ANTHONY SEQUEIRA

Cert Guide

Learn, prepare, and practice for exam success



AWS Certified Solutions Architect -Associate (SAA-C01)

PEARSON IT CERTIFICATION



AWS Certified Solutions Architect - Associate (SAA-C01) Cert Guide

Anthony Sequeira, CCIE No. 15626



AWS Certified Solutions Architect - Associate (SAA-C01) Cert Guide

Copyright © 2019 by Pearson Education, Inc.

All rights reserved. No part of this book shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher. No patent liability is assumed with respect to the use of the information contained herein. Although every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions. Nor is any liability assumed for damages resulting from the use of the information contained herein.

ISBN-13: 978-0-7897-6049-4

ISBN-10: 0-7897-6049-5

Library of Congress Control Number: 2018963110

01 19

Trademarks

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Pearson IT Certification cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Warning and Disclaimer

Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied. The information provided is on an "as is" basis. The author and the publisher shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book. Editor-in-Chief Mark Taub

Product Line Manager Brett Bartow

Acquisitions Editor Paul Carlstroem

Development Editor Christopher Cleveland

Managing Editor Sandra Schroeder

Senior Project Editor Tonya Simpson

Copy Editor Chuck Hutchinson

Indexer Erika Millen

Proofreader Abigail Manheim

Technical Editor Ryan Dymek

Publishing Coordinator Cindy J. Teeters

Cover Designer Chuti Prasertsith

Compositor codemantra

Contents at a Glance

Introduction xx

Part I: Domain 1: Design Resilient Architectures

- CHAPTER 1 The Fundamentals of AWS 3
- CHAPTER 2 Designing Resilient Storage 39
- CHAPTER 3 Designing Decoupling Mechanisms 59
- CHAPTER 4 Designing a Multitier Infrastructure 69
- CHAPTER 5 Designing High Availability Architectures 81

Part II: Domain 2: Define Performant Architectures

- CHAPTER 6 Choosing Performant Storage 97
- CHAPTER 7 Choosing Performant Databases 111
- CHAPTER 8 Improving Performance with Caching 139
- CHAPTER 9 Designing for Elasticity 157
- Part III: Domain 3: Specify Secure Applications and Architectures
- CHAPTER 10 Securing Application Tiers 167
- CHAPTER 11 Securing Data 179
- CHAPTER 12 Networking Infrastructure for a Single VPC Application 190
- Part IV: Domain 4: Design Cost-Optimized Architectures
- CHAPTER 13 Cost-Optimized Storage 211
- CHAPTER 14 Cost-Optimized Compute 225
- Part V: Domain 5: Define Operationally Excellent Architectures
- CHAPTER 15 Features for Operational Excellence 235
- **Part VI: Final Preparation**
- CHAPTER 16 Final Preparation 245

Part VII: Appendixes

iv

Glossary 255

- APPENDIX A Answers to the "Do I Know This Already?" Quizzes and Q&A Sections 263
- APPENDIX B AWS Certified Solutions Architect Associate (SAA-C01) Cert Guide Exam Updates 273

Index 275

Table of Contents

Introduction xx

Part I: Domain 1: Design Resilient Architectures Chapter 1 The Fundamentals of AWS 3 "Do I Know This Already?" Quiz 3 Advantages of Cloud Technologies 6 An Overview of Key AWS Services 7 Compute Services 8 Elastic Compute Cloud 8 Lambda 10 Elastic Container Service 11 Elastic Container Registry 12 Elastic Container Service for Kubernetes 12 Fargate 13 Serverless Application Repository 13 Lightsail 13 AWS Batch 14 Elastic Beanstalk 14 Elastic Load Balancing 16 Auto Scaling 16 CloudFormation 17 Application Services 18 OpsWorks 18 CloudFront 18 Simple Queue Service 19 Simple Notification Service 19 Kinesis 19 Database Services 20 Aurora 20 Relational Database Service 20 DynamoDB 21 ElastiCache 22

