by another resource in another step.

5. Which process requires no human intervention so long as the pipeline does not encounter any problems?

   Answer: d. Continuous deployment

   Explanation: CD (continuous delivery) pauses the process after testing to wait for manual approval from a human and then delivers the code changes to the next environment in the pipeline. **CD (continuous deployment)** skips the manual approval step and automatically deploys the code as soon as testing is successfully completed. CI (continuous integration) refers to the frequent merging of code changes into the central repository where several tests are performed automatically.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6. Which of these items is subject to PEMDAS?

Answer: a. Operators

Explanation: With arithmetic **operators**, the PEMDAS (Please Excuse My Dear Aunt Sally) principle requires that arithmetic operations be performed in a certain order. Operators perform logical operations on certain data types. A variable stores data for later use. A function is a block of code that runs when activated.

Section 10-3: Cloud Maintenance

7. What factor primarily determines the division of responsibility for maintenance tasks?

Answer: b. Service model

Explanation: The responsibility for maintenance tasks will vary by cloud **service model**, just as the responsibility for security varies. Because maintenance activities can impact the performance of affected resources, much of this work must be completed within a maintenance window to minimize disruption to cloud services. Although some updates apply small fixes to minor bugs and others address more critical updates that patch security vulnerabilities, the responsibility for applying these updates is not determined by how critical the update is. Different OSs on VMs must be updated from different sources with different settings and steps, but this does not determine who is responsible for making the update.

8. What is the opposite of an update?

Answer: d. Rollback

Explanation: The opposite of an update is a **rollback**, which will revert the system to its state from before the update was applied. A virtual patch is a collection of rules that can temporarily provide protection from related threats. A hotfix is typically a minor update that addresses a specific issue, receives less testing, might not require restarting the system, or might be intended only for temporary use. Regular and small Microsoft updates grouped together over time in a larger update is called a build.

Section 10-4: Troubleshooting Automation Issues

9. An improperly documented server name change is a type of _____.

Answer: d. change management failure

Explanation: Improperly documented server name, IP address, or location changes might be caused by a **change management failure**. A patch failure is the result of a patch process that was not successfully implemented, such as when there is not enough available storage space. A deprecated feature is an older feature that has been replaced by a newer feature or otherwise discontinued. API version incompatibility results from the vendor discontinuing support for older APIs in favor of newer APIs.

10. What is caused by manual portions of a workflow?

Answer: b. Bottleneck

Explanation: Incompletely automated workflows apply automation for part of the process but not all of it; this causes **bottlenecks** in the workflow and continuing issues with errors. An account mismatch might be caused by updated credentials that were not adequately documented in the workflow. Job validation is required for an automation system to complete the workflow. A mis-scheduled task dependency can result in batch job failures.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Appendix A

CompTIA Cloud+ Exam Objectives Mapped to Chapters

These tables provide a complete list of the latest CompTIA Cloud+ CV0-004 certification exam objectives. The official list of objectives is available on CompTIA's website: comptia.org. For your reference, the following tables list each exam objective and the chapter and section that addresses the objective, plus the amount of the exam that will cover each certification domain.

Domain 1.0 Cloud Architecture—23% of Exam

1.1 Given a scenario, use the appropriate cloud service model.

Objective	Chapter	Section
• Cloud service models	1 2	Section 1-3: Cloud Service Models Section 2-3: VM Alternatives
– Infrastructure as a service (IaaS)	1	Section 1-3: Cloud Service Models
– Platform as a service (PaaS)	1	Section 1-3: Cloud Service Models
– Software as a service (SaaS)	1	Section 1-3: Cloud Service Models
– Function as a service (FaaS)	2	Section 2-3: VM Alternatives
• Shared responsibility model	6	Section 6-1: Cloud Security Configurations

1.2 Explain concepts related to service availability.

Objective	Chapter	Section
• Resource availability	1, 4, 5, 9	
– Region	4	Section 4-3: Networking in AWS Section 4-4: Networking in Azure Section 4-5: Networking in GCP

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Objective	Chapter	Section
– Availability zone	4	Section 4-3: Networking in AWS Section 4-4: Networking in Azure Section 4-5: Networking in GCP
– Cloud bursting	1	Section 1-2: Cloud Deployment Models
– Edge computing	5	Section 5-3: Extending the Cloud
– Availability monitoring	9	Section 9-1: Monitoring Resources
• Disaster recovery (DR)	3	Section 3-4: Planning for Problems
– Recovery time objective (RTO)	3	Section 3-4: Planning for Problems
– Recovery point objective (RPO)	3	Section 3-4: Planning for Problems
– Hot site	3	Section 3-4: Planning for Problems
– Warm site	3	Section 3-4: Planning for Problems
– Cold site	3	Section 3-4: Planning for Problems
• Multicloud tenancy	1	Section 1-2: Cloud Deployment Models

1.3 Explain cloud networking concepts.

Objective	Chapter	Section
• Public and private connections to the cloud	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking
– Virtual private network (VPN)	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking
– Dedicated connections	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking
• Network functions, components, and services	4, 5, 6	
– Application load balancer	5	Section 5-2: Extending Network Services
– Network load balancer	5	Section 5-2: Extending Network Services
– Application gateway	5	Section 5-2: Extending Network Services
– Content delivery network (CDN)	5	Section 5-3: Extending the Cloud
– Firewalls	6	Section 6-2: Virtual Network Security
– Virtual private cloud (VPC)	4, 5	
▪ Peering	4 5	Section 4-3: Networking in AWS Section 5-1: Hybrid Cloud and Multi-Cloud Networking
▪ Transit gateway	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1.3 Explain cloud networking concepts. *(Continued)*

Objective	Chapter	Section
– Subnets	4	Section 4-1: Networking Concepts in the Cloud Section 4-2: IP Address Spaces Section 4-3: Networking in AWS Section 4-4: Networking in Azure Section 4-5: Networking in GCP
– Routing and switching	4, 5	
▪ Virtual local area network (VLAN)	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking
▪ Software-defined network (SDN)	4	Section 4-1: Networking Concepts in the Cloud
▪ Border Gateway Protocol (BGP)	5	Section 5-2: Extending Network Services
▪ Static routes	4 5	Section 4-4: Networking in Azure Section 4-5: Networking in GCP Section 5-2: Extending Network Services
▪ Route tables	4 5	Section 4-3: Networking in AWS Section 4-4: Networking in Azure Section 4-5: Networking in GCP Section 5-2: Extending Network Services Section 5-4: Troubleshooting Cloud Connectivity

1.4 Compare and contrast storage resources and technologies.

Objective	Chapter	Section
• Tiered storage	8	Section 8-2: Storage Configuration
– Hot	8	Section 8-2: Storage Configuration
– Warm	8	Section 8-2: Storage Configuration
– Cold	8	Section 8-2: Storage Configuration
– Archive	8	Section 8-2: Storage Configuration
• Disk types	2 8	Section 2-2: VMs in the Cloud Section 8-2: Storage Configuration
– Solid-state drive (SSD)	2 8	Section 2-2: VMs in the Cloud Section 8-2: Storage Configuration
– Hard disk drive (HDD)	2 8	Section 2-2: VMs in the Cloud Section 8-2: Storage Configuration

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Objective	Chapter	Section
• Storage types	8	Section 8-1: Storage Types
– Object storage	8	Section 8-1: Storage Types
– Block storage	8	Section 8-1: Storage Types
– File storage	8	Section 8-1: Storage Types
• Performance implications	8	Section 8-2: Storage Configuration
• Cost implications	8	Section 8-2: Storage Configuration Section 8-4: Creating and Storing Backups

1.5 Explain the purpose of cloud-native design concepts.

Objective	Chapter	Section
• Cloud-provided managed services	1 3	Section 1-3: Cloud Service Models Section 3-2: Cloud-Native Design
• Microservices	2 3	Section 2-3: VM Alternatives Section 3-2: Cloud-Native Design
• Loosely coupled architecture	3	Section 3-2: Cloud-Native Design
• Fan-out	3	Section 3-2: Cloud-Native Design
• Service discovery	3	Section 3-2: Cloud-Native Design

1.6 Compare and contrast containerization concepts.

Objective	Chapter	Section
• Stand-alone	2	Section 2-3: VM Alternatives
• Workload orchestration	10	Section 10-1: Automation
• Networking	2	Section 2-3: VM Alternatives
• Port mapping	2	Section 2-3: VM Alternatives
• Storage types	2 8	Section 2-3: VM Alternatives Section 8-3: Cloud Storage and Database Services
• Persistent volumes	2 8	Section 2-3: VM Alternatives Section 8-3: Cloud Storage and Database Services
• Ephemeral storage	2 8	Section 2-3: VM Alternatives Section 8-3: Cloud Storage and Database Services
• Image registries	2	Section 2-3: VM Alternatives

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1.7 Compare and contrast virtualization concepts.

Objective	Chapter	Section
• Stand-alone	2	Section 2-2: VMs in the Cloud
• Clustering	2 5	Section 2-2: VMs in the Cloud Section 5-2: Extending Network Services
• Cloning	2 8	Section 2-2: VMs in the Cloud Section 8-4: Creating and Storing Backups
• Host affinity	2	Section 2-2: VMs in the Cloud
• Hardware pass-through	2	Section 2-1: Virtualization Technologies
• Network types	2, 4, 5	
– Overlay networks	4 5	Section 4-1: Networking Concepts in the Cloud Section 5-1: Hybrid Cloud and Multi-Cloud Networking
– Virtual machine (VM) networks	2	Section 2-1: Virtualization Technologies
• Storage	8	Section 8-1: Storage Types
– Local	8	Section 8-1: Storage Types
– Storage area network (SAN)	8	Section 8-1: Storage Types
– Network-attached storage (NAS)	8	Section 8-1: Storage Types

1.8 Summarize cost considerations related to cloud storage.

Objective	Chapter	Section
• Billing models	2	Section 2-2: VMs in the Cloud
– Dedicated host	2	Section 2-2: VMs in the Cloud
– Reserved resources	2	Section 2-2: VMs in the Cloud
– Pay-as-you-go	2	Section 2-2: VMs in the Cloud
– Spot instance	2	Section 2-2: VMs in the Cloud
• Resource metering	9	Section 9-1: Monitoring Resources
• Tagging	9	Section 9-1: Monitoring Resources
• Rightsizing	2 3 9	Section 2-2: VMs in the Cloud Section 3-4: Planning for Problems Section 9-4: Troubleshooting Problems and Performance

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1.9 Explain the importance of database concepts.

Objective	Chapter	Section
• Types	8	Section 8-1: Storage Types
– Relational	8	Section 8-1: Storage Types
– Non-relational	8	Section 8-1: Storage Types
• Deployment options	8	Section 8-1: Storage Types
– Self-managed	8	Section 8-1: Storage Types
– Provider-managed	8	Section 8-1: Storage Types

1.10 Compare and contrast methods for optimizing workloads using cloud resources.

Objective	Chapter	Section
• Compute resources	2	Section 2-1: Virtualization Technologies Section 2-3: VM Alternatives
– VM	2	Section 2-1: Virtualization Technologies Section 2-3: VM Alternatives
– Container	2	Section 2-3: VM Alternatives
– Serverless	2	Section 2-3: VM Alternatives
• Orchestration	10	Section 10-1: Automation
• Workflow	3 10	Section 3-2: Cloud-Native Design Section 10-1: Automation
• Network	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking Section 5-3: Extending the Cloud
– Latency	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking Section 5-3: Extending the Cloud
– Throughput	5	Section 5-1: Hybrid Cloud and Multi-Cloud Networking
• Storage	8	Section 8-2: Storage Configuration
– Input/output operations per second (IOPS)	8	Section 8-2: Storage Configuration
– Throughput	8	Section 8-2: Storage Configuration
• Managed services	1	Section 1-3: Cloud Service Models

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1.11 Identify evolving technologies in the cloud.

Objective	Chapter	Section
• Machine learning and artificial intelligence (AI)	1	Section 1-4: Cloud Services
– Text recognition	1	Section 1-4: Cloud Services
– Text translation	1	Section 1-4: Cloud Services
– Visual recognition	1	Section 1-4: Cloud Services
– Sentiment analysis	1	Section 1-4: Cloud Services
– Voice-to-text	1	Section 1-4: Cloud Services
– Text-to-voice	1	Section 1-4: Cloud Services
– Generative AI	1	Section 1-4: Cloud Services
• Internet of Things (IoT)	1 5	Section 1-4: Cloud Services Section 5-3: Extending the Cloud
– Sensors	5	Section 5-3: Extending the Cloud
– Gateways	5	Section 5-3: Extending the Cloud
– Communication	5	Section 5-3: Extending the Cloud
– Transmission protocols	5	Section 5-3: Extending the Cloud

Domain 2.0 Deployment—19% of Exam

2.1 Compare and contrast cloud deployment models.

Objective	Chapter	Section
• Public	1	Section 1-2: Cloud Deployment Models
• Private	1	Section 1-2: Cloud Deployment Models
– On premises	1	Section 1-2: Cloud Deployment Models
• Hybrid	1	Section 1-2: Cloud Deployment Models
• Community	1	Section 1-2: Cloud Deployment Models

2.2 Given a scenario, implement appropriate deployment strategies.

Objective	Chapter	Section
• Blue-green	3	Section 3-3: System Development
• Canary	3	Section 3-3: System Development
• Rolling	3	Section 3-3: System Development
• In-place	3	Section 3-3: System Development

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.3 Summarize aspects of cloud migration.

Objective	Chapter	Section
• Migration types	3	Section 3-1: Cloud Migration
– On-premises-to-cloud	3	Section 3-1: Cloud Migration
– Cloud-to-on-premises	3	Section 3-1: Cloud Migration
– Cloud-to-cloud	3	Section 3-1: Cloud Migration
• Resource allocation	2	Section 2-2: VMs in the Cloud
• Considerations	2, 3, 4, 6, 8, 9	
– Storage	8	Section 8-2: Storage Configuration
– Platform compatibility	3	Section 3-1: Cloud Migration
– Compute	2 3	Section 2-2: VMs in the Cloud Section 3-1: Cloud Migration
– Cost	3	Section 3-1: Cloud Migration
– Networking	4	Section 4-2: IP Address Spaces
– Management overhead	3	Section 3-1: Cloud Migration
– Service availability	9	Section 9-1: Monitoring Resources
– Vendor lock-in	3	Section 3-1: Cloud Migration
– Environmental	3	Section 3-1: Cloud Migration
▪ Power and cooling	3	Section 3-1: Cloud Migration
– Regulatory	3 6 8	Section 3-1: Cloud Migration Section 6-1: Cloud Security Configurations Section 8-2: Storage Configuration
– Compliance	3 6 8	Section 3-1: Cloud Migration Section 6-1: Cloud Security Configurations Section 8-2: Storage Configuration
• Application migration strategies	3	Section 3-1: Cloud Migration
– Rehost	3	Section 3-1: Cloud Migration
– Replatform	3	Section 3-1: Cloud Migration
– Re-architect	3	Section 3-1: Cloud Migration
– Retain	3	Section 3-1: Cloud Migration
– Retire	3	Section 3-1: Cloud Migration
– Refactor	3	Section 3-1: Cloud Migration

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.4 Given a scenario, use code to deploy and configure cloud resources.

Objective	Chapter	Section
• Infrastructure as code (IaC)	10	Section 10-1: Automation
• Configuration as code (CaC)	10	Section 10-1: Automation
• Scripting logic	10	Section 10-2: Automating Code Pipelines
– Variables	10	Section 10-2: Automating Code Pipelines
– Conditionals	10	Section 10-2: Automating Code Pipelines
– Operators	10	Section 10-2: Automating Code Pipelines
– Data types	10	Section 10-2: Automating Code Pipelines
– Functions	10	Section 10-2: Automating Code Pipelines
• Repeatability	10	Section 10-1: Automation
• Drift detection	10	Section 10-1: Automation
• Versioning	10	Section 10-1: Automation
• Testing	3 10	Section 3-3: System Development Section 10-1: Automation Section 10-2: Automating Code Pipelines
• Documentation	10	Section 10-1: Automation
– Formats	10	Section 10-1: Automation Section 10-2: Automating Code Pipelines
– JavaScript Object Notation (JSON)	10	Section 10-1: Automation Section 10-2: Automating Code Pipelines
– Yet Another Markup Language (YAML)	10	Section 10-1: Automation Section 10-2: Automating Code Pipelines

2.5 Given a set for requirements, provision the appropriate cloud resources.

Objective	Chapter	Section
• Storage requirements	2 3 8	Section 2-1: Virtualization Technologies Section 2-2: VMs in the Cloud Section 3-1: Cloud Migration Section 8-2: Storage Configuration Section 8-3: Cloud Storage and Database Services
• Performance requirements	3 9	Section 3-1: Cloud Migration Section 9-3: Cloud Monitoring Services

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Objective	Chapter	Section
• Security requirements	6 7	Section 6-1: Cloud Security Configurations Section 6-2: Virtual Network Security Section 6-3: Compute Security Section 6-4: Data Security Section 7-1: Cloud Accounts
• Cost requirements	1 3 9	Section 1-1: Characteristics of Cloud Computing Section 3-1: Cloud Migration Section 9-3: Cloud Monitoring Services
• Availability requirements	4 5 8	Section 4-3: Networking in AWS Section 4-4: Networking in Azure Section 4-5: Networking in GCP Section 5-1: Hybrid Cloud and Multi-Cloud Networking Section 5-4: Troubleshooting Cloud Connectivity Section 8-2: Storage Configuration Section 8-3: Cloud Storage and Database Services
• Compliance requirements	3 6 8	Section 3-1: Cloud Migration Section 6-1: Cloud Security Configurations Section 6-5: Troubleshooting Cloud Security Section 8-2: Storage Configuration Section 8-4: Creating and Storing Backups Section 8-5: Storage Security
• Network requirements	4 5	Section 4-1: Networking Concepts in the Cloud Section 4-3: Networking in AWS Section 4-4: Networking in Azure Section 4-5: Networking in GCP Section 5-2: Extending Network Services Section 5-4: Troubleshooting Cloud Connectivity
• Compute requirements	2	Section 2-1: Virtualization Technologies Section 2-2: VMs in the Cloud

Domain 3.0 Operations—17% of Exam

3.1 Given a scenario, configure appropriate resources to achieve observability.

Objective	Chapter	Section
• Logging	9	Section 9-2: Events, Logs, and Traces
– Collection	9	Section 9-2: Events, Logs, and Traces

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3.1 Given a scenario, configure appropriate resources to achieve observability. *(Continued)*

Objective	Chapter	Section
– Aggregation	9	Section 9-2: Events, Logs, and Traces
– Retention	9	Section 9-2: Events, Logs, and Traces
• Tracing	9	Section 9-2: Events, Logs, and Traces
• Monitoring	9	Section 9-1: Monitoring Resources Section 9-3: Cloud Monitoring Services
– Metrics	9	Section 9-1: Monitoring Resources Section 9-3: Cloud Monitoring Services
• Alerting	9	Section 9-2: Events, Logs, and Traces Section 9-3: Cloud Monitoring Services
– Triage	9	Section 9-2: Events, Logs, and Traces
– Response	9	Section 9-3: Cloud Monitoring Services

3.2 Given a scenario, configure appropriate scaling approaches.

Objective	Chapter	Section
• Approaches	3 9	Section 3-4: Planning for Problems Section 9-4: Troubleshooting Problems and Performance
– Triggered	3 9	Section 3-4: Planning for Problems Section 9-4: Troubleshooting Problems and Performance
▪ Trending	3	Section 3-4: Planning for Problems
▪ Load	3	Section 3-4: Planning for Problems
▪ Event	3 9	Section 3-4: Planning for Problems Section 9-4: Troubleshooting Problems and Performance
– Scheduled	3	Section 3-4: Planning for Problems
– Manual	3	Section 3-4: Planning for Problems
• Types	3	Section 3-4: Planning for Problems
– Horizontal	3	Section 3-4: Planning for Problems
– Vertical	3	Section 3-4: Planning for Problems

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3.3 Given a scenario, use appropriate backup and recovery methods.

Objective	Chapter	Section
• Backup types	8	Section 8-4: Creating and Storing Backups
– Incremental	8	Section 8-4: Creating and Storing Backups
– Full	8	Section 8-4: Creating and Storing Backups
– Differential	8	Section 8-4: Creating and Storing Backups
• Backup locations	8	Section 8-4: Creating and Storing Backups
– On site	8	Section 8-4: Creating and Storing Backups
– Off site	8	Section 8-4: Creating and Storing Backups
• Schedule	8	Section 8-4: Creating and Storing Backups
• Retention	8	Section 8-4: Creating and Storing Backups
• Replication	8	Section 8-4: Creating and Storing Backups
• Encryption	8	Section 8-4: Creating and Storing Backups
• Testing	8	Section 8-4: Creating and Storing Backups
– Recoverability	8	Section 8-4: Creating and Storing Backups
– Integrity	8	Section 8-4: Creating and Storing Backups
• Recovery types	9	Section 9-4: Troubleshooting Problems and Performance
– In-place	9	Section 9-4: Troubleshooting Problems and Performance
– Parallel	9	Section 9-4: Troubleshooting Problems and Performance
• Recovery options	8	Section 8-4: Creating and Storing Backups
– Bulk	8	Section 8-4: Creating and Storing Backups
– Granular	8	Section 8-4: Creating and Storing Backups

3.4 Given a scenario, manage the life cycle of cloud resources.

Objective	Chapter	Section
• Patches	10	Section 10-3: Cloud Maintenance
• Updates	10	Section 10-3: Cloud Maintenance
– Major	10	Section 10-3: Cloud Maintenance
– Minor	10	Section 10-3: Cloud Maintenance
• Testing	10	Section 10-3: Cloud Maintenance

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3.4 Given a scenario, manage the life cycle of cloud resources. *(Continued)*

Objective	Chapter	Section
• Data	2 8	Section 2-3: VM Alternatives Section 8-3: Cloud Storage and Database Services
– Ephemeral	2 8	Section 2-3: VM Alternatives Section 8-3: Cloud Storage and Database Services
– Persistent	2 8	Section 2-3: VM Alternatives Section 8-3: Cloud Storage and Database Services
• Decommissioning	10	Section 10-3: Cloud Maintenance
– End of life	10	Section 10-3: Cloud Maintenance
– End of support	10	Section 10-3: Cloud Maintenance

Domain 4.0 Security—19% of Exam

4.1 Explain vulnerability management concepts.

Objective	Chapter	Section
Steps	6	Section 6-1: Cloud Security Configurations
– Scanning scope	6	Section 6-1: Cloud Security Configurations
– Identification	6	Section 6-1: Cloud Security Configurations
– Assessment	6	Section 6-1: Cloud Security Configurations
– Remediation	6	Section 6-1: Cloud Security Configurations
• Common Vulnerabilities and Exposures (CVEs)	6	Section 6-1: Cloud Security Configurations

4.2 Compare and contrast aspects of compliance and regulation.

Objective	Chapter	Section
• Data sovereignty	8	Section 8-5: Storage Security
• Data ownership	8	Section 8-5: Storage Security
• Data locality	8	Section 8-5: Storage Security
• Data classification	7 8	Section 7-3: Authorization Section 8-5: Storage Security
• Data retention	8	Section 8-2: Storage Configuration Section 8-4: Creating and Storing Backups Section 8-5: Storage Security
– Litigation hold	8	Section 8-4: Creating and Storing Backups

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Objective	Chapter	Section
– Contractual	8	Section 8-5: Storage Security
– Regulatory	8	Section 8-5: Storage Security
• Industry standards	1 3	Section 1-1: Characteristics of Cloud Computing Section 3-1: Cloud Migration
– Systems and Organization Controls 2 (SOC2)	3	Section 3-1: Cloud Migration
– Payment Card Industry Data Security Standards (PCI DSS)	3	Section 3-1: Cloud Migration
– International Organization for Standardization (ISO) 27001	3	Section 3-1: Cloud Migration
– Cloud Security Alliance	1 3	Section 1-1: Characteristics of Cloud Computing Section 3-1: Cloud Migration

4.3 Given a scenario, implement identity and access management.

Objective	Chapter	Section
• Secure access to the cloud management environment	1 7	Section 1-1: Characteristics of Cloud Computing Section 7-2: Authentication
– Programmatic access	7	Section 7-2: Authentication
▪ Application programming interface (API)	7	Section 7-2: Authentication
▪ Software development kit (SDK)	7	Section 7-2: Authentication
– Common Language Infrastructure (CLI)	7	Section 7-2: Authentication
– Web portal	1 7	Section 1-1: Characteristics of Cloud Computing Section 7-2: Authentication
• Secure access to the cloud resources	3, 4, 6, 7	
– API	6 7	Section 6-4: Data Security Section 7-2: Authentication
– Secure Shell (SSH)	3 6	Section 3-4: Planning for Problems Section 6-3: Compute Security
– Remote Desktop Protocol (RDP)	3 6	Section 3-4: Planning for Problems Section 6-3: Compute Security
– Bastion host	4	Section 4-3: Networking in AWS
• Authentication models	7	Section 7-2: Authentication Section 7-3: Authorization Section 7-4: IAM for Hybrid Clouds
– Local users	7	Section 7-2: Authentication

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4.3 Given a scenario, implement identity and access management. *(Continued)*

Objective	Chapter	Section
– Federation	7	Section 7-4: IAM for Hybrid Clouds
▪ Security Assertion Markup Language (SAML)	7	Section 7-4: IAM for Hybrid Clouds
– Token-based	7	Section 7-3: Authorization
– Directory-based	7	Section 7-2: Authentication Section 7-4: IAM for Hybrid Clouds
– Multifactor authentication (MFA)	7	Section 7-2: Authentication
– OpenID Connect	7	Section 7-4: IAM for Hybrid Clouds
• Authorization models	7	Section 7-3: Authorization Section 7-4: IAM for Hybrid Clouds
– Role-based access control	7	Section 7-3: Authorization
– Group-based access control	7	Section 7-3: Authorization
– OAuth 2.0	7	Section 7-3: Authorization Section 7-4: IAM for Hybrid Clouds
– Discretionary	7	Section 7-3: Authorization
• Accounting	7 9	Section 7-2: Authentication Section 9-2: Events, Logs, and Traces
– Audit trail	9	Section 9-2: Events, Logs, and Traces

4.4 Given a scenario, apply security best practices.

Objective	Chapter	Section
• Zero Trust	6	Section 6-2: Virtual Network Security
• Benchmark	6	Section 6-1: Cloud Security Configurations
– Center for Internet Security (CIS)	6	Section 6-1: Cloud Security Configurations
– Vendor-specific	6	Section 6-1: Cloud Security Configurations
• Hardening	6	Section 6-3: Compute Security
• Patching	6	Section 6-3: Compute Security
• Encryption	6 8	Section 6-4: Data Security Section 8-5: Storage Security
– Data in transit	6	Section 6-4: Data Security
– Data at rest	6 8	Section 6-4: Data Security Section 8-5: Storage Security
• Secrets management	7	Section 7-2: Authentication
• API security	6	Section 6-4: Data Security

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Objective	Chapter	Section
• Principle of least privilege	7	Section 7-1: Cloud Accounts Section 7-5: Troubleshooting Cloud IAM
• Container security	6	Section 6-3: Compute Security
– Privileged	6	Section 6-3: Compute Security
– Unprivileged	6	Section 6-3: Compute Security
– File access permissions	6	Section 6-3: Compute Security
• Storage security	6 8	Section 6-4: Data Security Section 8-5: Storage Security
– Object storage	8	Section 8-5: Storage Security
– File storage	8	Section 8-5: Storage Security

4.5 Given a scenario, apply security controls in the cloud.

Objective	Chapter	Section
• Endpoint protection	6	Section 6-3: Compute Security
• Data loss prevention (DLP)	6 8	Section 6-2: Virtual Network Security Section 8-5: Storage Security
• Intrusion prevention system/intrusion detection system (IPS/IDS)	6	Section 6-2: Virtual Network Security
• Distributed denial-of-service (DDoS) protection	6	Section 6-2: Virtual Network Security
• Identity and access management (IAM) policies	7	Section 7-3: Authorization
• Firewall	6	Section 6-2: Virtual Network Security Section 6-3: Compute Security
– Network access control list (ACL)	6	Section 6-2: Virtual Network Security
– Web application firewall (WAF)	6	Section 6-3: Compute Security
– Network security group	6	Section 6-2: Virtual Network Security

4.6 Given a scenario, monitor suspicious activities to identify common attacks.

Objective	Chapter	Section
• Event monitoring	9	Section 9-2: Events, Logs, and Traces Section 9-3: Cloud Monitoring Services
• Deviation from the baseline	9	Section 9-3: Cloud Monitoring Services
• Unnecessary open ports	6	Section 6-2: Virtual Network Security

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4.6 Given a scenario, monitor suspicious activities to identify common attacks. ***(Continued)***

Objective	Chapter	Section
• Attack types	6	Section 6-1: Cloud Security Configurations
– Vulnerability exploitation	6	Section 6-1: Cloud Security Configurations
▪ Human error	6	Section 6-1: Cloud Security Configurations
▪ Outdated software	6	Section 6-1: Cloud Security Configurations
– Social engineering	6	Section 6-1: Cloud Security Configurations
▪ Phishing	6	Section 6-1: Cloud Security Configurations
– Malware	6	Section 6-1: Cloud Security Configurations
▪ Ransomware	6	Section 6-1: Cloud Security Configurations
– DDoS	6	Section 6-1: Cloud Security Configurations
– Cryptojacking	6	Section 6-1: Cloud Security Configurations
– Zombie instances	6	Section 6-1: Cloud Security Configurations
– Metadata	6	Section 6-1: Cloud Security Configurations

Domain 5.0 DevOps Fundamentals—10% of Exam

5.1 Explain source control concepts.

Objective	Chapter	Section
• Version management	3 10	Section 3-3: System Development Section 10-1: Automation
• Code review	3	Section 3-3: System Development
• Pull request	3	Section 3-3: System Development
• Code push	3	Section 3-3: System Development
• Code commit	3	Section 3-3: System Development
• Code merge	3	Section 3-3: System Development
• Branch management	3	Section 3-3: System Development

5.2 Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.

Objective	Chapter	Section
• Automation	9 10	Section 9-4: Troubleshooting Problems and Performance Section 10-1: Automation Section 10-2: Automating Code Pipelines
• Code integration	10	Section 10-2: Automating Code Pipelines

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Objective	Chapter	Section
• Code deployment	10	Section 10-2: Automating Code Pipelines
– Build	10	Section 10-2: Automating Code Pipelines
• Testing	3 10	Section 3-3: System Development Section 10-1: Automation Section 10-2: Automating Code Pipelines
• Security	6	Section 6-3: Compute Security
• Workflow	10	Section 10-1: Automation
• Artifacts	10	Section 10-2: Automating Code Pipelines
– Images	10	Section 10-2: Automating Code Pipelines
▪ VM	10	Section 10-2: Automating Code Pipelines
▪ Container	10	Section 10-2: Automating Code Pipelines
– Packages	10	Section 10-2: Automating Code Pipelines
▪ Red Hat Package Manager (RPM)	10	Section 10-2: Automating Code Pipelines
▪ Debian	10	Section 10-2: Automating Code Pipelines
▪ ZIP	10	Section 10-2: Automating Code Pipelines
▪ tar	10	Section 10-2: Automating Code Pipelines
– Flat file	10	Section 10-2: Automating Code Pipelines
• Repositories	2 10	Section 2-3: VM Alternatives Section 10-2: Automating Code Pipelines
– Public	2 10	Section 2-3: VM Alternatives Section 10-2: Automating Code Pipelines
– Private	2 10	Section 2-3: VM Alternatives Section 10-2: Automating Code Pipelines

5.3 Explain concepts related to integration of systems.

Objective	Chapter	Section
• Event-driven architectures	3	Section 3-2: Cloud-Native Design
• Web services	3	Section 3-2: Cloud-Native Design
– Representational state transfer (REST)	3	Section 3-2: Cloud-Native Design
– Simple Object Access Protocol (SOAP)	3	Section 3-2: Cloud-Native Design
– Remote procedure call (RPC)	3	Section 3-2: Cloud-Native Design
• Web sockets	3	Section 3-2: Cloud-Native Design
• GraphQL	3	Section 3-2: Cloud-Native Design

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

5.4 Explain the importance of tools used in DevOps environments.

Objective	Chapter	Section
• Ansible	10	Section 10-1: Automation
• Docker	2	Section 2-3: VM Alternatives
• Elasticsearch, Logstash, and Kibana (ELK) stack	9	Section 9-2: Events, Logs, and Traces
• Git	3	Section 3-3: System Development
• GitHub actions	3	Section 3-3: System Development
• Grafana	9	Section 9-2: Events, Logs, and Traces
• Jenkins	3	Section 3-3: System Development
• Kubernetes	2	Section 2-3: VM Alternatives
• Terraform	10	Section 10-1: Automation

Domain 6.0 Troubleshooting—12% of Exam

6.1 Given a scenario, troubleshoot deployment issues.

Objective	Chapter	Section
• Incompatibility	10	Section 10-4: Troubleshooting Automation Issues
• Misconfigurations	2, 3, 7, 9	
– Resource allocation	2	Section 2-2: VMs in the Cloud
– Permission issues	7	Section 7-5: Troubleshooting Cloud IAM
– Oversubscription	9	Section 9-4: Troubleshooting Problems and Performance
– Sizing issues	2 3 9	Section 2-2: VMs in the Cloud Section 3-4: Planning for Problems Section 9-4: Troubleshooting Problems and Performance
• Outdated component definitions	10	Section 10-4: Troubleshooting Automation Issues
• Deprecation of functionality	10	Section 10-4: Troubleshooting Automation Issues
• Outages	3, 4, 5	
– Full	3	Section 3-4: Planning for Problems
– Partial	4 5	Section 4-3: Networking in AWS Section 5-4: Troubleshooting Cloud Connectivity

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Objective	Chapter	Section
• Resource limits	3, 9	
– API throttling	9	Section 9-4: Troubleshooting Problems and Performance
– Service quotas	3	Section 3-4: Planning for Problems
• Regional service availability	4 5	Section 4-3: Networking in AWS Section 5-4: Troubleshooting Cloud Connectivity

6.2 Given a scenario, troubleshoot network issues.

Objective	Chapter	Section
Network service unavailability	4, 5	
– Dynamic Host Configuration Protocol (DHCP)	5	Section 5-2: Extending Network Services
– Domain Name System (DNS)	5	Section 5-2: Extending Network Services
– Network Time Protocol (NTP)	5	Section 5-4: Troubleshooting Cloud Connectivity
– Network Address Translation (NAT)	4 5	Section 4-3: Networking in AWS Section 5-4: Troubleshooting Cloud Connectivity
– Hypertext Transfer Protocol (HTTP)	5	Section 5-4: Troubleshooting Cloud Connectivity
▪ Status codes	5	Section 5-4: Troubleshooting Cloud Connectivity
• Latency	1 5	Section 1-5: Troubleshooting Methodology Section 5-4: Troubleshooting Cloud Connectivity
• Bandwidth/throughput issues	5	Section 5-4: Troubleshooting Cloud Connectivity
• Network device misconfiguration	5	Section 5-4: Troubleshooting Cloud Connectivity
• Protocol incompatibility	5	Section 5-4: Troubleshooting Cloud Connectivity
• Protocol deprecations	6	Section 6-4: Data Security
• IP addressing issues	4	Section 4-2: IP Address Spaces
– Scope exhaustion	4	Section 4-2: IP Address Spaces
– Network overlap	4	Section 4-2: IP Address Spaces
• Routing issues	5	Section 5-4: Troubleshooting Cloud Connectivity
– Missing routes	5	Section 5-4: Troubleshooting Cloud Connectivity
– Misconfigured routes	5	Section 5-4: Troubleshooting Cloud Connectivity
• Switching issues	5	Section 5-4: Troubleshooting Cloud Connectivity
– VLAN issues	5	Section 5-4: Troubleshooting Cloud Connectivity
▪ Misconfigured tags	5	Section 5-4: Troubleshooting Cloud Connectivity
– Access vs. truck ports	5	Section 5-4: Troubleshooting Cloud Connectivity

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.3 Given a scenario, troubleshoot security issues.

Objective	Chapter	Section
• Cipher suite deprecations	6	Section 6-4: Data Security Section 6-5: Troubleshooting Cloud Security
• Authorization issues	7	Section 7-5: Troubleshooting Cloud IAM
– Privilege escalation	7	Section 7-5: Troubleshooting Cloud IAM
– Unauthorized access	7	Section 7-5: Troubleshooting Cloud IAM
• Authentication issues	6 7	Section 6-5: Troubleshooting Cloud Security Section 7-5: Troubleshooting Cloud IAM
– Leaked credentials	6	Section 6-5: Troubleshooting Cloud Security
• Software vulnerability issues	6	Section 6-1: Cloud Security Configurations Section 6-3: Compute Security
• Unauthorized software	6	Section 6-3: Compute Security

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Appendix B

Chapters Mapped to CompTIA Cloud+ Exam Objectives

These tables provide a complete list of the latest CompTIA Cloud+ CV0-004 certification exam objectives mapped to each section within each chapter. The official list of objectives is available on CompTIA's website: comptia.org. For your reference, the following tables list each line item from the objectives covered in each chapter and section.

Chapter 1: Introduction to Cloud Computing

Sections and headings	Objectives
Section 1-1: Characteristics of Cloud Computing • CompTIA Cloud+ Certification • Other Cloud Certifications • What Is Cloud Computing? • What Do I Need to Know?	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Cost requirements Obj. 4.2: Compare and contrast aspects of compliance and regulation. • Industry standards – Cloud Security Alliance Obj. 4.3: Given a scenario, implement identity and access management. • Secure access to the cloud management environment – Web portal
Section 1-2: Cloud Deployment Models • Public Cloud • Private Cloud • Hybrid Cloud • Multi-Cloud • Community Cloud	Obj. 1.2: Explain concepts related to service availability. • Resource availability – Cloud bursting • Multicloud tenancy Obj. 2.1: Compare and contrast cloud deployment models. • Public • Private – On premises • Hybrid • Community

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 1 Introduction to Cloud Computing *(Continued)*

Sections and headings	Objectives
Section 1-3: Cloud Service Models • Common Cloud Service Models • Managed Services	Obj. 1.1: Given a scenario, use the appropriate cloud service model. • Cloud service models – Infrastructure as a service (IaaS) – Platform as a service (PaaS) – Software as a service (SaaS) Obj. 1.5: Explain the purpose of cloud-native design concepts. • Cloud-provided managed services Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources. • Managed services
Section 1-4: Cloud Services • PaaS and IaaS Providers • Common Cloud Services • AI/ML (Artificial Intelligence/ Machine Learning) Services • IoT (Internet of Things)	Obj. 1.11: Identify evolving technologies in the cloud. • Machine learning and artificial intelligence (AI) – Text recognition – Text translation – Visual recognition – Sentiment analysis – Voice-to-text – Text-to-voice – Generative AI • Internet of Things (IoT)
Section 1-5: Troubleshooting Methodology • Common Cloud Computing Problems • Troubleshooting Steps • Preventive Measures	Obj. 6.2: Given a scenario, troubleshoot network issues. • Latency

Chapter 2: Cloud Compute

Sections and headings	Objectives
Section 2-1: Virtualization Technologies • VMs (Virtual Machines) • VM Configuration • Network Connection Types	Obj. 1.7: Compare and contrast virtualization concepts. • Hardware pass-through • Network types – Virtual machine (VM) networks Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources. • Compute resources – VM

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Storage requirements • Compute requirements
Section 2-2: VMs in the Cloud • VM Instance Types • Cloning VM Instances • Billing Models • Resource Allocation Factors • Affinity	**Obj. 1.4: Compare and contrast storage resources and technologies.** • Disk types – Solid-state drive (SSD) – Hard disk drive (HDD) **Obj. 1.7: Compare and contrast virtualization concepts.** • Stand-alone • Clustering • Cloning • Host affinity **Obj. 1.8: Summarize cost considerations related to cloud usage.** • Billing models – Dedicated host – Reserved resources – Pay-as-you-go – Spot instance • Rightsizing **Obj. 2.3: Summarize aspects of cloud migration.** • Resource allocation • Considerations – Compute **Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Storage requirements • Compute requirements **Obj. 6.1: Given a scenario, troubleshoot deployment issues.** • Misconfigurations – Resource allocation – Sizing issues
Section 2-3: VM Alternatives • Serverless Computing • Containers	**Obj. 1.1: Given a scenario, use the appropriate cloud service model.** • Cloud service models – Function as a service (FaaS) **Obj. 1.5: Explain the purpose of cloud-native design concepts.** • Microservices