Redshift 22 Database Migration Service 23 Networking Services 23 The AWS Global Infrastructure 23 Virtual Private Cloud 24 Direct Connect 25 Route 53 26 Storage Services 26 Simple Storage Service 26 Elastic Block Store 28 Elastic File System 29 Glacier 30 Snowball 30 AWS Storage Gateway 31 Security Services 31 Identity and Access Management 31 Web Application Firewall 32 Key Management Service 33 Directory Services 33 Management Services 33 Trusted Advisor 33 CloudWatch 34 CloudTrail 34 Review All Key Topics 35 Complete Tables and Lists from Memory 35 Define Key Terms 35 Q&A 36 Chapter 2 Designing Resilient Storage 39 "Do I Know This Already?" Quiz 39 Designing Resilient S3 Services 42 S3 Storage Classes 42 Lab: Creating an S3 Bucket 44 Lab Cleanup 47

Designing Resilient EBS Services 47 EBS Versus Instance Stores 47 Elastic Block Store 48 Lab: Creating an EBS Volume 48 Lab Cleanup 49 Elastic Block Store 49 Designing Resilient EFS Services 51 Lab: A Basic EFS Configuration 51 Lab Cleanup 53 Designing Resilient Glacier Services 53 Lab: Creating a Vault 54 Lab Cleanup 55 Review All Key Topics 56 Complete Tables and Lists from Memory 56 Define Key Terms 56 Q&A 57 Chapter 3 **Designing Decoupling Mechanisms** 59 "Do I Know This Already?" Quiz 59 Decoupling Demystified 61 Advantages of Decoupled Designs 62 Synchronous Decoupling 62 Asynchronous Decoupling 62 Lab: Configure SQS 63 Lab Cleanup 66 Review All Key Topics 67 Complete Tables and Lists from Memory 67 Define Key Terms 67 Q&A 67 Chapter 4 Designing a Multitier Infrastructure 69 "Do I Know This Already?" Quiz 69 Single-Tier Architectures 71 Lab: Building a Single-Tier Architecture with EC2 72 Lab Cleanup 74

Multitier Architectures 74 The Classic Three-Tier Architecture 76 Review All Key Topics 78 Complete Tables and Lists from Memory 78 Define Key Terms 78 Q&A 79 Chapter 5 Designing High Availability Architectures 81 "Do I Know This Already?" Quiz 81 High Availability Compute 84 Lab: Provisioning EC2 Instances in Different Availability Zones 85 Lab Cleanup 88 High Availability Application Services 88 High Availability Database Services 88 High Availability Networking Services 91 High Availability Storage Services 92 High Availability Security Services 92 High Availability Monitoring Services 93 Review All Key Topics 93 Complete Tables and Lists from Memory 94 Define Key Terms 94 O&A 94

Part II: Domain 2: Define Performant Architectures

Chapter 6 Choosing Performant Storage 97

"Do I Know This Already?" Quiz 97 Performant S3 Services 99 Performant EBS Services 101 Performant EFS Services 103 Performant Glacier Services 108 Review All Key Topics 108 Complete Tables and Lists from Memory 109 Define Key Terms 109 Q&A 109

```
Chapter 7
            Choosing Performant Databases 111
            "Do I Know This Already?" Quiz 111
            Aurora 114
                Which DB Instance Are You Connected To? 115
                When to Use T2 Instances 115
                Work with Asynchronous Key Prefetch 116
                Avoid Multithreaded Replication 116
                Use Scale Reads 117
                Consider Hash Joins 117
                Use TCP Keepalive Parameters 117
            RDS 118
            DynamoDB 119
                Burst Capacity 121
                Adaptive Capacity 122
                Secondary Indexes 122
                Querying and Scanning Data
                                             124
            ElastiCache 126
                Lazy Loading Versus Write Through 127
                Scenario 1: Cache Hit 127
                Scenario 2: Cache Miss 127
                Advantages and Disadvantages of Lazy Loading 127
                Advantages and Disadvantages of Write Through 128
                What Is TTL? 129
                Background Write Process and Memory Usage 129
                Avoid Running Out of Memory When Executing a Background Write 129
                How Much Reserved Memory Do You Need? 130
                Parameters to Manage Reserved Memory 130
                Online Cluster Resizing 131
            Redshift 132
                Amazon Redshift Best Practices for Designing Queries 133
                Work with Recommendations from Amazon Redshift Advisor 134
            Review All Key Topics 135
            Complete Tables and Lists from Memory 136
            Define Key Terms 136
            Q&A 136
```