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 2 Cloud Compute *(Continued)*

Sections and headings	Objectives
	Obj. 1.6: Compare and contrast containerization concepts. • Stand-alone • Networking – Port mapping • Storage types – Persistent volumes – Ephemeral storage • Image registries Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources. • Compute resources – VM – Container – Serverless Obj. 3.4: Given a scenario, manage the life cycle of cloud resources. • Data – Ephemeral – Persistent Obj. 5.2: Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines. • Repositories – Public – Private Obj. 5.4: Explain the importance of tools used in DevOps environments. • Docker • Kubernetes

Chapter 3: Cloud System Design

Sections and headings	Objectives
Section 3-1: Cloud Migration • Migration Types • Migration Strategies • Budgeting for Cloud • Regulations and Compliance • Project Management	Obj. 2.3: Summarize aspects of cloud migration. • Migration types – On-premises-to-cloud – Cloud-to-on-premises – Cloud-to-cloud

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
	• Considerations – Platform compatibility – Compute – Cost – Management overhead – Vendor lock-in – Environmental ▪ Power and cooling – Regulatory – Compliance • Application migration strategies – Rehost – Replatform – Re-architect – Retain – Retire – Refactor **Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Storage requirements • Performance requirements • Cost requirements • Compliance requirements **Obj. 4.2: Compare and contrast aspects of compliance and regulation.** • Industry standards – Systems and Organization Controls 2 (SOC 2) – Payment Card Industry Data Security Standards (PCI DSS) – International Organization for Standardization (ISO) 27001 – Cloud Security Alliance
Section 3-2: Cloud-Native Design • Building with Microservices • System Integration	**Obj. 1.5: Explain the purpose of cloud-native design concepts.** • Cloud-provided managed services • Microservices • Loosely coupled architectures • Fan-out • Service discovery **Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources.** • Workflow

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 3 Cloud System Design *(Continued)*

Sections and headings	Objectives
	Obj. 5.3: Explain concepts related to integration of systems. • Event-driven architectures • Web services – Representational state transfer (REST) – Simple Object Access Protocol (SOAP) – Remote procedure call (RPC) • Web sockets • GraphQL
Section 3-3: System Development • DevOps • Version Control • Deployment Strategies • Testing Types	Obj. 2.2: Given a scenario, implement appropriate deployment strategies. • Blue-green • Canary • Rolling • In-place Obj. 2.4: Given a scenario, use code to deploy and configure cloud resources. • Testing Obj. 5.1: Explain source control concepts. • Version management • Code review • Pull request • Code push • Code commit • Code merge • Branch management Obj. 5.2: Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines. • Testing Obj. 5.4: Explain the importance of tools used in DevOps environments. • Git • GitHub actions • Jenkins
Section 3-4: Planning for Problems • Common Deployment Issues • Sizing Issues • Capacity Planning • Auto Scaling Types • Business Continuity Planning • Disaster Recovery	Obj. 1.2: Explain concepts related to service availability. • Disaster recovery (DR) – Recovery time objective (RTO) – Recovery point objective (RPO) – Hot site – Warm site – Cold site Obj. 1.8: Summarize cost considerations related to cloud usage. • Rightsizing

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
	Obj. 3.2: Given a scenario, configure appropriate scaling approaches. • Approaches – Triggered ▪ Trending ▪ Load ▪ Event – Scheduled – Manual • Types – Horizontal – Vertical Obj. 4.3: Given a scenario, implement identity and access management. • Secure access to the cloud resources – Secure Shell (SSH) – Remote Desktop Protocol (RDP) Obj. 6.1: Given a scenario, troubleshoot deployment issues. • Misconfigurations – Sizing issues • Outages – Full • Resource limits – Service quotas

Chapter 4: Cloud Networking

Sections and headings	Objectives
Section 4-1: Networking Concepts in the Cloud • Networking Concepts • From OSI Model to Cloud Stack • Software-Defined Networking in the Cloud	Obj. 1.3: Explain cloud networking concepts. • Network functions, components, and services – Virtual private cloud (VPC) – Subnets – Routing and switching ▪ Software-defined network (SDN) Obj. 1.7: Compare and contrast virtualization concepts. • Network types – Overlay networks Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Network requirements

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 4 Cloud Networking *(Continued)*

Sections and headings	Objectives
Section 4-2: IP Address Spaces • IP Addressing • Subnetting • Cloud Network Interfaces	Obj. 1.3: Explain cloud networking concepts. • Network functions, components, and services – Subnets Obj. 2.3: Summarize aspects of cloud migration. • Considerations – Networking Obj. 6.2: Given a scenario, troubleshoot network issues. • IP addressing issues – Scope exhaustion – Network overlap
Section 4-3: Networking in AWS • Regions in AWS • Availability Zones in AWS • VPCs in AWS • Subnets in AWS • Gateways and Route Tables in AWS	Obj. 1.2: Explain concepts related to service availability. • Resource availability – Region – Availability zone Obj. 1.3: Explain cloud networking concepts. • Network functions, components, and services – Virtual private cloud (VPC) ▪ Peering – Subnets – Routing and switching ▪ Route tables Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Availability requirements • Network requirements Obj. 4.3: Given a scenario, implement identity and access management. • Secure access to the cloud resources – Bastion host Obj. 6.1: Given a scenario, troubleshoot deployment issues. • Outages – Partial • Regional service availability Obj. 6.2: Given a scenario, troubleshoot network issues. • Network service unavailability – Network Address Translation (NAT)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
Section 4-4: Networking in Azure • Regions and Availability Zones in Azure • VNets and Subnets in Azure • Route Tables in Azure	Obj. 1.2: Explain concepts related to service availability. • Resource availability – Region – Availability zone Obj. 1.3: Explain cloud networking concepts. • Network functions, components, and services – Virtual private cloud (VPC) – Subnets – Routing and switching ▪ Static routes ▪ Route tables Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Availability requirements • Network requirements
Section 4-5: Networking in GCP • Regions and Zones in GCP • VPCs and Subnets in GCP • Routes in GCP	Obj. 1.2: Explain concepts related to service availability. • Resource availability – Region – Availability zone Obj. 1.3: Explain cloud networking concepts. • Network functions, components, and services – Virtual private cloud (VPC) – Subnets – Routing and switching ▪ Static routes ▪ Route tables Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Availability requirements • Network requirements

Chapter 5: Cloud Connectivity

Sections and headings	Objectives
Section 5-1: Hybrid Cloud and Multi-Cloud Networking • Connecting Networks • Connecting Cloud Networks	Obj. 1.3: Explain cloud networking concepts. • Public and private connections to the cloud – Virtual private network (VPN) – Dedicated connections

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 5 Cloud Connectivity *(Continued)*

Sections and headings	Objectives
	• Network functions, components, and services – Virtual private cloud (VPC) ▪ Peering ▪ Transit gateway – Routing and switching ▪ Virtual local area network (VLAN) **Obj. 1.7: Compare and contrast virtualization concepts.** • Network types – Overlay networks **Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources.** • Network – Latency – Throughput **Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Availability requirements
Section 5-2: Extending Network Services • DHCP Services • DNS Services • Routing • Load Balancing	**Obj. 1.3: Explain cloud networking concepts.** • Network functions, components, and services – Application load balancer – Network load balancer – Application gateway – Routing and switching ▪ Border Gateway Protocol (BGP) ▪ Static routes ▪ Route tables **Obj. 1.7: Compare and contrast virtualization concepts.** • Clustering **Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Network requirements **Obj. 6.2: Given a scenario, troubleshoot network issues.** • Network service unavailability – Dynamic Host Configuration Protocol (DHCP) – Domain Name System (DNS)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
Section 5-3: Extending the Cloud • The Case for Edge Computing • CDN (Content Delivery Networks) • IoT Cloud Networking	**Obj. 1.2: Explain concepts related to service availability.** • Resource availability – Edge computing **Obj. 1.3: Explain cloud networking concepts.** • Network functions, components, and services – Content delivery network (CDN) **Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources.** • Network – Latency **Obj. 1.11: Identify evolving technologies in the cloud.** • Internet of Things (IoT) – Sensors – Gateways – Communication – Transmission protocols
Section 5-4: Troubleshooting Cloud Connectivity • Common Connectivity Problems • HTTP Status Codes • Misconfigured Routes • NTP (Network Time Protocol) • VLAN Issues	**Obj. 1.3: Explain cloud networking concepts.** • Network functions, components, and services – Routing and switching ▪ Route tables **Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Availability requirements • Network requirements **Obj. 6.1: Given a scenario, troubleshoot deployment issues.** • Outages – Partial • Regional service availability **Obj. 6.2: Given a scenario, troubleshoot network issues.** • Network service unavailability – Network Time Protocol (NTP) – Network Address Translation (NAT) – Hypertext Transfer Protocol (HTTP) ▪ Status codes • Latency • Bandwidth/throughput issues • Network device misconfiguration • Protocol incompatibility

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 5 Cloud Connectivity *(Continued)*

Sections and headings	Objectives
	• Routing issues – Missing routes – Misconfigured routes • Switching issues – VLAN issues ▪ Misconfigured tags – Access vs. trunk ports

Chapter 6: Securing Cloud Resources

Sections and headings	Objectives
Section 6-1: Cloud Security Configurations • Threats to Cloud Security • Vulnerability Management • Shared Responsibility for Cloud Security	**Obj. 1.1: Given a scenario, use the appropriate cloud service model.** • Shared responsibility model **Obj. 2.3: Summarize aspects of cloud migration.** • Considerations – Regulatory – Compliance **Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Security requirements • Compliance requirements **Obj. 4.1: Explain vulnerability management concepts.** • Steps – Scanning scopes – Identification – Assessment – Remediation • Common Vulnerabilities and Exposures (CVE) **Obj. 4.4: Given a scenario, apply security best practices.** • Benchmark – Center for Internet Security (CIS) – Vendor-specific

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
	Obj. 4.6: Given a scenario, monitor suspicious activities to identify common attacks. • Attack types – Vulnerability exploitation ▪ Human error ▪ Outdated software – Social engineering ▪ Phishing – Malware ▪ Ransomware – DDoS – Cryptojacking – Zombie instances – Metadata **Obj. 6.3: Given a scenario, troubleshoot security issues.** • Software vulnerability issues
Section 6-2: Virtual Network Security • Allow and Deny Rules • Zero Trust Security • AWS VPCs and Subnets • Azure VNets • GCP VPCs • Securing Hybrid and Multi-Clouds	**Obj. 1.3: Explain cloud networking concepts.** • Network functions, components, and services – Firewalls **Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Security requirements **Obj. 4.4: Given a scenario, apply security best practices.** • Zero Trust **Obj. 4.5: Given a scenario, apply security controls in the cloud.** • Data loss protection (DLP) • Intrusion prevention system/intrusion detection system (IPS/IDS) • Distributed denial-of-service (DDoS) protection • Firewall – Network access control list (ACL) – Network security group **Obj. 4.6: Given a scenario, monitor suspicious activities to identify common attacks.** • Unnecessary open ports

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 6 Securing Cloud Resources *(Continued)*

Sections and headings	Objectives
Section 6-3: Compute Security • Endpoint Protection • Application Security • Pipeline Security	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Security requirements Obj. 4.3: Given a scenario, implement identity and access management. • Secure access to the cloud resources – Secure Shell (SSH) – Remote Desktop Protocol (RDP) Obj. 4.4: Given a scenario, apply security best practices. • Hardening • Patching • Container security – Privileged – Unprivileged – File access permissions Obj. 4.5: Given a scenario, apply security controls in the cloud. • Endpoint protection • Firewall – Web application firewall (WAF) Obj. 5.2: Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines. • Security Obj. 6.3: Given a scenario, troubleshoot security issues. • Software vulnerability issues • Unauthorized software
Section 6-4: Data Security • Encrypting Data • Key Management	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Security requirements Obj. 4.3: Given a scenario, implement identity and access management. • Secure access to the cloud resources – API Obj. 4.4: Given a scenario, apply security best practices. • Encryption – Data in transit – Data at rest • API security • Storage security Obj. 6.2: Given a scenario, troubleshoot network issues. • Protocol deprecations Obj. 6.3: Given a scenario, troubleshoot security issues. • Cipher suite deprecations

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
Section 6-5: Troubleshooting Cloud Security • Security Misconfigurations • Key Management Issues	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Compliance requirements Obj. 6.3: Given a scenario, troubleshoot security issues. • Cipher suite deprecations • Authentication issues – Leaked credentials

Chapter 7: Identity and Access Management

Sections and headings	Objectives
Section 7-1: Cloud Accounts • Identity • Account Types • Account Management	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Security requirements Obj. 4.4: Given a scenario, apply security best practices. • Principle of least privilege
Section 7-2: Authentication • Authentication Processes • Password Policies • MFA (Multifactor Authentication) • Certificate-Based Authentication • Programmatic Access • Security Secrets	Obj. 4.3: Given a scenario, implement identity and access management. • Secure access to the cloud management environment – Programmatic access ▪ Application programming interface (API) ▪ Software development kit (SDK) – Common Language Infrastructure (CLI) – Web portal • Secure access to the cloud resources – API • Authentication models – Local users – Directory-based – Multifactor authentication (MFA) • Accounting Obj. 4.4: Given a scenario, apply security best practices. • Secrets management

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 7 Identity and Access Management *(Continued)*

Sections and headings	Objectives
Section 7-3: Authorization • Authorization Processes • AWS IAM • Azure IAM • GCP IAM	**Obj. 4.2: Compare and contrast aspects of compliance and regulation.** • Data classification **Obj. 4.3: Given a scenario, implement identity and access management.** • Authentication models – Token-based • Authorization models – Role-based access control – Group-based access control – OAuth 2.0 – Discretionary **Obj. 4.5: Given a scenario, apply security controls in the cloud.** • Identity and access management (IAM) policies
Section 7-4: IAM for Hybrid Clouds • SSO (Single Sign-On) • AWS IAM in a Hybrid Cloud • Azure IAM in a Hybrid Cloud • GCP IAM in a Hybrid Cloud	**Obj. 4.3: Given a scenario, implement identity and access management.** • Authentication models – Federation ▪ Security Assertion Markup Language (SAML) – Directory-based – OpenID Connect • Authorization models – OAuth 2.0
Section 7-5: Troubleshooting Cloud IAM • Common IAM Issues • Troubleshooting Privilege Issues	**Obj. 4.4: Given a scenario, apply security best practices.** • Principle of least privilege **Obj. 6.1: Given a scenario, troubleshoot deployment issues.** • Misconfigurations – Permission issues **Obj. 6.3: Given a scenario, troubleshoot security issues.** • Authorization issues – Privilege escalation – Unauthorized access • Authentication issues

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 8: Cloud Storage and Databases

Sections and headings	Objectives
Section 8-1: Storage Types • Data Types • On-Prem Storage Technologies • Cloud Storage Technologies • Types of Databases	Obj. 1.4: Compare and contrast storage resources and technologies. • Storage types – Object storage – Block storage – File storage Obj. 1.7: Compare and contrast virtualization concepts. • Storage – Local – Storage area network (SAN) – Network-attached storage (NAS) Obj. 1.9: Explain the importance of database concepts. • Types – Relational – Non-relational • Deployment options – Self-managed – Provider-managed
Section 8-2: Storage Configuration • Tiering • Storage Performance • Storage Migration • Data Destruction	Obj. 1.4: Compare and contrast storage resources and technologies. • Tiered storage – Hot – Warm – Cold – Archive • Disk types – Solid-state drive (SSD) – Hard disk drive (HDD) • Performance implications • Cost implications Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources. • Storage – Input/output operations per second (IOPS) – Throughput Obj. 2.3: Summarize aspects of cloud migration. • Considerations – Storage – Regulatory – Compliance

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 8 Cloud Storage and Databases *(Continued)*

Sections and headings	Objectives
	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Storage requirements • Availability requirements • Compliance requirements Obj. 4.2: Compare and contrast aspects of compliance and regulation. • Data retention
Section 8-3: Cloud Storage and Database Services • AWS Storage Services • Azure Storage Services • GCP Storage Services	Obj. 1.6: Compare and contrast containerization concepts. • Storage types – Persistent volumes – Ephemeral storage Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Storage requirements • Availability requirements Obj. 3.4: Given a scenario, manage the life cycle of cloud resources. • Data – Ephemeral – Persistent
Section 8-4: Creating and Storing Backups • Backup Strategies • Backup Retention • Backup Types • Backup Considerations	Obj. 1.4: Compare and contrast storage resources and technologies. • Cost implications Obj. 1.7: Compare and contrast virtualization concepts. • Cloning Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Compliance requirements Obj. 3.3: Given a scenario, use appropriate backup and recovery methods. • Backup types – Incremental – Full – Differential • Backup locations – On site – Off site • Schedule • Retention • Replication • Encryption

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
	• Testing – Recoverability – Integrity • Recovery options – Bulk – Granular **Obj. 4.2: Compare and contrast aspects of compliance and regulation.** • Data retention – Litigation hold
Section 8-5: Storage Security • Storage Security Techniques • Data Classification • Data Obfuscation • Regulations for Storage Security	**Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources.** • Compliance requirements **Obj. 4.2: Compare and contrast aspects of compliance and regulation.** • Data sovereignty • Data ownership • Data locality • Data classification • Data retention – Contractual – Regulatory **Obj. 4.4: Given a scenario, apply security best practices.** • Encryption – Data at rest • Storage security – Object storage – File storage **Obj. 4.5: Given a scenario, apply security controls in the cloud.** • Data loss protection (DLP)

Chapter 9: Observability and Monitoring

Sections and headings	Objectives
Section 9-1: Monitoring Resources • Targets to Monitor • Monitoring Tools • Dashboards	**Obj. 1.2: Explain concepts related to service availability.** • Resource availability – Availability monitoring **Obj. 1.8: Summarize cost considerations related to cloud usage.** • Resource metering • Tagging

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 9 Observability and Monitoring *(Continued)*

Sections and headings	Objectives
	Obj. 2.3: Summarize aspects of cloud migration. • Considerations – Service availability Obj. 3.1: Given a scenario, configure appropriate resources to achieve observability. • Monitoring – Metrics
Section 9-2: Events, Logs, and Traces • Events • Logs • Traces	Obj. 3.1: Given a scenario, configure appropriate resources to achieve observability. • Logging – Collection – Aggregation – Retention • Tracing • Alerting – Triage Obj. 4.3: Given a scenario, implement identity and access management. • Accounting – Audit trail Obj. 4.6: Given a scenario, monitor suspicious activities to identify common attacks. • Event monitoring Obj. 5.4: Explain the importance of tools used in DevOps environments. • Elasticsearch, Logstash, and Kibana (ELK) stack • Grafana
Section 9-3: Cloud Monitoring Services • Monitoring in AWS • Monitoring in Azure • Monitoring in GCP	Obj. 2.5: Given a set of requirements, provision the appropriate cloud resources. • Performance requirements • Cost requirements Obj. 3.1: Given a scenario, configure appropriate resources to achieve observability. • Monitoring – Metrics • Alerting – Response Obj. 4.6: Given a scenario, monitor suspicious activities to identify common attacks. • Event monitoring • Deviation from baseline

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
Section 9-4: Troubleshooting Problems and Performance • Quotas and Throttling • Utilization Issues • Scaling and Sizing Issues • Restoration from Backup	Obj. 1.8: Summarize cost considerations related to cloud usage. • Rightsizing Obj. 3.2: Given a scenario, configure appropriate scaling approaches. • Approaches – Triggered ▪ Event Obj. 3.3: Given a scenario, use appropriate backup and recovery methods. • Recovery types – In-place – Parallel Obj. 5.2: Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines. • Automation Obj. 6.1: Given a scenario, troubleshoot deployment issues. • Misconfigurations – Oversubscription – Sizing issues • Resource limits – API throttling

Chapter 10: Cloud Automation

Sections and headings	Objectives
Section 10-1: Automation • Why Automate? • IaC (Infrastructure as Code) • CaC (Configuration as Code) • Comparing Tools • Cloud CLIs	Obj. 1.6: Compare and contrast containerization concepts. • Workload orchestration Obj. 1.10: Compare and contrast methods for optimizing workloads using cloud resources. • Orchestration • Workflow Obj. 2.4: Given a scenario, use code to deploy and configure cloud resources. • Infrastructure as code (IaC) • Configuration as code (CaC) • Repeatability • Drift detection • Versioning • Testing • Documentation • Formats – JavaScript Object Notation (JSON) – Yet Another Markup Language (YAML)

(continues)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 10 Cloud Automation *(Continued)*

Sections and headings	Objectives
	Obj. 5.1: Explain source control concepts. • Version management **Obj. 5.2: Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.** • Automation • Testing • Workflow **Obj. 5.4: Explain the importance of tools used in DevOps environments.** • Ansible • Terraform
Section 10-2: Automating Code Pipelines • CI/CD Pipelines • Scripting Logic	**Obj. 2.4: Given a scenario, use code to deploy and configure cloud resources.** • Scripting logic – Variables – Conditionals – Operators – Data types – Functions • Testing • Formats – JavaScript Object Notation (JSON) – Yet Another Markup Language (YAML) **Obj. 5.2: Explain concepts related to continuous integration/continuous deployment (CI/CD) pipelines.** • Automation • Code integration • Code deployment – Build • Testing • Artifacts – Images ▪ VM ▪ Container – Packages ▪ Red Hat Package Manager (RPM) ▪ Debian ▪ ZIP ▪ tar – Flat file • Repositories – Public – Private

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sections and headings	Objectives
Section 10-3: Cloud Maintenance • Types of Updates • Patch and Update Methodologies • Decommissioning	Obj. 3.4: Given a scenario, manage the life cycle of cloud resources. • Patches • Updates – Major – Minor • Testing • Decommissioning – End of life – End of support
Section 10-4: Troubleshooting Automation Issues • Common Automation Issues • Breakdowns in Workflow	Obj. 6.1: Given a scenario, troubleshoot deployment issues. • Incompatibility • Outdated component definitions • Deprecation of functionality

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Appendix C

CompTIA Cloud+ Acronyms

Table C-1 provides a complete list of acronyms listed in the CompTIA Cloud+ CV0-004 certification exam objectives. The official list of objectives is available on CompTIA's website: comptia.org.