Chapter 8	Improving Performance with Caching 139
	"Do I Know This Already?" Quiz 139
	ElastiCache 142
	Lab: Configuring a Redis ElastiCache Cluster 142
	Lab Cleanup 145
	DynamoDB Accelerator 145
	CloudFront 147
	Lab: Configuring CloudFront 147
	Lab Cleanup 149
	Greengrass 149
	Route 53 150
	Lab: Creating a Hosted Domain and DNS Records in Route 53 152
	Lab Cleanup 153
	Review All Key Topics 154
	Complete Tables and Lists from Memory 154
	Define Key Terms 154
	Q&A 155
Chapter 9	Designing for Elasticity 157
	"Do I Know This Already?" Quiz 157
	Elastic Load Balancing 159
	Auto Scaling 160
	Target Tracking Scaling Policies 162
	The Cooldown Period 163
	Review All Key Topics 164
	Complete Tables and Lists from Memory 164
	Define Key Terms 164
	Q&A 164
Part III: Don	nain 3: Specify Secure Applications and Architectures

Chapter 10 Securing Application Tiers 167

"Do I Know This Already?" Quiz 167 Using IAM 169 IAM Identities 171

Securing the OS and Applications 173 Security Groups 174 Network ACLs 175 Systems Manager Patch Manager 175 Review All Key Topics 176 Complete Tables and Lists from Memory 177 Define Key Terms 177 Q&A 177 Chapter 11 Securing Data 179 "Do I Know This Already?" Quiz 179 Resource Access Authorization 181 Storing and Managing Encryption Keys in the Cloud 182 Protecting Data at Rest 183 Decommissioning Data Securely 184 Protecting Data in Transit 185 Review All Key Topics 185 Complete Tables and Lists from Memory 185 Define Key Terms 186 Q&A 186 Chapter 12 Networking Infrastructure for a Single VPC Application 190 "Do I Know This Already?" Quiz 190 Introducing the Basic AWS Network Infrastructure 193 Lab: Checking Default Networking Components in a Region 194 Network Interfaces 198 Route Tables 199 Internet Gateways 201 Egress-Only Internet Gateways 201 DHCP Option Sets 202 DNS 202 Elastic IP Addresses 203 VPC Endpoints 204 Interface Endpoints (Powered by AWS PrivateLink) 204 Gateway Endpoints 205

NAT 205 VPC Peering 206 ClassicLink 206 Review All Key Topics 207 Complete Tables and Lists from Memory 207 Define Key Terms 207 Q&A 208

Part IV: Domain 4: Design Cost-Optimized Architectures

Chapter 13	Cost-Optimized Storage 211
	"Do I Know This Already?" Quiz 211
	S3 Services 214
	Lab: Estimating AWS S3 Costs 214
	Lab: Implementing Lifecycle Management 216
	Lab Cleanup 218
	EBS Services 218
	EFS Services 218
	Glacier Services 219
	Lab: Changing the Retrieval Rate in Glacier 220
	Lab Cleanup 221
	Review All Key Topics 221
	Complete Tables and Lists from Memory 221
	Define Key Terms 221
	Q&A 222
Chapter 14	Cost-Optimized Compute 225
	"Do I Know This Already?" Quiz 225
	Cost-Optimized EC2 Services 227
	Lab: Using the Cost Explorer 228
	Lab: Creating a Billing Alarm 230
	Cost-Optimized Lambda Services 231
	Review All Key Topics 233
	Complete Tables and Lists from Memory 233
	Define Key Terms 233
	Q&A 233

Part V: Domain 5: Define Operationally Excellent Architectures

Chapter 15 Features for Operational Excellence 235

"Do I Know This Already?" Quiz 235 Introduction to the AWS Well-Architected Framework 237 Prepare 239 **Operational Priorities** 239 Design for Operations 239 Operational Readiness 240 Operate 241 Understanding Operational Health 241 Responding to Events 241 Evolve 242 Learning from Experience 242 Share Learnings 242 Review All Key Topics 242 Complete Tables and Lists from Memory 243 Define Key Terms 243 Q&A 243

Part VI: Final Preparation

Chapter 16 Final Preparation 245 Exam Information 245 Getting Ready 248 Tools for Final Preparation 249 Pearson Test Prep Practice Test Engine and Questions on the Website 249 Accessing the Pearson Test Prep Practice Test Software Online 249 Accessing the Pearson Test Prep Practice Test Software Offline 250 Customizing Your Exams 251 Updating Your Exams 252 Premium Edition 252 Memory Tables 252 Chapter-Ending Review Tools 253 Suggested Plan for Final Review/Study 253 Summary 254

Part VII: Appendixes

Glossary 255

Appendix A Answers to the "Do I Know This Already?" Quizzes and Q&A Sections 263

Appendix B AWS Certified Solutions Architect – Associate (SAA-C01) Cert Guide Exam Updates 273

Index 275

About the Author

Anthony Sequeira, CCIE No. 15626, is a seasoned trainer and author regarding various levels and tracks of Cisco, Microsoft, and AWS certifications. In 1994, Anthony formally began his career in the information technology industry with IBM in Tampa, Florida. He quickly formed his own computer consultancy, Computer Solutions, and then discovered his true passion—teaching and writing about information technologies.

Anthony joined Mastering Computers in 1996 and lectured to massive audiences around the world about the latest in computer technologies. Mastering Computers became the revolutionary online training company KnowledgeNet, and Anthony trained there for many years.

Anthony is currently pursuing his second CCIE in the area of Cisco Data Center! He is a full-time instructor at CBT Nuggets.

Dedication

I dedicate this book to my incredible daughter, Bella Sequeira. While she may never read it to master AWS, she can at least use it as a rather expensive coaster in her beautiful new Oregon home.

Acknowledgments

This manuscript was made truly great by the incredible technical review of Ryan Dymek. Sometimes I think he might have invented AWS.

I would also like to express my gratitude to Chris Cleveland, development editor of this book. I was so incredibly lucky to work with him again on this text. Like Ryan, he made this book several cuts above the rest.

About the Technical Reviewer

Ryan Dymek has been working with Amazon Web Services (AWS) for more than 9 years and holds all nine AWS certifications as well as various Google Cloud Platform (GCP) certifications. Ryan trains and advises some of the largest companies in the world on sound architectural practices in cloud strategy and DevOps principles. While working with business leaders, developers, and engineers, Ryan bridges the gap between business and technology, maintaining the understanding and skills required to be able to perform at a deep technical level. Ryan runs his own cloud consulting practice advising more than 20 companies on the Fortune 500 list and has helped many startups find their way in the cloud.

In addition to cloud and technical acumen, Ryan is a certified business coach personally trained by John Maxwell. He uses these professional skills not only to advise companies on best cloud practices but also on how to align with a business's needs and culture, making confident business and technical decisions and cultivating a transformation into DevOps.

We Want to Hear from You!

As the reader of this book, *you* are our most important critic and commentator. We value your opinion and want to know what we're doing right, what we could do better, what areas you'd like to see us publish in, and any other words of wisdom you're willing to pass our way.

We welcome your comments. You can email to let us know what you did or didn't like about this book—as well as what we can do to make our books better.

Please note that we cannot help you with technical problems related to the topic of this book.

When you write, please be sure to include this book's title and author as well as your name and email address. We will carefully review your comments and share them with the author and editors who worked on the book.

Email: feedback@pearsonitcertification.com

Introduction

The AWS Certified Solutions Architect - Associate is a cloud-related certification that tests a candidate's ability to architect effective solutions calling upon the most popular aspects of Amazon Web Services. The solutions architect candidates must demonstrate their skills on how to plan and implement a sophisticated design that saves costs, is secure, and perhaps most importantly, operates with excellence. Candidates are also required to know the most important facts regarding various services and their capabilities.

The AWS Certified Solutions Architect is an Associate-level cloud career certification. This certification is an excellent second step after the achievement of the AWS Certified Cloud Practitioner certification, although seasoned AWS users might choose to skip that entry-level exam. Following this certification, AWS offers a Professional level of certification for the solutions architect.