Table C-1 CompTIA Cloud+ Acronyms

Acronym	Chapter	Spellout
ACL	6	Access Control List
AES	6	Advanced Encryption Standard
AI	1	Artificial Intelligence
API	1	Application Programming Interface
AZ	4	Availability Zone
BGP	5	Border Gateway Protocol
BYOD	1	Bring Your Own Device
CaC	10	Configuration as Code
CDN	8	Content Delivery Network
CI/CD	10	Continuous Integration/Continuous Deployment
CIS	6	Center for Internet Security
CLI	7	Common Language Infrastructure
CPU	2	Central Processing Unit
CRM	1	Customer Relationship Management
CRUD	3	Create, Read, Update, Delete
CSA	1	Cloud Security Alliance
CSP	1	Cloud Service Provider
CVE	6	Common Vulnerabilities and Exposures

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Acronym	Chapter	Spellout
CVSS	6	Common Vulnerability Scoring System
CWE	6	Common Weakness Enumeration
CWSS	6	Common Weakness Scoring System
DBaaS	1	Database as a Service
DDoS	6	Distributed Denial of Service
DHCP	2	Dynamic Host Configuration Protocol
DLP	6	Data Loss Prevention
DNS	4	Domain Name System
DR	3	Disaster Recovery
DSS	3	Data Security Standard
ELK	9	Elasticsearch, Logstash, and Kibana
FaaS	2	Function as a Service
GDPR	3	General Data Protection Regulation
GPU	2	Graphics Processing Unit
HDD	2	Hard Disk Drive
HTTP	3	Hypertext Transfer Protocol
IaaS	1	Infrastructure as a Service
IaC	10	Infrastructure as Code
IAM	7	Identity and Access Management
ICMP	5	Internet Control Management Protocol
IDS	6	Intrusion Detection System
IOPS	8	Input/Output Operations Per Second
IP	2	Internet Protocol
IPS	6	Intrusion Prevention System
iSCSI	8	Internet Small Computer System Interface
ISO	3	International Organization for Standardization
ISP	1	Internet Service Provider
ITIL	3	Information Technology Infrastructure Library
JSON	8	JavaScript Object Notation
LAN	5	Local Area Network
LDAP	7	Lightweight Directory Access Protocol
LUN	8	Logical Unit Number
MFA	7	Multifactor Authentication

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Acronym	Chapter	Spellout
ML	1	Machine Learning
MTU	4	Maximum Transmission Unit
NAS	8	Network Attached Storage
NAT	2	Network Address Translation
NIC	2	Network Interface Card
NoSQL	8	Not Only Structured Query Language
NTP	3	Network Time Protocol
NVME	8	Non-volatile Memory Express
OAuth	7	Open Authorization
OIDC	7	OpenID Connect Protocol
OS	2	Operating System
PaaS	1	Platform as a Service
PCI	3	Payment Card Industry
RACI	6	Responsible, Accountable, Consulted, Informed
RAID	8	Redundant Array of Inexpensive Disks
RAM	2	Random-access Memory
RDP	3	Remote Desktop Protocol
REST	3	Representational State Transfer
RPC	3	Remote Procedure Call
RPM	10	Red Hat Package Manager
RPO	3	Recovery Point Objective
RTMP	5	Real-time Messaging Protocol
RTO	3	Recovery Time Objective
SaaS	1	Software as a Service
SAML	7	Security Assertion Markup Language
SAN	8	Storage Area Network
SDK	7	Software Development Kit
SDN	4	Software-defined Network
SOAP	3	Simple Object Access Protocol
SOC2	3	System and Organization Controls 2
SQL	8	Structured Query Language
SSD	2	Solid-state Drive
SSH	3	Secure Shell

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Acronym	Chapter	Spellout
SSL	6	Secure Sockets Layer
SSO	7	Single Sign-On
STAR	3	Security, Trust, Assurance, Risk
TCP	5	Transmission Control Protocol
TLS	6	Transport Layer Security
USB	2	Universal Serial Bus
vCPU	2	Virtual CPU
VDI	2	Virtual Desktop Interface
VLAN	5	Virtual LAN
VM	1	Virtual Machine
vNIC	2	Virtual NIC
VPC	4	Virtual Private Cloud
VPN	1	Virtual Private Network
WAF	6	Web Application Firewall

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Molibdenis-Studio/Shutterstock.com

Appendix D

Rubric for Hands-On Projects and Challenge Projects

Criteria	Beginning	Developing	Proficient	Exemplary	Score
Responses to questions	All missing or incorrect [0 points]	Most missing or incorrect [15 points]	Few missing or incorrect [20 points]	All complete [25 points]	
Other deliverables	Missing [0 points]	Present but missing most or all the required information [15 points]	Present and missing some of the required information [20 points]	Present and contains all the required information [25 points]	
Critical thinking and engagement	Student shows little to no evidence of attempting to meet the performance requirements of the assignment [0 points]	Student retains their existing understanding while attempting to meet the performance requirements of the assignment [15 points]	Student challenges their existing understanding and shows evidence of new learning [20 points]	Student challenges their existing understanding and displays creative and original insights [25 points]	
Mechanics	Grammar, spelling, punctuation, and formatting make student's message difficult to understand [0 points]	Grammar, spelling, punctuation, and formatting detract from student's message [15 points]	Grammar, spelling, punctuation, and formatting support student's message [20 points]	Grammar, spelling, punctuation, and formatting enhance student's message [25 points]	
				Total	

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Calculating Subnets

Chapter 4 provides context for how subnetting works and how subnets contribute to cloud network management and security. If you need additional details on how to calculate subnets, this appendix guides you through these calculation processes. Note that the examples in this appendix use physical network devices, such as routers and switches on a LAN. However, the same concepts and calculations apply in a cloud environment where IP addressing is used for virtual devices such as route tables and VPCs.

Subnetting Questions on Exams

Although it's impossible to know for sure, you're likely to encounter two types of subnet calculation problems on some cloud certification exams:

- Given certain network requirements (such as required number of hosts or required number of subnets), calculate possible subnets and host IP address ranges.
- Given an IP address, determine its subnet's network ID, broadcast address, and first/last host addresses.

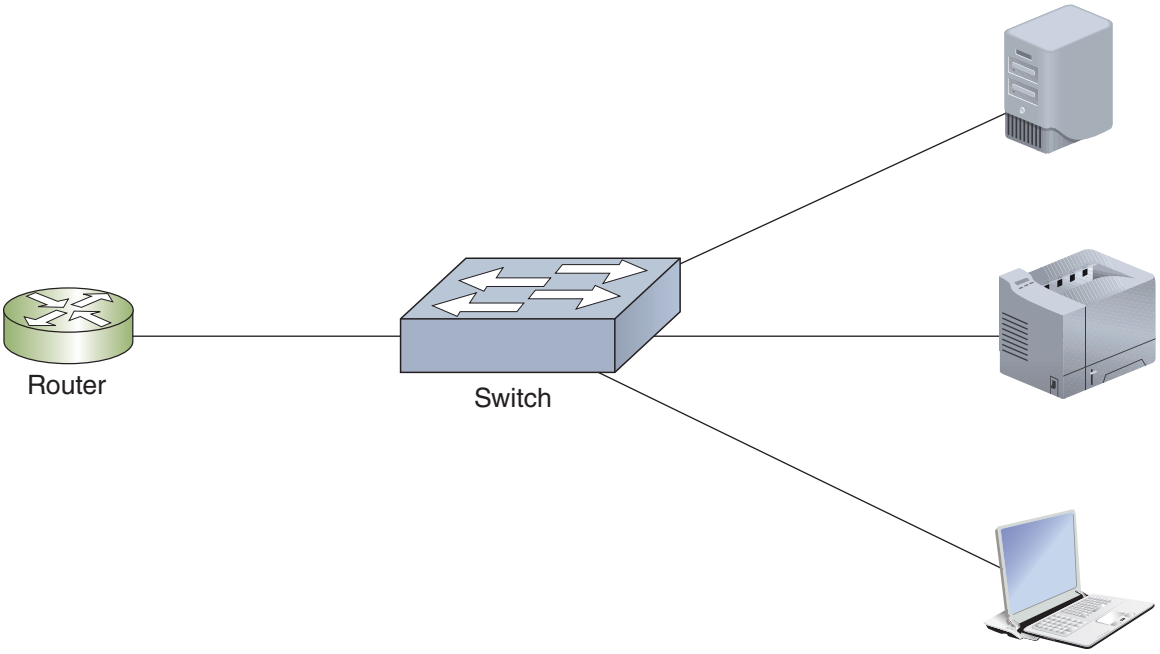
Consider the following example of a network that initially contains only a single class C subnet, and all its hosts (up to 254) are on the same subnet. By borrowing only one bit from the host ID, you can have two subnets with up to 126 hosts (a more manageable number) per subnet. Let's step through how this works.

IPv4 Subnet Calculation in Binary

Suppose you have a network, as illustrated in Figure E-1, with one router, a switch connected to one interface on the router, and your endpoint devices connected to that one switch.

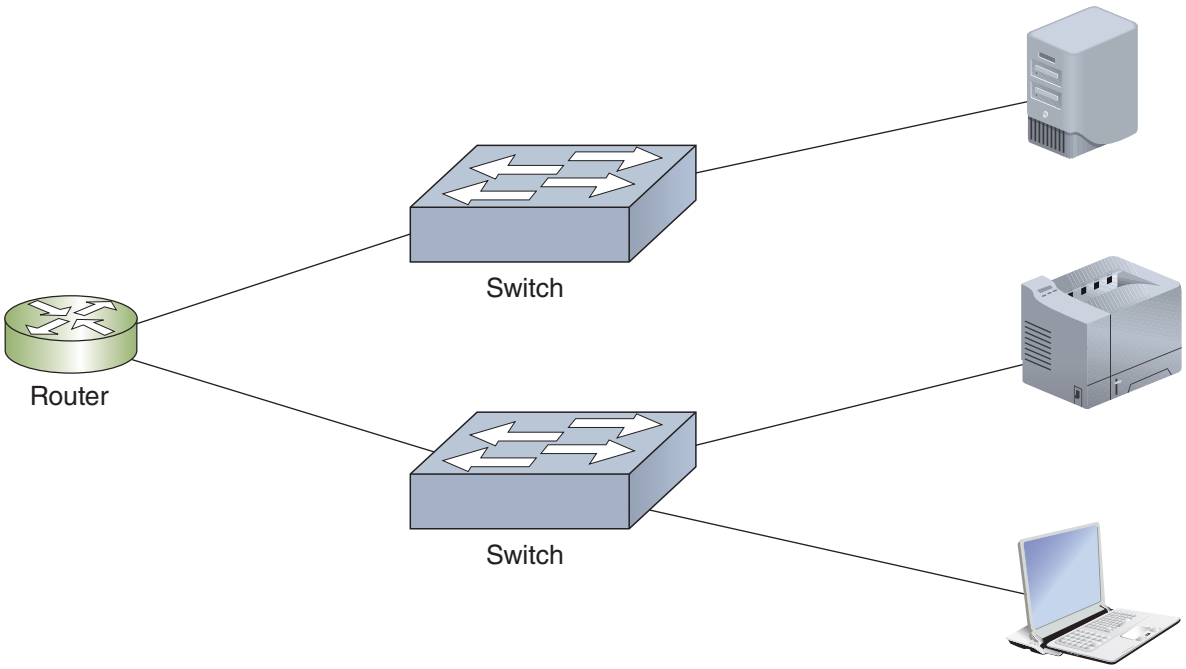
Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure E-1 Network with one switch (one LAN) connected to one router



You then add a second switch to a different interface on the router to divide your local network into two LANs—each router interface connects to a single LAN. You can then redistribute your network devices so some of them stay connected to the LAN on the first switch, and others are moved to the new LAN on the second switch (Figure E-2).

Figure E-2 Network with two switches (two LANs) connected to one router



Note E-1 | In this example and in many others in this appendix, a bold red font is used for the network ID portion of an IP address.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The network ID of the original network is 192.168.89.0, and its subnet mask is 255.255.255.0. From this IP address block, let's create two subnets of IP addresses, one for each LAN:

1. **Borrow from host bits**—Initially, the network ID is 24 bits. First convert the network ID to binary:

   Network ID 192.168.89 in binary:

   - **11000000.10101000.01011001**

   Borrow one bit from the host portion to give to the network ID, which will then have 25 bits (notice one additional red, underlined bit in the fourth octet). Here, the borrowed bit is an "x" to show it can have a value of 0 or 1:

   - **11000000.10101000.01011001.x**

     How many subnets do you have now? The underlined red bit can be either a 0 or a 1, which gives you the possibility of two subnets:

   - **11000000.10101000.01011001.0**
   - **11000000.10101000.01011001.1**

2. **Determine the subnet mask**—Recall that the subnet mask marks the bits in an IP address that belong to the network ID. Therefore, the subnet mask for both subnets is as follows:

   - **11111111.11111111.11111111.1**0000000 or decimal 255.255.255.128

   To calculate that last octet, you convert binary 10000000 to decimal, which is 128. You can use a calculator to do the conversion (if you have access to a calculator), manually calculate it, or memorize this and a handful of other common binary-to-decimal conversions.

3. **Determine the network IDs**—Recall that in the network ID, the underlined red bit can be a 0 or 1—these two possibilities yield two subnets. Therefore, the network ID for each subnet is as follows:

   - Subnet A: **11000000.10101000.01011001.0**0000000 or decimal 192.168.89.0
   - Subnet B: **11000000.10101000.01011001.1**0000000 or decimal 192.168.89.128

   In CIDR notation, which identifies the first 25 bits as the network portion, the network ID for each subnet is the following:

   - Subnet A: 192.168.89.0/25
   - Subnet B: 192.168.89.128/25

4. **Determine the ranges of IP addresses for hosts in the subnet**—Start with the range of available IP addresses for subnet A. For host addresses, use the last seven bits in the last octet. (The first bit for this octet in Subnet A is always 0 and belongs to the network ID.) Start counting in binary and convert to decimal:

   - **0**0000000 is not used because it's the network ID for this subnet
   - **0**0000001 or decimal 1
   - **0**0000010 or decimal 2
   - **0**0000011 or decimal 3
   - . . .
   - **0**1111110 or decimal 126
   - **0**1111111 or decimal 127, which is reserved for broadcasting within the subnet and cannot be used as a host address

   The range of possible IP addresses, then, is 192.168.89.0 through 192.168.89.127 (which is 128 possibilities). However, the first and last addresses cannot be used. Therefore, the range of host IP addresses for subnet A is 192.168.89.1 through 192.168.89.126 (that's 128 – 2, yielding 126 possibilities).

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For subnet B, the first bit of the last octet is always 1, and the range of host addresses is as follows:

- 10000000 is not used because it's the network ID for this subnet
- 10000001 in decimal: 129
- 10000010 in decimal: 130
- 10000011 in decimal: 131
- . . .
- 11111110 in decimal: 254
- 11111111 in decimal: 255, which is reserved for broadcasting within the subnet and cannot be used as a host address

Therefore, the range of host IP addresses for subnet B is 192.168.89.129 through 192.168.89.254 (another 126 possibilities).

The results of each of the preceding steps are summarized in Table E-1.

Table E-1 Steps to divide IP addresses for network ID 192.168.89.0 into two subnets

Step 1: Borrow from host bits.				
Network ID	192	168	89	0
In binary	11000000	10101000	01011001	00000000
Borrow 1 bit	11000000	10101000	01011001	x0000000
Step 2: Determine the subnet mask.				
In binary	11111111	11111111	11111111	10000000
In decimal	255	255	255	128
Step 3: Determine the network IDs.				
Network ID A	11000000	10101000	01011001	00000000
In decimal	192	168	89	0
In CIDR notation	192.168.89.0/25			
Network ID B	11000000	10101000	01011001	10000000
In decimal	192	168	89	128
In CIDR notation	192.168.89.128/25			
Step 4: Determine range of host IP addresses.				
Subnet A:				
First host, binary	11000000	10101000	01011001	00000001
First host, decimal	192	168	89	1
Last host, binary	11000000	10101000	01011001	01111110
Last host, decimal	192	168	89	126

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Subnet B:				
First host, binary	11000000	10101000	01011001	10000001
First host, decimal	192	168	89	129
Last host, binary	11000000	10101000	01011001	11111110
Last host, decimal	192	168	89	254

Notice that you gained a network (the one larger network became two smaller subnets), but you lost some host addresses (each subnet offers a possible 126 host addresses for a total of 252 hosts, which is two fewer than the original network's possible 254 hosts). The difference is because each subnet needs its own network ID and broadcast address. Instead of the original network's one network ID and one broadcast address, you now have two network IDs and two broadcast addresses, which reduces the total number of available host addresses. Subnetting offers many advantages, but one disadvantage is that you lose possible host addresses each time you divide a network into more subnets.

IPv4 Subnet Calculations Using Formulas

Now you're ready to move on to a more complicated example, performing calculations using formulas and without so much binary involved. Exploring the process in binary helps you understand what's happening "under the hood," so to speak, to clarify the math of what is happening. Realistically, you'll more likely use formulas if you need to calculate subnets on an exam because using formulas typically is a lot faster.

Suppose you want to divide your local network, which has a network ID of 192.168.89.0, into six subnets to correspond to an office building's six floors. The following steps guide you through the process:

1. **Decide how many bits to borrow**—How many bits must you borrow from the host portion of the original network ID to get six subnets? Use this formula to determine the number of bits:

$$2^n = Y$$

   - **n** is the number of bits that must be switched from the host address to the network ID.
   - **Y** is the number of subnets that result.

   The given scenario requires six separate subnets, meaning that **Y**, in this case, should be equal to or greater than 6. Experiment with different values for **n** until you find a value large enough to give you at least the number of subnets you need. For example, $2^2 = 4$; however, 4 is not high enough. Instead consider that $2^3 = 8$; this will give you enough subnets to meet your current needs and allow room for future growth. With 3 being the exponent, replace **n** in the formula with 3. This tells you that three bits in the host addresses of your class C network must become network ID bits for subnets. This also means that three bits in your subnet mask must change from 0 to 1.

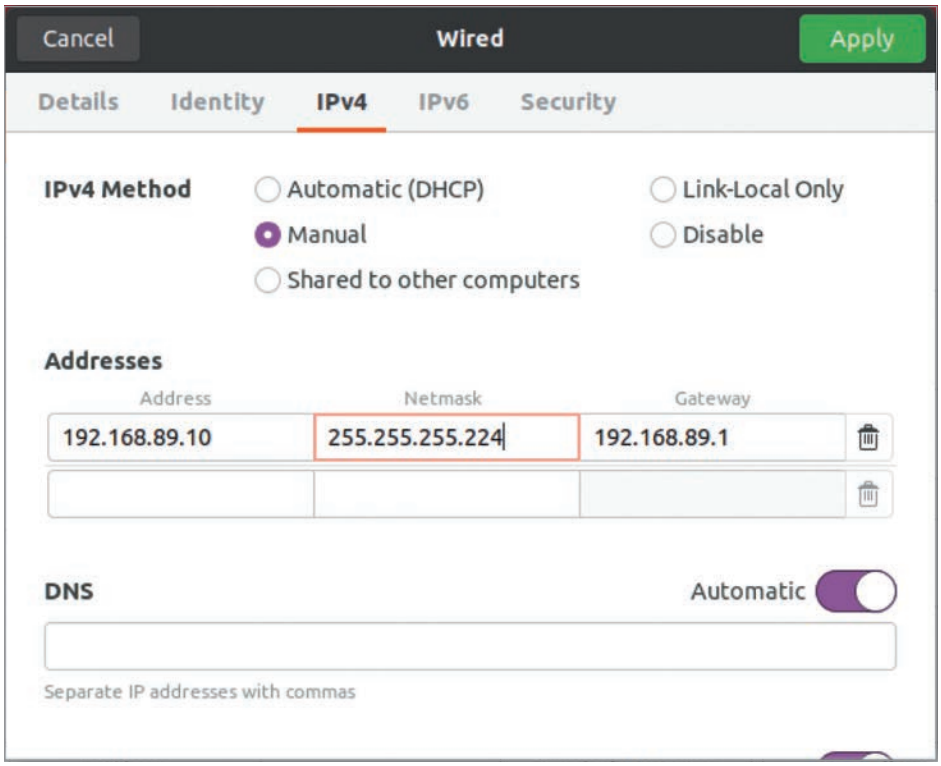
2. **Determine the subnet mask**—The default subnet mask for a class C network is **255.255.255**.0, or **11111111 11111111 11111111** 00000000. In this default subnet mask, the first 24 bits indicate the position of network information.

   Changing three of the default subnet mask's bits from host to network information gives you the subnet mask **11111111 11111111 11111111 111**00000. In this modified subnet mask, the first 27 bits indicate the bits for the network ID. For the original class C network whose network ID is 192.168.89.0, the slash notation for its first subnet would now be 192.168.89.0/27 because 27 bits of the subnet's address are used to provide network information. All other network IDs and IP addresses for the remaining subnets will also require the /27 CIDR block.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Converting from binary to the more familiar dotted decimal notation, this subnet mask becomes 255.255.255.224 because 11100000 in binary in the final octet is 224 in decimal. When you configure the TCP/IP properties of clients on your network (Figure E-3), you'll specify this new subnet mask.

Figure E-3 IPv4 client configuration for a subnet in Ubuntu Desktop



Source: Canonical Group Limited

Note E-2 | When examining the subnet mask for a network, if any octet is not the usual value of 255 or 0, you know that this network is a subnet and classful addressing is not used. The unusual number (224 in the example) is often called the "interesting octet." Subtract the interesting octet value from 256 and you get what is called the magic number. In this example, the magic number is 256 - 224 = 32. This magic number can be used to calculate the network IDs in all the subnets of the larger network, which you'll practice next.

3. **Calculate the network ID for each subnet**—The first three octets of the network ID for the original class C network (**192.168.89**) are the same for all eight possible subnets. The network IDs differ in the last octet. Use the magic number (as described in Note E-2) to calculate each subnet's network ID as follows:

- **Subnet 1**: **192.168.89.0**
- **Subnet 2**: 192.168.89.0 + 32 yields **192.168.89.32**
- **Subnet 3**: 192.168.89.32 + 32 yields **192.168.89.64**
- **Subnet 4**: 192.168.89.64 + 32 yields **192.168.89.96**
- **Subnet 5**: 192.168.89.96 + 32 yields **192.168.89.128**
- **Subnet 6**: 192.168.89.128 + 32 yields **192.168.89.160**
- **Subnet 7**: 192.168.89.160 + 32 yields **192.168.89.192**
- **Subnet 8**: 192.168.89.192 + 32 yields **192.168.89.224**

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This method of adding on the same number over and over is called skip-counting. You probably learned this technique in elementary school. For example, skip-counting by twos gives you 0, 2, 4, 6, 8, 10, and so on. Skip-counting by threes gives 0, 3, 6, 9, 12, and so on. In the last octet, you can skip-count by the magic number to get the subnets' final octets. For this scenario, the magic number is 32 and you get the eight subnets' final octets of 0, 32, 64, 96, 128, 160, 192, and 224.

4. **Determine the IP address range for hosts in each subnet**—Recall that you have borrowed three bits from what used to be host information in the IP address. This leaves five bits instead of eight available in the last octet to identify hosts. To calculate the number of possible addresses, keep in mind that each of the five bits has two possible values, a 0 or a 1. Therefore, the number of possible addresses is $2^5 = 2 \times 2 \times 2 \times 2 \times 2 = 32$. But you can't use two of these addresses for hosts because one is used for the subnet's network ID (the one where all five host bits are 0 in binary) and one for the broadcast address (the one where all five host bits are 1 in binary). This leaves you 30 host addresses in each subnet. As a shortcut to calculating the number of hosts per subnet, you can use the following formula:

$$2^h - 2 = Z$$

- **h** is the number of bits remaining in the host portion.
- **Z** is the number of hosts available in each subnet.

In summary, $2^5 - 2$ yields 30 possible hosts per subnet.

In this example, you can have a maximum of 8 (number of subnets) × 30 (number of hosts per subnet), or 240, unique host addresses on the entire, larger network. Recall that each time you subnet a network, you lose two possible host addresses with each new subnet. This overhead is the price you pay for subnetting a network in exchange for the advantages you gain.

When calculating subnets, you'll work with the following information, some of which is initially known and some of which must be calculated:

- Number of subnets
- Number of host addresses per subnet
- Network ID for each subnet
- Broadcast address for each subnet
- Range of possible host addresses within each subnet

In this example, you've already calculated the number of subnets, number of host addresses, and each subnet's network ID. Once you know the network ID of the subnets, calculating the address range of hosts and each subnet's broadcast address is relatively simple. For example, take subnet 5. The network ID is 192.168.89.128. The broadcast address is the last address before the next subnet's network ID. In this case, that's 192.168.89.160 – 1, which is 192.168.89.159.

For the host addresses in this subnet, start back at the network ID. You won't use the network ID for a host address, so you start with the next value and keep going until you reach the broadcast address for the subnet. Therefore, the available host address range for subnet 5 includes 192.168.89.129 through 192.168.89.158, yielding for this particular subnet a total of 30 available host addresses.

Table E-2 lists the network ID, broadcast address, and the range of usable host addresses for each of the eight subnets in this subnetted class C network. Together, the existing network ID along with the additional bits used for subnet information are sometimes called the extended network prefix.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table E-2 Subnet information for eight subnets in the sample IPv4 class C network

Subnet number	Network ID (extended network prefix)	Range of host addresses	Broadcast address
1	192.168.89.0 or 11000000 10101000 01011001 000**00000**	192.168.89.1-30	192.168.89.31 or 11000000 10101000 01011001 000**11111**
2	192.168.89.32 or 11000000 10101000 01011001 001**00000**	192.168.89.33-62	192.168.89.63 or 11000000 10101000 01011001 001**11111**
3	192.168.89.64 or 11000000 10101000 01011001 010**00000**	192.168.89.65-94	192.168.89.95 or 11000000 10101000 01011001 010**11111**
4	192.168.89.96 or 11000000 10101000 01011001 011**00000**	192.168.89.97-126	192.168.89.127 or 11000000 10101000 01011001 011**11111**
5	192.168.89.128 or 11000000 10101000 01011001 100**00000**	192.168.89.129-158	192.168.89.159 or 11000000 10101000 01011001 100**11111**
6	192.168.89.160 or 11000000 10101000 01011001 101**00000**	192.168.89.161-190	192.168.89.191 or 11000000 10101000 01011001 101**11111**
7	192.168.89.192 or 11000000 10101000 01011001 110**00000**	192.168.89.193-222	192.168.89.223 or 11000000 10101000 01011001 110**11111**
8	192.168.89.224 or 11000000 10101000 01011001 111**00000**	192.168.89.225-254	192.168.89.255 or 11000000 10101000 01011001 111**11111**

Note E-3 | You can also calculate the magic number by raising 2 to the power of the number of bits in the host portion of the subnet mask's interesting octet. Use this formula:

$$2^h = \text{magic number}$$

In this example, the host portion has five bits, meaning **h** is 5. Therefore, the magic number is $2^5 = 32$. You can then use this number to determine the subnets' network IDs.

Several good websites and apps can give you a variety of practice subnet questions so you can become especially comfortable with calculating subnets. Search online for "subnetting practice" or try an app on your phone. A good subnetting practice app will give you options on the types of problems to work with, tell you the correct answers, and explain why those answers are correct.

Calculate IPv4 Subnets and Host IP Address Ranges

Recall that you'll typically find two types of subnet calculation problems on some networking exams:

- Given certain network requirements (such as required number of hosts or required number of subnets), calculate possible subnets and host IP address ranges.
- Given an IP address, determine its subnet's network ID, broadcast address, and first/last host addresses.

Now that you've worked through steps for solving the first type of problem using a class C network ID, you're ready to work through another example of the same type of problem, but this time you'll begin with a class B network ID. Then you'll work through an example of the second type of subnetting problem.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Suppose your organization uses the class B network ID of 172.20.0.0 for its entire network, and you need to create 15 subnets. Complete the following steps, responding to the questions as you go:

1. First decide how many bits to borrow from the host address bits. Recall that you can use the formula $\mathbf{2^n = Y}$. To get at least 15 new subnets (without creating more subnets than necessary), how many bits must be borrowed from the host address portion? How many bits total will be used for identifying a host's subnet?
2. With this information, you can now calculate the subnet mask. The default subnet mask for a class B network is 255.255.0.0, and so the third octet is the one that will change. What is the subnet mask for these subnets, written in dotted decimal notation?
3. The magic number will tell you by what amount to skip-count when you're listing the subnets' network IDs. There are two ways to calculate the magic number: Subtract the interesting octet's value from 256, or use the formula $\mathbf{2^h}$ where **h** is the number of host bits in the interesting octet. What is the magic number you can use to calculate the network IDs?
4. Now you can calculate the network IDs for each subnet. Begin with the original network ID. Then, in the third octet, count up by the magic number with each iteration. The last subnet's network ID will be equal to 256 minus the magic number, because you can't use 256 itself in any IP address. What is the CIDR notation for the first subnet's network ID? For the second subnet's network ID? For the last subnet's network ID?
5. If 20 bits are used to identify the network and subnet, this leaves 12 bits to identify each host. Using the formula $\mathbf{2^h - 2 = Z}$, where **h** is the total number of host bits, how many host addresses are possible in each subnet? (You might need a calculator for this step, such as the Windows Calculator app in Scientific mode. Note that it's unusual to get an exam question where you're working with bits in the third octet if you're not allowed to use a calculator during the exam, unless you're taking an advanced networking exam.)
6. The range of available host addresses consists of all the possible IP addresses between the network ID and the broadcast address (which is 1 below the network ID for the next subnet). What is the range of host addresses for the first subnet? For the second subnet? For the last subnet?

Now you're ready to work through the other type of subnetting problem that you'll likely find on some exams: calculating a host's network information.

Calculate an IPv4 Host's Network Information

This time, you'll work backwards in your calculations by starting with one host's IP address information. Suppose a server on your network displays the following IPv4 network configuration:

IPv4 address: 192.168.89.130

Subnet mask: 255.255.255.224

Your task is to determine the network ID of the subnet this server is located on, the broadcast address, and the range of available host addresses on this subnet. Complete the following steps, responding to the questions as you go:

1. You don't necessarily need to use binary for these calculations. Find the magic number and go from there. As with your earlier calculations, you need to subtract the interesting octet's value from 256 to get the magic number. What is the magic number?
2. If the interesting octet is the final octet of the subnet mask, you can assume the first three octets of the IP address identify the classful network ID before this network was subnetted. This network ID also serves as the network ID for the first subnet. What is the network ID of the first subnet?

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. You can now use the magic number to calculate the remaining subnets' network IDs. What is the second subnet's network ID? What is the final subnet's network ID?
4. To narrow this down to your server's subnet, either skip-count up from a lower numbered subnet or skip-count down from a higher numbered subnet. Either way will work. You're trying to find a network ID that is as close to the server's IP address as possible without going over. What is the network ID of the server's subnet?
5. You can check the next higher subnet's network ID and subtract 1 to determine the broadcast address of the server's subnet. What is the broadcast address?
6. Finally, any IP address between the subnet's network ID and its broadcast address is the range of available host IP addresses. What is this range?

Subnet Masks and the Logical ANDing Function

To calculate a host's network ID given its IPv4 address and subnet mask, computers follow a logical process of combining bits known as ANDing. In ANDing, a bit with a value of 1 combined, or ANDed, with another bit with a value of 1 results in a 1. A bit with a value of 0 ANDed with any other bit results in a 0. If you think of 1 as "true" and 0 as "false," the logic of ANDing makes sense: ANDing a true statement to a true statement still results in a true statement. But ANDing a true statement to a false statement results in a false statement.

ANDing logic is demonstrated in Table E-3, which provides every possible combination of having a 1 or 0 bit in an IPv4 address or subnet mask.

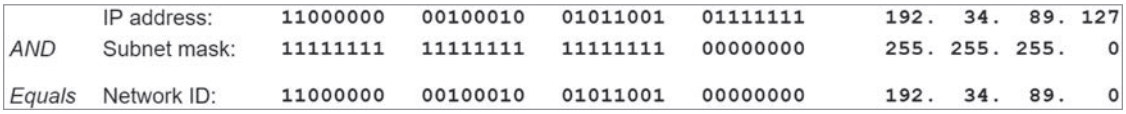
Table E-3 ANDing

IP address bit	1	1	0	0
Subnet mask bit	1	0	1	0
Resulting bit when ANDed	1	0	**0**	0

A sample IPv4 host address, its default subnet mask, and its network ID are illustrated in Figure E-4 in both binary and dotted decimal notation. Notice that the address's fourth octet could have been composed of any combination of 1s and 0s, and the network ID's fourth octet would still result in all 0s.

Figure E-4 shows how ANDing logic is applied to an IPv4 address combined with a default subnet mask. It works the same way for networks that are subnetted with classless subnet masks.

Figure E-4 Example of ANDing a host's network ID



Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Self-Check

1. What is the minimum number of bits that should be borrowed to create 14 subnets?
   - **a.** 4 bits
   - **b.** 2 bits
   - **c.** 16 bits
   - **d.** 8 bits

2. Suppose you're calculating the range of host IP addresses for a subnet (the targeted subnet). If the next subnet's network ID is 192.168.42.128, what is the targeted subnet's broadcast address?
   - **a.** 192.168.42.96
   - **b.** 192.168.42.0
   - **c.** 192.168.42.127
   - **d.** 192.168.42.255

Check your answers at the end of this appendix.

Solutions to Self-Check Questions

Appendix E: Calculating Subnets

1. What is the minimum number of bits that should be borrowed to create 14 subnets?

   Answer: a. 4 bits

   Explanation: Use the formula $2^n = Y$ to determine how many bits to borrow to create a desired number of subnets, where **n** is the number of bits to borrow and **Y** is the number of subnets that number of borrowed bits will create. 2^3 equals 8, which is not enough subnets for the targeted 14 subnets. 2^4 equals 16, which is sufficient for the targeted 14 subnets. Therefore, **4 bits** should be borrowed from the host portion for the network portion.

2. Suppose you're calculating the range of host IP addresses for a subnet (the targeted subnet). If the next subnet's network ID is 192.168.42.128, what is the targeted subnet's broadcast address?

   Answer: c. 192.168.42.127

   Explanation: The broadcast address is always 1 below the network ID for the next subnet. Therefore, 192.168.42.(128 – 1) yields **192.168.42.127**.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Glossary

A

access key A string of characters used to authenticate to a system, similar in concept to a password. *Ch. 6*

access port A switch port used to connect VLAN (virtual local area network) client devices, such as servers or workstations, that are intended to communicate only on a single VLAN. *Ch. 5*

actuator A device that converts digital commands into physical actions. *Ch. 1*

AD (Active Directory) A collection of services on Windows Server that manages access to resources on a network. *Ch. 7*

affinity The requirement that guest VMs reside on the same physical host. *Ch. 2*

agent A lightweight application that runs in the background, automatically performing tasks on behalf of a user or another program. *Ch. 9*

AI (artificial intelligence) The ability of computers to adapt to changing circumstances. *Ch. 1*

ALB (application load balancer) A load balancer that examines application-layer data to determine where to send incoming traffic across available targets. *Ch. 5*

Ansible A cloud-agnostic automation tool for managing infrastructure configuration. *Ch. 10*

anti-affinity The requirement that guest VMs reside on different physical hosts. *Ch. 2*

API (application programming interface) A request made in a specific format to a program. *Ch. 1*

API (application programming interface) call A message sent between system components using each system's API. *Ch. 3*

API (application programming interface) throttling A restriction on the number of API requests allowed to each customer's APIs. *Ch. 9*

application gateway A service that handles user authentication to give users remote access to applications without requiring a VPN (virtual private network). *Ch. 5*

artifact A single file or a collection of data and files produced by one resource in one step of a pipeline and consumed by another resource in another step of the pipeline. *Ch. 10*

audit trail A collection of data in logs that is consistent and thorough enough to retroactively prove compliance and also to defensibly prove user actions. *Ch. 9*

authentication A process usually managed by a server that proves the identity of a client and determines whether that client is allowed to access a secured system. *Ch. 7*

authorization A process that determines what resources a user or an application can access after their identity has been confirmed. *Ch. 7*

auto scaling The ability of a virtual resource to automatically increase or decrease in capacity as demand on the resource changes. *Ch. 1*

automation The use of technology to perform a task with minimal human intervention. *Ch. 1*

AWS (Amazon Web Services) A subsidiary of Amazon that is headquartered in Seattle, Washington, and provides extensive cloud computing services to businesses and individuals. *Ch. 1*

AZ (availability zone) One or more data centers in a geographic region that provides power, cooling, and network infrastructure independently of other AZs. *Ch. 2*

Azure Microsoft's cloud computing platform that is designed for optimal compatibility with existing Microsoft products. *Ch. 1*

B

bastion host A VM residing in a public subnet that can be accessed over the Internet from an administrator's local computer, providing the ability to remote into other VMs protected in private subnets. *Ch. 4*

BC (business continuity) A company's ability to weather a failure, crisis, or disaster while maintaining continuity of operations, especially for critical services to customers and income-generating activities. *Ch. 3*

BGP (Border Gateway Protocol) Dubbed the "protocol of the Internet," this routing protocol is the only current EGP (exterior gateway protocol) and is capable of considering many factors in its routing metrics. *Ch. 5*

big data Large amounts of data generated at high speeds and collected from many sources, resulting in data of such high volume, variety, and velocity that special tools are required to capture, analyze, and store the data. *Ch. 8*

broad network access The ability to connect to cloud-hosted resources from anywhere on the Internet using a variety of device types. *Ch. 1*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

C

CaC (configuration as code) A process that uses code to monitor and enforce the configuration of a monitored resource. *Ch. 10*

CapEx (capital expenditures) Costs whose benefits are experienced over a long period of time after the investment is made. *Ch. 3*

CD (continuous delivery) A system that, upon manual approval to begin, automatically delivers code changes to the next environment in the pipeline. *Ch. 10*

CD (continuous deployment) A system that automatically deploys code changes to the next environment in the pipeline as soon as the changes pass all required tests in the previous stage. *Ch. 10*

CDN (content delivery network) A distributed storage structure that allows customers to store files in locations closer to where their users are located. *Ch. 5*

chain of custody Documentation that describes evidence, including when it was collected, who collected it, its condition, and how it was secured and transferred from one responsible party to the next. *Ch. 3*

CI (continuous integration) A practice of frequently merging code changes into a central repository where tests are performed automatically. *Ch. 10*

CIA (confidentiality, integrity, and availability) triad A three-tenet, standard security model describing the primary ways that encryption protects data. Confidentiality ensures that data can only be viewed by its intended recipient or at its intended destination. Integrity ensures that data was not modified after the sender transmitted it and before the receiver picked it up. Availability ensures that data is available to and accessible by the intended recipient when needed. *Ch. 6*

CIDR (classless interdomain routing) notation A shorthand method for identifying network and host bits in an IP address. *Ch. 4*

CIS (Center for Internet Security) A nonprofit organization that draws on expertise from IT professionals in government, business, and academia to define best practices for cybersecurity across a wide range of technologies and platforms. *Ch. 6*

CIS benchmark A guideline from CIS (Center for Internet Security) that specifies secure configurations of various technologies categorized by vendor, such as AWS, Microsoft, and GCP. *Ch. 6*

clone A copy of a VM's design and data. *Ch. 2*

cloud bursting Temporary reliance on cloud resources to accommodate short spikes in demand. *Ch. 1*

cluster A group of devices or resources that appear as a single device or resource to the rest of the network. *Ch. 2*

CM (configuration management) tool An automation tool used to configure, update, and reconfigure resources over time. *Ch. 10*

commit A version of a project. *Ch. 3*

community cloud A cloud deployment model in which data storage, applications, or services are hosted on hardware shared by a specific group of organizations and centrally managed by a service provider. *Ch. 1*

conditional statement A statement that must be evaluated to determine the next step. *Ch. 10*

consumer account An account designed for external users to access web or mobile apps without also needing to manage their access to backend resources. *Ch. 7*

container A lightweight, self-contained environment that provides only the essential OS (operating system) services needed to run an embedded application. *Ch. 2*

CRM (customer relationship management) A product that assists a business with managing data from interactions with customers, potential customers, vendors, and associates. *Ch. 1*

cryptojacking The illegal use of someone else's compute resources for cryptomining. *Ch. 6*

cryptomining The typically legitimate process of running calculations to enable blockchain transactions, which earns cryptocurrency. *Ch. 6*

CSA (Cloud Security Alliance) A vendor-neutral nonprofit organization that helps establish and maintain IT industry standards and best practices for cloud security. *Ch. 1*

CSP (cloud service provider) A business that offers one or more cloud services to other businesses or individuals. *Ch. 1*

CVE (Common Vulnerabilities and Exposures) A standardized catalog of known vulnerabilities that aids in tracking and researching these issues. *Ch. 6*

CVSS (Common Vulnerability Scoring System) A standardized framework for rating vulnerabilities on a scale from 0 to 10, with higher scores representing more severe vulnerabilities. *Ch. 6*

CWE (Common Weakness Enumeration) A standardized taxonomy of known weaknesses in software, firmware, hardware, and service components. *Ch. 6*

CWSS (Common Weakness Scoring System) A standardized framework for rating weaknesses in three metric groups: Base Finding (basic characteristics of the weakness), Attack Surface (options for exploiting the weakness), and Environmental (potential impact on the system's environment or other context, such as business impact). *Ch. 6*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

D

DAS (direct attached storage) Storage that is directly attached to the computer or server that uses it. Also called local storage. *Ch. 8*

data ownership The possession of and responsibility for data. *Ch. 8*

data retention The storage of data for a defined period. *Ch. 8*

data sovereignty The right to determine what laws apply to data. *Ch. 8*

data type A category of data that determines what can be done with that data. *Ch. 10*

DDoS (distributed denial of service) attack An attack in which multiple computers simultaneously flood a target with traffic, rendering the target unable to function. Most often, this type of attack is performed by a botnet made up of numerous malware-infected machines without the owners' knowledge. *Ch. 6*

DEB (Debian) artifact A package used to install or update software on Debian-based Linux distributions. *Ch. 10*

dedicated host A configuration option to run a cloud-hosted VM on a physical server that is not shared with other customers. *Ch. 2*

DevOps (development and operations) A cultural shift toward continuous collaboration between development teams and operations teams that brings highly responsive application updates. *Ch.3*

DHCP (Dynamic Host Configuration Protocol) A protocol that manages the dynamic distribution of IP addresses on a network. *Ch. 2*

digital certificate A small file containing verified identification information and the public key of the entity whose identity is being authenticated. *Ch. 7*

DLP (data loss prevention) A security technique that uses software to monitor confidential data, track data access and ownership, and prevent it from being copied or transmitted off the network. *Ch. 6*

DNS (Domain Name System or **Domain Name Service)** A hierarchical approach to tracking domain names and their addresses, devised in the mid-1980s. *Ch. 5*

DoS (denial of service) attack An attack in which a legitimate user is unable to access normal network resources because of an attacker's intervention. Most often, this type of attack is achieved by flooding a system with so many requests for services that it can't respond to any of them. *Ch. 6*

DR (disaster recovery) Strategies and tools designed to facilitate an organization's recovery from an adverse event. *Ch. 2*

durability The resistance of data to errors due to corruption or loss. *Ch. 8*

E

edge computing The placement of computing devices, data, or connections at or near the location of their use. *Ch. 5*

EDR (endpoint detection and response) A service that detects and monitors endpoints in a network or cloud deployment. *Ch. 6*

egress traffic Traffic exiting a network or system. *Ch. 2*

elasticity A cloud resource's ability to expand or contract automatically without requiring human intervention. *Ch. 2*

ELK (Elasticsearch, Logstash, and Kibana) stack A popular combination of services primarily used to explore log data. *Ch. 9*

encryption The use of an algorithm to scramble data into a format that can be read only by reversing the algorithm—that is, by decrypting the data—to keep the information private. *Ch. 6*

encryption key A string of characters combined with a block of data during encryption to obfuscate the data. *Ch. 6*

EOL (end of life) The date past which a product is no longer sold or service subscriptions renewed. *Ch. 10*

EOS (end of support) The date past which support is no longer provided for a product or service. *Ch. 10*

ephemeral storage A storage resource that does not continue to exist on its own after its host resource, such as a container, is shut down. *Ch. 2*

event A detectable change in a monitored cloud resource's state. *Ch. 3*

event correlation The process of identifying relationships between events to identify event patterns. *Ch. 9*

event-driven architecture A design strategy that uses events to trigger communications between microservices. *Ch. 3*

F

FaaS (Function as a Service) The management of individual processes or functions in the cloud rather than managing an entire OS environment. *Ch. 2*

false positive An indication of a problem that isn't actually a problem. *Ch. 9*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

fan-out architecture A pattern of sending a single message from a producer to several consumers. *Ch. 3*

federation The process of managing user identities across organizations in a trust-based relationship. *Ch. 7*

flat file A file that contains plaintext data in a list where data fields and records are separated by a symbol such as a comma, period, space, or hard return. *Ch. 10*

FM (foundational model) A trained dataset used by GenAI to generate content. *Ch. 1*

function A block, or portion, of code that runs when activated. *Ch. 10*

functional testing A testing strategy that confirms a system works as expected and meets predefined requirements. *Ch. 3*

G

GCP (Google Cloud Platform) Google's collection of public cloud computing services, designed to take advantage of Google's own extensive physical infrastructure around the globe. *Ch. 1*

GDPR (General Data Protection Regulation) A European Union law establishing broad-reaching data protection standards for any information that could be tied to a single individual. *Ch. 3*

GenAI (generative AI) Technology that generates content based on training from one or more datasets. *Ch. 1*

Git A free and open-source version control tool. *Ch. 3*

GitHub An online service that allows developers to run Git in the cloud. *Ch. 3*

GitHub Actions A feature of GitHub that supports automated Git workflows. *Ch. 3*

graph database A type of database that stores data according to each entity's relationship with other entities. *Ch. 8*

GraphQL A query language and an API (application programming interface) standard known to return more specific results than REST (representational state transfer) can do. *Ch. 3*

H

HA (high availability) The percentage of time during which a system, resource, or network functions reliably. *Ch. 2*

hashing The transformation of data through an algorithm that is mathematically irreversible and generally reduces the amount of space needed for the data. Hashing is mostly used to ensure data integrity—that is, to verify the data has not been altered. *Ch. 6*

HIPAA (Health Insurance Portability and Accountability Act) A U.S. law that, among other things, establishes data protection standards for medical information. *Ch. 3*

horizontal scaling The process of increasing capacity of a cluster by increasing the number of resources in the cluster. *Ch. 3*

HTTP (Hypertext Transfer Protocol) status code A three-digit code, in the range of 100 to 599, returned from a server to an HTTP client to indicate the status of an HTTP request and some information about the nature of a problem when there is an error. *Ch. 5*

hub-and-spoke topology A VPC (virtual private cloud) peering architecture where a central device, such as AWS's Transit Gateway, or a central network, such as Azure's hub virtual network, connect many VPCs to each other. *Ch. 5*

hybrid cloud A cloud deployment model in which both private and public clouds or both cloud-based and on-prem services are used simultaneously. *Ch. 1*

I

IaaS (Infrastructure as a Service) A cloud service that allows consumers to deploy a cloud-based network with services such as operating systems, network configurations, storage, and virtual devices. *Ch. 1*

IaC (infrastructure as code) The provisioning and management of IT infrastructure through computer-readable configuration files that can be automated rather than interactive tools that require manual configuration. *Ch. 10*

IAM (identity and access management) A framework of techniques and tools for managing the identities of people and applications that allow for access to cloud resources. *Ch. 7*

identity A digital entity that can be given access to resources through roles, policies, and permissions. *Ch. 7*

IDS (intrusion detection system) A physical or virtual device or service that monitors network traffic, generating alerts about suspicious activity. *Ch. 6*

IG or IGW (Internet gateway) A virtual gateway device that provides a connection between internal, private resources and the public Internet. *Ch. 4*

indexing A compute-intensive process in log management that organizes and labels data saved to a log for easier search and analysis. *Ch. 9*

in-place recovery A recovery process that restores backup data to the same resources that hosted the original data. *Ch. 9*

ingress traffic Traffic entering a network or system. *Ch. 2*

instance A specific resource, such as a VM, hosted in a cloud platform. *Ch. 2*

IoT (Internet of Things) The connection of all kinds of devices to the Internet, even types of devices that normally wouldn't be expected to do so, such as lamps, bathroom scales, or doorbells. *Ch. 1*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

IoT (Internet of Things) gateway A physical or virtual device that serves as an intermediary connecting IoT devices to the cloud, similar to a network router. *Ch. 5*

IPS (intrusion prevention system) A physical or virtual device or service that stands in-line between a potential attacker and the targeted network or host where it can prevent malicious traffic from reaching the target. *Ch. 6*

ITIL (IT Infrastructure Library) A collection of publications that define best practices in IT. *Ch. 3*

ITSM (IT service management) A collection of tools and strategies used to help IT teams manage several customer-facing processes to better meet user needs. *Ch. 9*

J

Jenkins A standalone tool that supports automated software development workflows and is compatible with many platforms. *Ch. 3*

K

key-value database A database that contains any number of key-value pairs for each record. *Ch. 8*

knowledge base A collection of information, configurations, troubleshooting scenarios, and helpful tips relevant to a specific network or service. *Ch. 1*

Kubernetes An open-source container management service developed by Google. *Ch. 2*

L

latency The delay between the transmission of a signal and its receipt. *Ch. 1*

legal hold A court-ordered notification requiring that identified data not be deleted. Also called a litigation hold. *Ch. 8*

life cycle rule Storage configurations that enforce the automatic transition of data from one storage tier to another based on various factors, such as age or time since last access. *Ch. 8*

load balancing The distribution of traffic over multiple components or links to optimize performance and fault tolerance. *Ch. 5*

load testing A testing strategy that pushes a system beyond its normal workload conditions to determine how well it adapts to higher demands at the upper end of its anticipated workload. *Ch. 3*

local repository In Git, a structured storage space on a local computer for project files. *Ch. 3*

local storage Storage that is directly attached to the computer or server that uses it. Also called DAS (direct attached storage). *Ch. 8*

local user authentication An authentication model that performs the authentication process on a single machine to access that machine's own resources. *Ch. 7*

log A collection of streams of data over time that show historical insights and reveal usage patterns or anomalies. *Ch. 9*

log aggregation The process of collecting, standardizing, and organizing log data for storage and analysis. *Ch. 9*

log pipeline A logical sequence and structure of individual processes that collect, process, enrich, and route log data from many sources to a central location for storage, analysis, and monitoring. *Ch. 9*

log scrubbing An important part of the logging process that removes sensitive data from logs before storing captured traffic. *Ch. 9*

loosely coupled architecture An infrastructure design where each system operates independently of other systems so that any individual system can be changed or replaced without affecting other systems in the infrastructure. *Ch. 3*

M

malware (malicious software) A program or piece of code designed to intrude upon or harm a system or its resources. *Ch. 6*

managed service A service where the provider handles most of the management and upkeep of the infrastructure for that service, and the customer also retains some control over where their resources run, how they're configured, and how they scale. *Ch. 2*

measured service The ability to charge for cloud resource usage according to an incremented schedule based on the type of service being used. *Ch. 1*

metadata Information about a resource that describes the resource, how it's used and configured, or what it can do. *Ch. 6*

MFA (multifactor authentication) An authentication process that requires information from two or more categories of authentication factors. *Ch. 7*

micro-segmentation A security technique that allows for granular control of traffic and workflows within a cloud-based network. *Ch. 6*

microservice A component of an application that can be run in a separate cloud resource from other application components. *Ch. 2*

ML (machine learning) The ability of computers to learn from data analysis rather than only from specific programming. *Ch. 1*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

multi-cloud A cloud deployment model in which a customer incorporates cloud services from multiple vendors at one time. *Ch. 1*

multitenancy The support of multiple customers' processes on the same hardware. *Ch. 1*

N

NAC (network access control) One or more security techniques for managing users' or resources' access to a network and its resources. *Ch. 6*

NAS (network attached storage) A specialized storage device or group of storage devices that provides centralized, fault-tolerant file system storage and relies on network infrastructure to provide file access. *Ch. 8*

NFV (network functions virtualization) A network architecture that virtualizes network devices and services. *Ch. 4*

NGW (NAT gateway) A virtual gateway device situated in a public subnet to provide resources in a private subnet with egress-only access to the Internet. *Ch. 4*

NIST (National Institute of Standards and Technology) A nonregulatory agency that is part of the U.S. Department of Commerce and promotes measurements, standards, and technology in support of innovation and industrial competition. *Ch. 1*

NLB (network load balancer) A load balancer that uses transport-layer information contained in packets to distribute traffic among systems in a target group. *Ch. 5*

nonrelational database A type of database that can store a wide variety of data types using an unstructured, semi-structured, or structured design, allowing for flexibility, scalability, and high performance to adapt to changing needs. *Ch. 8*

O

OAuth 2.0 (Open Authorization, version 2.0) The industry-standard protocol for token-based authorization processes. *Ch. 7*

obfuscation Techniques that make data difficult or impossible to interpret even if it is discovered or stolen. *Ch. 8*

observability A systematic design that ensures transparency of activity and resource state information in a distributed system to support thorough monitoring and automated responses. *Ch. 9*

on demand A billing model where the customer pays for services used as they're needed with no commitment to future use. Also called pay-as-you-go. *Ch. 2*

on-demand self-service The ability to add, adjust, or remove cloud resources at any time by the service subscriber or other approved users. *Ch. 1*

operator A symbol used to define an operation on variables. *Ch. 10*

OpEx (operational expenditures) Costs whose benefits are experienced in about the same period as when the expenses are paid for those benefits. *Ch. 3*

orchestration The design, development, and optimization of automation processes. *Ch. 1*

organization An identity management resource at the top of a hierarchy of accounts that is used to control access to all of a company's resources in that cloud. *Ch. 7*

orphan resource A resource that was created and then abandoned without being deleted, which often continues to cost money and could present security vulnerabilities. *Ch. 6*

overlay network A virtual network that uses its own logical configurations to direct traffic on the underlying physical infrastructure. *Ch. 4*

P

PaaS (Platform as a Service) An intermediate layer of cloud capability that allows consumers to deploy applications on various platforms without having to manage lower-layer infrastructure. *Ch. 1*

PAM (privileged access management) A subset of IAM (identity and access management) that applies stricter rules and safety precautions specifically to users who are given elevated permissions to do their jobs. *Ch. 7*

parallel recovery A recovery process that restores a system using backup data hosted in a different location. Also called side-by-side recovery. *Ch. 9*

password vault Secure storage for passwords or other credentials and information. *Ch. 7*

pay-as-you-go A billing model where the customer pays for services used as they're needed with no commitment to future use. Also called on demand. *Ch. 2*

PCI DSS (Payment Card Industry Data Security Standard) A data protection standard applying to any company that accesses, stores, processes, or transmits credit card information. *Ch. 3*

peering The joining of VPCs (virtual private clouds) in such a way that traffic is routed via private IP addresses and not over the Internet. *Ch. 4*

pen (penetration) testing A testing strategy that scans a network for vulnerabilities and attempts to take advantage of those gaps to investigate potential security flaws. *Ch. 3*

performance testing A testing strategy that challenges a system to determine its responsiveness and stability under various conditions. *Ch. 3*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

persistent storage A virtual storage device that continues to exist even when its connected instance or container is terminated. *Ch. 2*

phishing A practice in which an attacker attempts to glean access or authentication credentials by posing as someone who needs that information. *Ch. 6*

pipeline A series of processes in the development or automation life cycle of an application or other system. *Ch. 6*

PKI (public key infrastructure) The use of certificate authorities to associate public keys with certain users. *Ch. 7*

placement group In AWS, a logical group of instances all hosted in the same AZ (availability zone) that are used to determine the affinity of these instances. *Ch. 2*

port mapping A configuration that bridges a specific port on a host to a corresponding port on a container. *Ch. 2*

principle of least privilege A security measure that ensures users are only given enough access and privileges to do their jobs, and these privileges are terminated as soon as the person no longer needs them. *Ch. 7*

private cloud A cloud deployment model in which data storage, applications, or services are hosted on hardware dedicated to one organization. *Ch. 1*

private IP address An IP address that can be used on a private network but not on the Internet. IEEE recommends the following IP address ranges for private use: 10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16. *Ch. 4*

private subnet A cloud-based subnet in which hosted resources do not have direct access to the Internet. *Ch. 4*

privilege escalation attack The use of existing privileges to gain information about a system or to further escalate access privileges. *Ch. 7*

privileged container A container that is associated with the host's root user and therefore is given some level of root access to the host. *Ch. 6*

programmatic access An access method that requires access keys for authentication and allows interaction with cloud resources through a command line, SDK (software development kit), API (application programming interface), or a third-party tool. *Ch. 7*

project An effort that has a clearly defined beginning and ending. *Ch. 3*

project management The application of specific skills, tools, and techniques to manage project-related processes in such a way that the desired outcome is achieved. *Ch. 3*

provisioning tool An automation tool that deploys new resources. *Ch. 10*

pub/sub (publish/subscribe) messaging Asynchronous communication between event producers and event consumers. *Ch. 9*

public cloud A cloud deployment model in which data storage, applications, or services are hosted on shared hardware centrally managed by service providers. *Ch. 1*

public IP address An IP address that is valid for use on public networks, such as the Internet. *Ch. 4*

public subnet A cloud-based subnet in which hosted resources are provided a direct route to the Internet. *Ch. 4*

R

RACI (responsible, accountable, consulted, informed) matrix A theoretical framework that identifies role and responsibility assignments for team members or groups. *Ch. 6*

RAID (Redundant Array of Independent Disks or **Redundant Array of Inexpensive Disks)** A distribution of data copies stored on multiple disk drives; provides a level of data security, performance improvement, or both. *Ch. 8*

ransomware Malicious software that encrypts data or locks a device until a ransom is paid. *Ch. 6*

rapid elasticity The ability to increase or decrease the number or size of cloud resources according to demand. *Ch. 1*

RBAC (role-based access control) A method of access control where a network administrator assigns only the privileges and permissions necessary for users to perform the role required by their position in an organization. Also called nondiscretionary access control. *Ch. 7*

region A collection of AZs (availability zones) in one geographic location, usually hosted in several of the cloud provider's physical data centers. *Ch. 2*

regression testing A testing strategy that tests preexisting functionality after a change to confirm that changes haven't negatively impacted other systems. *Ch. 3*

relational database A type of database that must conform to a rigid structure consisting of tables that each holds different kinds of information, which yields insight into relationships between the tables. *Ch. 8*

remote repository In Git, a structured storage space for project files on a remote system. *Ch. 3*

resiliency The ability of a system to withstand and compensate for failures or outages while continuing to function. *Ch. 2*

resource metering A way to measure usage of individual resources, such as network bandwidth or CPU cycles, and enforce a pay-as-you-go pricing model. *Ch. 9*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

resource pooling The availability of physical and virtual cloud resources to multiple subscribers according to consumer demand without regard to geographic location. *Ch. 1*

REST (representational state transfer) A popular, simple, flexible, and scalable API (application programming interface) protocol that relies on a client-server structure and can incorporate caching on either the server side or the client side. *Ch. 3*

RFC (Request for Comments) 1918 An IETF (Internet Engineering Task Force) publication specifying best practices for address allocation on private networks. *Ch. 4*

RI (reserved instance) Also called a reserved resource, a billing model where the customer pays a lower rate for services based on a commitment to future use of the service. *Ch. 2*

rightsizing The process of finding an optimal balance for a workload or specific resources to meet performance demands while keeping within budgetary or other restrictions. *Ch. 2*

risk register A project management tool that can be used to document cybersecurity and other types of risks, background information on these risks, mitigation strategies specific to each risk, plans for responding should a vulnerability be attacked, and team members with responsibilities related to each risk. *Ch. 6*

role A resource that represents a collection of permissions, which can be assigned to an identity to grant short-term access to resources. *Ch. 7*

root The master user in a cloud account or compute system that retains complete access and action permissions, no matter what permissions are also given to other users. *Ch. 7*

root-cause analysis A formal procedure that identifies all the contributing factors that allowed an incident to happen. *Ch. 3*

route table A list of routes used to manage traffic to and from cloud-based resources. *Ch. 4*

RPC (remote procedure call) An older API (application programming interface) type that is primarily designed to perform remote functions on servers; RPC is less structured or standardized than REST (representational state transfer) or SOAP (Simple Object Access Protocol) and supports fewer types of requests. *Ch. 3*

RPM (Red Hat Package Manager) artifact A package used to install or update software on Red Hat–based Linux distributions. *Ch. 10*

runbook A collection of physical or digital documents that outline the precise steps to complete procedures or operations in a specific IT environment. *Ch. 10*

S

SaaS (Software as a Service) The provision of application services through the cloud where those applications can be accessed from many types of devices without the customer having to manage any of the underlying infrastructure. *Ch. 1*

SAN (storage area network) A distinct and specialized network of storage devices and the infrastructure that supports them in providing centralized, fault-tolerant block storage. *Ch. 8*

scaling A process that increases or decreases the capacity of a resource or group of resources, often automatically. *Ch. 1*

scanning scope A statement that defines the boundaries of a scan, including a list of the scan's targets, objectives, and depth. *Ch. 6*

scope exhaustion The problem of running out of available IP addresses in a defined CIDR (classless interdomain routing) block. *Ch. 4*

screened subnet An area on the perimeter of a network that is less protected than the internal network and provides internal resources with a transitive connection to external resources. Previously called DMZ (demilitarized zone). *Ch. 6*

script A document defining a simple, static set of instructions for a series of tasks. *Ch. 10*

SDK (software development kit) A platform-specific collection of tools and resources developers use to create, deploy, and troubleshoot their applications. *Ch. 7*

SDN (software-defined networking) A centralized approach to networking that removes most of the decision-making power from network devices and instead handles that responsibility at a software level. *Ch. 4*

secrets management A system that handles the secure storage and distribution of keys, passwords, authorization tokens, and other files used to secure access to resources. *Ch. 7*

sensor A device that detects and monitors metrics related to its environment. *Ch. 1*

sentiment analysis A service that evaluates text to extract information about the emotional tone of the text. *Ch. 1*

serverless service A cloud-native, streamlined technology where a customer's cloud-hosted resource runs on a server managed by the cloud provider. *Ch. 1*

service account An account designed for a resource to use. *Ch. 7*

service discovery The process of finding and monitoring services and their health throughout a network. *Ch. 3*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

service quota A maximum number of resources or a maximum value for a feature that each customer can deploy within a single account. *Ch. 9*

shared responsibility model The division of labor between a cloud service provider and a cloud customer for the management, configuration, and security of a cloud service. *Ch. 6*

SLA (service-level agreement) A contract defining service standards from a company to its customer and options for recourse should those service levels not be met. *Ch. 1*

SOAP (Simple Object Access Protocol) An API (application programming interface) protocol that is more structured and formalized than REST (representational state transfer), tends to be slower, and is more difficult to set up. *Ch. 3*

SOC (System and Organization Controls) report A report produced by an audit performed according to the standards defined by SSAE 18. *Ch. 3*

social engineering The act of manipulating social relationships to circumvent network security measures and gain access to a system. *Ch. 6*

spot instance A billing model where a resource's price fluctuates based on supply and demand, and the customer sets a maximum threshold for cost of a resource. The customer is given use of the resource only when its cost falls below that threshold. *Ch. 2*

SSO (single sign-on) A form of authentication in which a client signs in one time to access multiple systems or resources. *Ch. 7*

storage tier A pricing schema that places data in various storage technologies where storage costs decrease as access costs increase. Also called storage class. *Ch. 8*

stratum A number that indicates an NTP (Network Time Protocol) server's location within the NTP hierarchy relative to a stratum-1 server. *Ch. 5*

subnetting The segmentation of networks into smaller networks called subnets, and the process of calculating address spaces for these subnets within the larger network. *Ch. 4*

synthetic full backup A backup created from the most recent full backup and any applicable incremental or differential backups. *Ch. 8*

T

tag A key-value pair used to track resources, especially in a cloud platform. *Ch. 9*

tar (tape archive) file A collection of files bundled into a single file for easy transport, but the files are not compressed. Also called a tarball. *Ch. 10*

template 1) A configuration file that determines specific parameters for deployment of identical resources. *Ch. 2.* 2) A document or collection of documents defining a specific outcome of automated actions, such as detailed configurations for all of an environment's resources. *Ch. 10*

tenant A cloud customer whose cloud services are running on specific hardware. *Ch. 1*

Terraform A cloud-agnostic automation tool for infrastructure provisioning. *Ch. 10*

text recognition A service that can interpret text in documents, prompts, transcripts, images, handwriting, and other media. Also known as OCR (optical character recognition). *Ch. 1*

text translation A service that translates text from one language to another. *Ch. 1*

text-to-voice A service that converts text to human-like speech. *Ch. 1*

token-based authentication A model that blends authentication and authorization; the system relies on generated tokens to validate a user's identity to services without requiring repeated authentication. *Ch. 7*

tokenization A data obfuscation technique that replaces sensitive data with random data called a token. *Ch. 8*

tracing The process of tracking a request's end-to-end progress across a distributed system, which might involve many services. *Ch. 9*

transit gateway A virtual device at the center of a hub-and-spoke topology that acts as a cloud router directing traffic across all connected networks. *Ch. 5*

transitory peering The ability of communications to cross multiple peering connections. *Ch. 5*

triage A process of prioritizing and organizing alerts in a way that ensures the most critical issues are addressed first. *Ch. 9*

trunk port A switch port used to connect a switch to a router or another switch, which are connections that must be able to handle traffic for multiple VLANs (virtual local area networks). *Ch. 5*

U

unprivileged container A container that is assigned a non-root user on the host and therefore does not have root access to the host. *Ch. 6*

usability testing A testing strategy that asks users to test the system while developers observe the users and the system. *Ch. 3*

user account An account designed for a human. *Ch. 7*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

V

variable A logical container for information stored in code for later use. *Ch. 10*

vCPU (virtual CPU) One logical thread of processing power allotted to a VM. *Ch. 2*

VDI (virtual desktop infrastructure) A remote desktop implementation that offers VM instances to remote access clients. *Ch. 2*

vector database A type of database that stores collections of values, called vector embeddings, in a way that can be searched quickly for values that are similar. *Ch. 8*

vendor lock-in The ominous expense of changing vendors that often prevents customers from considering competitors' products. *Ch. 3*

versioning The technique of keeping older versions of the data as a backup or to document changes in the data, such as for compliance or regulatory purposes. *Ch. 8*

vertical scaling The process of increasing the capacity of an existing resource by adding more components (such as adding more memory) or upgrading its components (such as replacing its CPU with a more powerful one). *Ch. 3*

vGPU (virtual GPU) A partial allocation of available GPU resources. *Ch. 2*

virtual patch A collection of rules designed to close a security gap from a recently discovered vulnerability. *Ch. 10*

virtualization The emulation of part or all of a computer or network. *Ch. 1*

visual recognition A service that analyzes high volumes of images and videos to identify desired information. *Ch. 1*

VM (virtual machine) A logical computer running its own operating system while borrowing hardware resources from its physical host computer. *Ch. 2*

VNet (virtual network) A software-defined portion of a larger, cloud-based network. *Ch. 4*

vNIC (virtual NIC) A logically defined network interface associated with a VM. *Ch. 2*

voice-to-text A service that interprets speech into text. *Ch. 1*

volume-based licensing Software licensing designed to allow many users under one license. *Ch. 2*

VPC (virtual private cloud) A software-defined portion of a larger, cloud-based network. *Ch. 4*

VPN (virtual private network) A virtual connection over the Internet between two geographically distant locations that provides access to network resources between a client and a remote network, two remote networks, or two remote hosts. *Ch. 5*

vRAM (virtual RAM) The amount of the host's physical memory reserved for a VM's use. *Ch. 2*

vulnerability management The process of identifying, assessing, and addressing security vulnerabilities in a system. *Ch. 6*

vulnerability testing A testing strategy that identifies a system's security weaknesses that could be compromised with or without malicious intent. *Ch. 3*

W

WAF (web application firewall) A firewall service that can be installed either on-prem or in the cloud and protects web apps by inspecting, monitoring, and filtering traffic coming into the web app. *Ch. 6*

web service A standardized method of offering network-based communications between software systems. *Ch. 3*

WebSocket An API type that is intended for interactive two-way communication over a stateful connection. *Ch. 3*

workflow A series of steps, tasks, or operations that can be automated, scheduled, and monitored. *Ch. 10*

workforce identity An internal user account for employees or contractors who do not require administrative access to cloud-hosted resources. *Ch. 7*

Z

zero trust A security model where everything in the network is considered untrustworthy until proven otherwise. *Ch. 6*

ZIP file A package of compressed files that contains a directory for accessing individual files within the package. *Ch. 10*

zombie resource An abandoned resource that is taken over by an attacker and used for their own purposes. Also called a zombie instance, such as a VM instance that is taken over. *Ch. 6*

ZTA (zero trust architecture) The implementation specific configurations and tools that support a zero trust security model. *Ch. 6*

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Index

A

access control list (ACL), 343
access control methods, 296
access key, 252, 292
access management, 352
access port, 202, 203
account hijacking, 225
account management, 225, 283–284
account types
  consumer accounts, 280–281
  organizations, 279–280
  workforce identities, 281–282
Active Directory (AD), 284, 299
actuator, 25
Advanced Message Queuing Protocol (AMQP), 196
advanced persistent threats (APTs), 225
affinity, 60–62
agent, 377
AI. *See* artificial intelligence (AI)
Alibaba Cloud, 17
Amazon Q AI-powered assistant, 281, 282
Amazon Q, AWS Management Console, 23–24
Amazon Web Services (AWS), 4, 17
  account with, 33–36
  Artifact service, 91
  auto scaling, 109, 110, 119–120
  availability zones in, 144
  CLI, 425–427
  cloud monitor, 406–408
  cloud peering in, 217–218
  configure security in, 266–267
  configure VPC and subnets in, 160–162
  cost budget in, 107
  deploy a VM in, 74–75
  Direct Connect, 174
  Directory Service, 304
  Elastic Block Store, 340
  Elastic Container Registry service, 67
  Elastic Container Service, 67–68
  Elastic Kubernetes Service, 67
  explore VPN options in, 208–212
  gateways and route tables in, 148–151
  hub-and-spoke VPC peering architecture, 178
  IAM, 297–299, 304–305
  IoT services, 195
  KMS and secrets manager, 313–314
  manage storage in, 360–362
  monitoring in, 390–395
  patching in, 438–439
  pricing calculator, 88
  regions in, 143–144
  RPO, RTO, and cost, 113
  security group, 129, 130
  serverless services, 63, 67
  storage services, 338–341
  subnets in, 139, 147, 238–243
  using code in, 447–448
  VM instance types for, 54
  VPCs in, 145–146, 238–243
Ansible tool, 424
anti-affinity rule, 61
API-based communication, 285
API throttling, 400
append blob, 342
application gateway, 173
application load balancer (ALB), 188
application programming interface (API), 10, 224
  calls, 96
applications, cloud storage, 323
application security, 252
application security group (ASG), 245–246
artifact, 431
artificial intelligence (AI), 21–24
  in IT, 5, 49, 110, 134–135, 196, 234, 340, 382, 442
audit trails, 388
authentication, 284
  backups, 350
  certificate-based, 290–292
  defined, 284
  issues, 307
  multifactor authentication, 288–290

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

authentication, (*Continued*)
password policies, 286–288
processes, 284–285
programmatic access, 292–294
securing secrets, 294–295
authentication, authorization, and accounting (AAA), 284
authorization, 295
AWS IAM, 297–299
Azure IAM, 299–300
GCP IAM, 300–301
processes, 295–297
authorization issues, 307
automating code pipelines
CI/CD pipelines, 430–432
scripting logic, 433–434
automation, 8, 418–419
CLIs, 425–429
code pipelines, 430–434
comparing tools, 425
configuration as code, 423–424
infrastructure as code, 419–423
troubleshooting automation issues, 441–443
auto scaling, 16, 109
availability zones (AZs), 61, 144
in AWS, 144
in Azure, 152
AWS. *See* Amazon Web Services (AWS)
AZs. *See* availability zones (AZs)
Azure, 4, 17
account with, 36–38
CLI, 427–428
cloud monitoring, 408–409
cloud peering in, 217–218
configure a VNet and subnets in, 162–164
configure security in, 268–270
deploy a VM in, 75–77
explore VPN options in, 212–215
hub-and-spoke VPC peering architecture, 179
IAM, 299–300, 305
Key Vault, 314–315
manage storage in, 363–365
monitoring in, 395–396
patching in, 439
regions and availability zones in, 152
route tables in, 154
serverless services, 63
storage services, 341–343
subnets in, 152–153
using code in, 448–449
virtual machine scale set in, 120–121
virtual networks, 243–246
VM instance types for, 54
VNet in, 142, 152–153
Azure Active Directory (Azure AD), 299
Azure Blob storage, 342

B

backup rotation, 350
backups, creating and storing
considerations, 350
retention, 347–348
strategies, 346–347
types, 348–349
bare-metal hypervisor, 46
bastion host, 150, 151
BCP. *See* business continuity planning (BCP)
big data, 324
billing models, 56–58
binary large objects (BLOBs), 342
blast radius, 283
block blobs, 342
block storage, 328
Border Gateway Protocol (BGP), 185
bridged mode, 51
broad network access, 6
budgeting for cloud, 87–89
business continuity (BC), 111
business continuity planning, 110–112

C

California Consumer Privacy Act (CCPA), 91
call tree, 111
capacity planning, 108–109
capital expenditures (CapEx), 87
CASB. *See* cloud access security broker (CASB)
Center for Internet Security (CIS), 233
central processing unit (CPU), 46, 48, 49
certificate-based authentication, 290–292
certificate management, 291–292

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Certified Cloud Security Professional (CCSP) certification, 4
Certified Information Systems Security Professional (CISSP) certification, 4
chain of custody, 114
chargeback model, 377
ChatGPT, 22, 24
chief information security officer (CISO), 232
ciphertext, 257
CIS benchmarks, 233
classless interdomain routing (CIDR) notation, 137, 139, 140
CLI. *See* command-line interface (CLI)
client access license (CAL), 48
clone, 54
cloning, 348
cloning VM instances, 54–56
cloud access security broker (CASB), 238, 250
cloud accounts, 278
  account management, 283–284
  account types, 279–282
  identity, 278–279
cloud architecture, 96
cloud bursting, 11
cloud computing
  certifications, 32–33
  characteristics of, 2–8
  description of, 5–7
cloud deployment models, 452–454
  community cloud, 13
  hybrid cloud, 11–12
  multi-cloud, 12
  private cloud, 10
  public cloud, 9–10
cloud IAM troubleshooting, 307–309
cloud identity, 305
Cloud Interconnect, 174
cloud maintenance
  decommissioning, 440
  patching and update methodologies, 437–440
  tasks, 435
  updates, types of, 436–437
cloud migration
  budgeting for cloud, 87–89
  project management, 92–93
  regulations and compliance, 89–92
  strategies, 86–87
  types, 85–86
cloud-native design
  building with microservices, 94–95
  system integration, 96–97
cloud network
  deployment, 165–166
  interfaces, 140–142
cloud peering, 178
cloud security
  monitoring for suspicious activities, 227–228
  shared responsibility for cloud security, 231–233
  threats to, 224–227
  troubleshooting, 260–262
  vulnerability management, 229–231
Cloud Security Alliance (CSA), 4, 91
cloud service models, 13–16
cloud service providers (CSPs), 4, 17, 54, 91, 227
  AI/MI, 21–24
  categories of, 18–21
  IOT, 25
  outage, 106
  PaaS and IaaS, 17–18
cloud service types, 40–41
cloud stack
  from OSI model to, 130–133
  theoretical model, 132–133
CloudStack, 18
cloud storage and database
  AWS storage services, 338–341
  Azure storage services, 341–343
  block storage, 328
  choosing between file, block, and object storage, 330
  data types, 322–323
  file system storage, 327–328
  GCP storage services, 343–344
  object storage, 328–329
  on-prem storage technologies, 324–326
  types comparison, 326
  types of databases, 330–333
cloud system design
  auto scaling types, 109–110
  business continuity planning, 110–112
  capacity planning, 108–109

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

cloud migration, 84–93
cloud-native design, 94–97
deployment issues, 105–107
disaster recovery, 112–115
sizing issues, 107–108
system development, 98–104
cloud-to-cloud migrations, 85–86
cloud–to–on-premises, 85
cloud WAN, 179–180
CloudWatch
alarm configuration, 392–394
events and rules, 395
cluster, 61
cluster across boxes (CAB), 60
cluster placement group, 61
colocation facility, 174
command-line interface (CLI)
AWS, 425–427
Azure, 427–428
GCP, 428–429
commit, 100
Common Vulnerabilities and Exposures (CVE), 229
Common Vulnerability Scoring System (CVSS), 231
Common Weakness Enumeration (CWE), 230
Common Weakness Scoring System (CWSS), 231
community cloud, 13
compliance auditing, 352
compression, 348
CompTIA Cloud + Certification, 3–4
compute security
application security, 252
endpoint protection, 250–251
pipeline security, 252–253
Computing Technology Industry Association (CompTIA), 3
conditional statement, 433
confidentiality, integrity, and availability (CIA) triad, 256
configuration as code (CaC), 423–424
configuration management (CM) tools, 423
connecting cloud networks
cloud peering, 178
cloud WAN, 179–180
transit gateways, 178–179
connecting networks, 172
direct connection, 174
network throughput capacity on direct connections, 174
segment direct connections, 175–176
virtual private networks, 172–173
Constrained Application Protocol (CoAP), 196
consumer accounts, 280–281
containers, 65–66
AWS storage services, 340–341
networking, 69
supporting, 66–69
troubleshoot applications, 69–70
container security, 251
content delivery network (CDN), 193
continuous delivery (CD), 430
continuous deployment (CD), 430
continuous integration (CI), 430
cost considerations, 60
CPU. *See* central processing unit (CPU)
create, read, update, and delete (CRUD), 97
cryptojacking, 225
cryptomining, 225
CSPs. *See* cloud service providers (CSPs)
customer relationship management (CRM), 12, 322

D

DAS. *See* direct attached storage (DAS)
dashboards, 380–382
database, cloud storage, 323
database migration, 337
databases
management types, 333
nonrelational, 331–332
relational, 330–331
data breaches, 224, 228–229
data classification, storage security, 353
data destruction, 337
Data Distribution Service (DDS), 196
data loss prevention (DLP), 238, 352
data obfuscation, storage security, 353–354
data ownership, 354
data retention, 355
data security
encrypting data, 254–257
key management, 257–259

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

data sovereignty, 354
data types, 433
Debian (DEB) artifact, 432
decommissioning, 440
dedicated host, 58
denial of service (DoS) attack, 224
development and operations (DevOps), 98–99
DHCP. *See* Dynamic Host Configuration Protocol (DHCP)
digital certificate, 290
direct attached storage (DAS), 324
direct connection, 174
  network throughput capacity, 174
  segment, 175–176
direct connect routing, 186
disaster recovery (DR), 59–60, 112–115
Disaster Recovery as a Service (DRaaS), 113
discretionary access control (DAC), 296
distributed denial of service (DDoS) attack, 224
Domain Name System/Domain Name Service (DNS), 184–185
DR. *See* disaster recovery (DR)
durability, 327
Dynamic Host Configuration Protocol (DHCP), 51, 180–181
  across hybrid cloud, 183
  in cloud, 181–183

E

EC2. *See* Elastic Compute Cloud (EC2)
EC2 Instance Storage, 340
EC2 template, 56, 57
edge computing, 191
  case for, 191–192
  CDN, 193
editing code, 101
egress traffic, 60
Elastic Block Store (EBS), 56, 328
Elastic Compute Cloud (EC2), 54
Elastic Container Registry (ECR) service, 63
Elastic Container Service (ECS), 67–68
elasticity, 51, 376
Elastic Kubernetes Service (EKS), 67
Elasticsearch, Logstash, and Kibana (ELK) stack, 388
Elastic Utility Computing Architecture for Linking Your Programs To Useful Systems (Eucalyptus), 18
encryption, 254, 255, 351
  data techniques, 255–256
encryption key, 252
end of life (EOL), 440
end of support (EOS), 440
endpoint detection and response (EDR), 251
endpoint protection, 250–251
energy savings, 60
Enhanced Interior Gateway Routing Protocol (EIGRP), 185
ephemeral storage, 66
event correlation, 383
event-driven architecture, 94, 391, 392
events, 94, 383–384
Extensible Messaging and Presence Protocol (XMPP), 196

F

failback, 114
failover, 114
failover cluster, 436
false positives, 383
fan-out architecture, 95
federated identity, 305
federation, 303
file system storage, 327–328
  AWS storage services, 341
  Azure storage services, 343
  GCP storage services, 344
Firewall as a Service (FWaaS), 13–14, 238
five-second rule, 380
flat file, 432
flat file database, 323
flow logs, 240–241
foundational model (FM), 23
fully managed/provider-managed databases, 16, 333
function, 433
functional testing, 103
Function as a Service (FaaS), 63
FWaaS. *See* Firewall as a Service (FWaaS)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

G

gateways and route tables, AWS, 148–151
applying routes, 149–150
Internet gateway, 148–149
NAT gateway, 149
remote access to cloud-hosted servers, 150–151
GCP. *See* Google Cloud Platform (GCP)
General Data Protection Regulation (GDPR), 90
generating access keys, 292
generative AI (GenAI), 22–24
Git, 99
GitHub, 99
GitHub Actions, 103
GitHub repository, 100, 123–124
git status, 101
global server load balancing (GSLB), 190
Google Cloud Platform (GCP), 4–5, 17
account with, 38–40
CLI, 428–429
cloud monitoring, 410–411
cloud peering in, 217–218
configure a VPC in, 164–165
configure security in, 270–272
deploy a VM in, 77–78
explore VPN options in, 215–217
IAM, 300, 301, 306
managed instance group, 121–122
manage storage in, 365–366
monitoring in, 396–399
patching in, 440
regions and zones in, 155–156
routes in, 157
Secret Manager, 315–316
serverless services, 63
storage services, 343–344
subnets in, 155–156
using code in, 450–452
VM instance types for, 54
VPCs, 156–157, 246–248
graph databases, 331–332
graphics processing unit (GPU), 49–50, 400
GraphQL, 97
group-based access control (GBAC), 296–297
GuardDuty, 227
guest, 46

H

hardened baseline, 251
hashing, 256
Health Insurance Portability and Accountability Act (HIPAA), 90
high availability (HA), 59, 348
HIPAA. *See* Health Insurance Portability and Accountability Act (HIPAA)
horizontal scaling, 109
host, 46
host-only mode, 52
HTTP status codes, 200–201
hub-and-spoke topology, 178, 179
human error, 225
hybrid cloud, 11–12
AWS IAM in, 304–305
Azure IAM in, 305
DHCP across, 183
DNS across, 185
GCP IAM in, 306
IAM for, 302–306
load balancing across, 190
Single Sign-On (SSO), 302–304
hybrid cloud and multi-cloud networking
connecting networks, 172–177
use cases, 171–172
hybrid cloud storage, 337
Hypertext Transfer Protocol (HTTP), 244, 256
Hyper-V, 46
hypervisor, 46, 47
type 1, 46
type 2, 46

I

IaaS. *See* Infrastructure as a Service (IaaS)
IaC. *See* infrastructure as code (IaC)
IAM. *See* identity and access management (IAM)
IBM Cloud, 17–18
identity, 278–279
identity and access management (IAM), 278
AWS, 297–299
Azure, 299–300
GCP, 300–301
for hybrid clouds, 302–306
policies, 316–317
troubleshooting cloud, 307–309

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

identity providers (IdPs), 302, 303
IG/IGW. *See* Internet gateway (IG/IGW)
indexing, 385
Infrastructure as a Service (IaaS), 14–15
  cloud service providers, 17–18
infrastructure as code (IaC), 419–422
  benefits, 422
  tools, 422–423
ingress traffic, 60
in-place recovery, 402
input/output operations per second (IOPS), 336
insider threats, 225
instances, 54
Internet Engineering Task Force (IETF), 256
Internet gateway (IG/IGW), 148
Internet of Things (IoT), 25
Internet Protocol Security (IPsec), 256
Internet service providers (ISPs) outage, 106
inter-VLAN routing, 175
intrusion detection system (IDS), 243
intrusion prevention system (IPS), 243
IoT. *See* Internet of Things (IoT)
IoT cloud networking, 193–196
IoT gateway, 194
IP addressing, 136–137
IP address management (IPAM), 183
IP address spaces
  cloud network interfaces, 140–142
  IP addressing, 136–137
  subnetting, 137–140
ISO/IEC 27001 standard, 9–10, 90
IT Infrastructure Library (ITIL), 92
IT service management (ITSM), 380

J

JavaScript Object Notation (JSON), 297, 298
Jenkins, 103

K

Kernel-based Virtual Machine (KVM), 46
key, 257
key management system (KMS), 259
key performance indicators (KPIs), 374
key-value databases, 331
knowledge base, 28
Kubernetes, 67

L

LastPass, 228
latency, 26, 176–177, 336
legal hold, 347
licensing outage, 106
life cycle rules, 335
Lightweight Directory Access Protocol (LDAP), 285
load balancing, 186–187
  across hybrid cloud, 190
  algorithms, 187
  in cloud, 187–189
  issues, 199–200
  sticky sessions, 187
load testing, 104
locally redundant storage, 341
local repository, 100
local storage, 324
local user authentication, 292
log aggregation, 387
log analytics, 386
log file, cloud storage, 323
log pipeline, 386–387
logs, 384
  analysis, 387–388
  audit trails, 388
  explore, 385–386
  pipeline, 386–387
  types of, 384
log scrubbing, 385
long-range WAN (LoRaWAN), 196
loosely coupled architecture, 95

M

machine learning (ML), 21–24, 378
malicious software (malware), 224
managed service, 16, 68
mandatory access control (MAC), 296
measured service, 6
memory utilization, 401
Message Queuing Telemetry Transport (MQTT), 195
metadata, 226
micro-segmentation, 236
microservices, 66
misconfigured routes, 201–202
ML. *See* machine learning (ML)

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

monitoring resources
in AWS, 390–395
in Azure, 395–396
capturing network traffic, 379
dashboards, 380–382
data collection and response, 390
events, 383–384
logs, 384–388
monitoring in, 396–399
resource metering, 377
tagging cloud resources, 377–378
targets to monitor, 374–377
traces, 388–389
troubleshooting problems and performance, 399–402
user experience, 380
multi-cloud model, 12
multi-cloud networking. *See* hybrid cloud and multi-cloud networking
multifactor authentication (MFA), 288–290
multitenancy, 6

N

NACL. *See* network access control list (NACL)
NAT gateway (NGW), 148, 149, 198
National Institute of Standards and Technology (NIST), 5
network access control (NAC), 238
network access control list (NACL), 237–239
network address translation (NAT), 51–52, 148
network and security settings, 56
network attached storage (NAS), 324
network connection types, 51–52
network firewalls, 242–243
network functions virtualization (NFV), 129
networking concepts, in cloud, 129–130
OSI model to cloud stack, 130–133
software-defined networking in, 133–134
networking containers, 69
network interface controllers (NICs), 51
network load balancer (NLB), 188
network redundancy, 348
network security, 352
network security group (NSG), 244–245
network services, extending
configure network services, 218–219
DHCP, 180–183
DNS, 184–185
load balancing, 186–189
routing, 185–186
Network Time Protocol (NTP), 202
network traffic, 379
network virtual appliance (NVA), 243–246
NGW. *See* NAT gateway (NGW)
nondisclosure agreement (NDA), 90
nonrelational databases, 331–332
graph databases, 331–332
key-value databases, 331
vector databases, 332
nonvolatile memory express over fabrics (NVMe-oF), 326
not-only SQL (NoSQL) database, 331
NVA. *See* network virtual appliance (NVA)

O

obfuscation, 353
object storage, 328–329
AWS storage services, 338–339
Azure storage services, 342–343
GCP storage services, 343, 344
observability, 374
office and media files, cloud storage, 322
on demand, 56–57
on-demand self-service, 5–6
Online Certificate Status Protocol (OCSP), 291
on-premises–to–cloud, 85
on-prem storage technologies, 324–326
Open Authorization, version 2.0 (OAuth 2.0), 303
OpenID Connect, 302
OpenShift, 18
Open Shortest Path First (OSPF), 185
OpenStack, 18
operational expenditures (OpEx), 87
operational level agreement (OLA), 111
operators, 433
optical character recognition (OCR) services, 21
Oracle Cloud, 18
orchestration, 8
organizations, 279–280
orphan resources, 225
overlay network, 132

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

P

PaaS. *See* Platform as a Service (PaaS)
packet broker, 379
Packet Mirroring service, 248
page blobs, 343
parallel recovery, 402
ParkMyCloud, 394
partition placement group, 61, 62
password policies, 286–288
password vault, 287
patching
  in AWS, 438–439
  in Azure, 439
  in GCP, 440
patching/update methodologies, 437–440
pay-as-you-go, 56
Payment Card Industry Data Security Standard (PCI DSS), 90
peering, 136, 137
peering connection, 149, 150
penetration testing (pen testing), 104
performance testing, 103
permission, 297
persistent storage, 66
phishing, 225
pipeline, 252
pipeline security, 252–253
placement groups, 61
Platform as a Service (PaaS), 14
  cloud service providers, 17–18
  security concerns, 14
platform integration, 106
playbook, 114
pluggable authentication module (PAM), 302
port mapping, 69
practice subnetting, 138–140
principle of least privilege, 283
private cloud, 10
private IP addresses, 135
private key encryption, 258
private subnets, 147
privileged access management (PAM), 283–284
privileged container, 251
privilege escalation attack, 309
programmatic access, 292–294
project, 92
project management, 92–93
provisioning tools, 423
public cloud, 9–10
public IP addresses, 135
public key encryption, 258
public key infrastructure (PKI), 290
public subnets, 147
pub/sub messaging, 391
pull model, 425
pull requests, 101–103
push model, 425
Python code, working with, 434

Q

Quality of Service (QoS), 198
quotas, 399–400

R

random access memory (RAM), 46
ransomware, 224
rapid elasticity, 6
Real-Time Messaging Protocol (RTMP), 196
Real-Time Streaming Protocol (RTSP), 196
recoverability testing, backup, 350
recovery options, backup, 350
recovery point objective (RPO), 112, 113, 349
recovery time objective (RTO), 112, 113, 349
Red Hat Package Manager (RPM) artifact, 432
Redundant Array of Independent Disks or Redundant Array of Inexpensive Disks (RAID), 324
regions, 61, 143–144
  in Azure, 152
  in GCP, 155–156
regression testing, 104
regulations for storage security, 354–355
relational databases, 330–331
remote access technology, 6
remote procedure call (RPC), 96
remote repository, 100
representational state transfer (REST), 96, 285
Request for Comments (RFC) 1918, 136
requests for proposal (RFPs), 93
research billing alarms and alerts, 411–413

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

research database, 344–345
reserved instance (RI), 57
Reserved Instance Marketplace, 57
resiliency, 53
resource allocation factors, 59–60
resource configurations, backup, 350
resource contention, 105–106
resource metering, 377
resource pooling, 6
responsible, accountable, consulted, informed (RACI) matrix, 232
RESTful API, 96
rightsizing, 59
risk register, 229
role, 296
role-based access control (RBAC), 296
root, 278
root cause analysis, 114
routes, misconfigured, 201–202
route table, 133, 148–150, 154
routing, 185
  direct connect, 186
  VPN, 185–186
Routing Information Protocol (RIP), 185
runbook, 419–420

S

SaaS. *See* Software as a Service (SaaS)
scalability, 376
scaling, 16
scanning scope, 229
scope, defined, 300
scope exhaustion, 136
screened subnet, 243
script, 421
scripting logic, 433–434
SDN. *See* software-defined networking (SDN)
secrets management, 294–295
secure access service edge (SASE), 237, 238
Secure Sockets Layer (SSL), 256–257
secure web gateway (SWG), 238
securing hybrid and multi-clouds, 248–249
securing protocols, 256–257
security allocation factor, 59
Security Assertion Markup Language (SAML), 302
security, backup, 350
security configurations
  monitoring for suspicious activities, 227–228
  shared responsibility for cloud security, 231–233
  threats to, 224–227
  vulnerability management, 229–231
security groups, 238–239
Security Hub, 227
security incident responses, 228
Security Information and Event Management (SIEM), 379
security issues, cloud, 260
security orchestration, automation, and response (SOAR) workflows, 227
Security, Trust, Assurance, and Risk (STAR) program, 91
segment direct connections, 175–176
self-managed cloud services, 16
self-managed databases, 333
sensor, 25
sentiment analysis, 21
serverless computing, 63–65
serverless database, 333
serverless service, 16
service accounts, 278
service discovery, 95
service level agreement (SLA), 10, 227, 350
service providers (SPs), 302
service quota, 399
shared responsibility for cloud security, 231–233
shared responsibility model, 231
showback model, 377
Simple Notification Service (SNS) notifications, 391–392
Simple Object Access Protocol (SOAP), 96
Simple Storage Service (S3), 338–339
single sign-on (SSO), 302–304
social engineering, 225
Software as a Service (SaaS), 14, 15
software-defined networking (SDN), 129, 133, 238
software-defined wide area networking (SD-WAN), 238
software development kit (SDK), 292
software vulnerabilities, 225
source composition analysis (SCA), 252

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

spot instance, 58
spread placement group, 61, 62
Statement on Standards for Attestation Engagements No. 21 (SSAE 21), 90
static application security testing (SAST), 252
static website, 366–368
storage area network (SAN), 324–326
storage configuration
  data destruction, 337
  storage migration, 336–337
  storage performance, 335–336
  tiering, 334–335
storage metrics, 401
storage migration, 336–337
storage performance, 335–336
storage replication, 348
storage security, 351
  data classification, 353
  data obfuscation, 353–354
  regulations for, 354–355
  techniques, 351–352
storage tiers, 334
stratum, 202
subnets
  in AWS, 147, 160–162, 238–243
  in Azure, 152–153, 162–164
  in GCP, 156–157
subnetting
  IP addressing, 137–140
  practice, 138–140
supporting containers, 66–69
synchronized identity, 305
synthetic full backup, 349
System and Organization Controls (SOC) reports, 90
system development
  deployment strategies, 103
  DevOps, 98–99
  testing types, 103–104
  version control, 99–103
system integration, 96–97

T

tags, 377–378
tape archive (tar) file, 432
template, 55, 421
  misconfiguration, 106
tenants, 6
Terraform, 423
test access point (TAP), 379
text recognition, 21
text-to-voice, 22
text translation, 21
throttling, 400
tiering, 334–335
time synchronization, 106
TLS. *See* Transport Layer Security (TLS)
token-based authentication, 303–304
tokenization, 353
total cost of ownership/total cost of operations (TCO), 88
tracing, 388
traditional network models, 131–132
traffic mirroring, 242
transit gateway, 178–179
transitory peering, 178
Transport Layer Security (TLS), 256–257
triage, 384
troubleshoot applications in containers, 69–70
troubleshooting cloud connectivity
  HTTP status codes, 200–201
  load-balancing issues, 199–200
  misconfigured routes, 201–202
  NAT issues, 198–199
  NTP, 202
  unreachable instance, 197–198
  VLAN issues, 202–203
troubleshooting cloud IAM
  authentication and authorization issues, 307
  privilege issues, 308–309
troubleshooting cloud security
  key management issues, 261–262
  security misconfigurations, 260–261
troubleshooting methodology
  common issues, 26–27
  preventive measures, 28
  troubleshooting steps, 27–28
troubleshooting problems and performance, in monitoring
  quotas and throttling, 399–400
  restoration from backup, 402

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

scaling and sizing issues, 401–402
utilization issues, 400–401
trunk port, 202–203
type 1 hypervisor, 46
type 2 hypervisor, 46

U

unprivileged containers, 251
usability testing, 104
user account, 278
user error, 225
user interface (UI), 225
user quotas, 348

V

variable, 433
vector databases, 332
vendor lock-in, 85
version control, 99–100
editing code, 101
pull requests, 101–103
versioning, 335
vertical scaling, 108
virtual CPUs (vCPUs), 48
virtual desktop infrastructure (VDI), 48
virtual firewall, 246–247
virtual GPU (vGPU), 49
virtual IP (VIP), 187
virtualization technology (VT), 10, 46
hypervisors, 46
network connection types, 51–52
virtual machines, 46–48
VM configuration, 48–51
virtual local area networks (VLANs), 174–176
virtual machines (VMs), 46–47
affinity, 60–62
billing models, 56–58
cloning instances, 54–56
containers, 65–70
deploy VM in AWS, 74–75
deploy VM in Azure, 75–77
deploy VM in GCP, 77–78
instance types, 54, 78–80
resource allocation factors, 59–60
serverless computing, 63–65
virtual machines configuration
CPUs, 48, 49
GPUs, 49–50
memory, 50
storage, 50–51
virtual network (VNet), in Azure, 142, 152–153, 162–164
virtual network security, 234–235
allow and deny rules, 235–236
AWS VPCs and subnets, 238–243
Azure virtual networks, 243–246
GCP VPCs, 246–248
securing hybrid and multi-clouds, 248–249
zero trust security, 236–238
virtual NICs (vNICs), 51
virtual patch, 437
virtual private cloud (VPC), 142
in AWS, 145–146, 160–162, 238–243
configure security across, 272–273
flow logs, 240, 241, 247
in GCP, 156–157, 164–165, 246–248
virtual private networks (VPNs), 172–173
in AWS, 208–212
in Azure, 212–215
in GCP, 215–217
routing, 185–186
virtual RAM (vRAM), 50
visual recognition, 22
VLAN issues, 202–203
VLANs. *See* virtual local area networks (VLANs)
VM-attached storage
AWS storage services, 341
Azure storage services, 343
GCP storage services, 343
VMs. *See* virtual machines (VMs)
VMware, 5, 18
voice-to-text, 22
volume-based licensing, 48
VPC. *See* virtual private cloud (VPC)
VPNs. *See* virtual private networks (VPNs)
vSwitch (virtual switch), 51

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

VT. *See* virtualization technology (VT)
vulnerability management, 229
vulnerability testing, 104

W

waterfall method, 98, 99
web application firewall (WAF), 252
web service, 97
websites, cloud storage, 323
WebSocket, 96
workflow, 420
workforce identities, 281–282
write once read many (WORM), 330

X

XMPP. *See* Extensible Messaging and Presence Protocol (XMPP)

Y

YAML Ain't Markup Language (YAML), 297

Z

zero trust, 236
zero trust architecture (ZTA), 237
zero trust network access (ZTNA), 238
zero trust security, 236–238
ZIP file, 432
zombie resources, 226

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2027 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.