AWS also offers certifications in which you might be interested in different tracks. For example, a Developer track for AWS also includes Associate and Professional levels. Amazon also uses specialty certifications to deep dive into many different areas such as security and advanced networking.

NOTE The AWS Certified Solutions Architect - Associate certification is globally recognized and does an excellent job of demonstrating that the holder has knowledge and skills across a broad range of AWS topics.

The Goals of the AWS Certified Solutions Architect - Associate Certification

The AWS Certified Solutions Architect - Associate certification is intended for individuals who perform a solutions architect role. The certification seeks to validate your ability to effectively demonstrate knowledge of how to architect and deploy secure and robust applications on AWS technologies.

You should be able to define a solution using architectural design principles based on customer requirements. You should also be able to provide implementation guidance based on best practices to the organization throughout the lifecycle of the project.

Ideal Candidates

Although this text provides you with the information required to pass this exam, Amazon considers ideal candidates to be those who possess the following:

 One year of hands-on experience designing available, cost-efficient, faulttolerant, and scalable distributed systems on AWS

- Hands-on experience using computing, networking, storage, and database AWS services
- Hands-on experience with AWS deployment and management services
- The ability to identify and define technical requirements for an AWS-based application
- The ability to identify which AWS services meet a given technical requirement
- Knowledge of recommended best practices for building secure and reliable applications on the AWS platform
- An understanding of the basic architectural principles of building on the AWS cloud
- An understanding of the AWS global infrastructure
- An understanding of network technologies as they relate to AWS
- An understanding of security features and tools that AWS provides and how they relate to traditional services

The Exam Objectives (Domains)

The AWS Certified Solutions Architect - Associate (SAA-C01) exam is broken down into five major domains. The contents of this book cover each of the domains and the subtopics included in them as illustrated in the following descriptions.

The following table breaks down each domain represented in the exam.

Domain	Percentage of Representation in Exam
1: Design Resilient Architectures	34%
2: Define Performant Architectures	24%
3: Specify Secure Applications and Architectures	26%
4: Design Cost-Optimized Architectures	10%
5: Define Operationally Excellent Architectures	6%
	Total 100%

1.0 Design Resilient Architectures

The Design Resilient Architectures domain is covered in Chapters 1 through 5. It covers critical information for designing and deploying highly available services and resources in the cloud. It comprises 34 percent of the exam. Topics include

- 1.1 Choose reliable/resilient storage.
- 1.2 Determine how to design decoupling mechanisms using AWS services.
- 1.3 Determine how to design a multitier architecture solution.
- 1.4 Determine how to design high availability and/or fault-tolerant architectures.

2.0 Define Performant Architectures

The Define Performant Architectures domain is covered in Chapters 6 through 9. This domain ensures that you consider and properly implement solutions that perform per the client requirements. This makes up 24 percent of the exam. Topics include

- 2.1 Choose performant storage and databases.
- 2.2 Apply caching to improve performance.
- 2.3 Design solutions for elasticity and scalability.

3.0 Specify Secure Applications and Architectures

The Specify Secure Applications and Architectures domain is covered in Chapters 10 through 12. This domain is critical, especially in today's landscape, because it assists you in an AWS implementation that features security at many layers. This builds a Defense in Depth type of strategy that increases your chances of warding off would-be attackers. It encompasses 26 percent of the exam. Topics include

- 3.1 Determine how to secure application tiers.
- 3.2 Determine how to secure data.
- **3.3** Define the networking infrastructure for a single VPC application.

4.0 Design Cost-Optimized Architectures

The Design Cost-Optimized Architectures domain is covered in Chapters 13 and 14. Here you learn about saving on costs where these costs could be the most substantial—in the areas of compute and storage. This domain embodies 10 percent of the exam. The topics include

- 4.1 Determine how to design cost-optimized storage.
- 4.2 Determine how to design cost-optimized compute.

5.0 Define Operationally Excellent Architectures

The Define Operationally Excellent Architectures domain is covered in Chapter 15. This chapter ensures you understand how to use AWS best practices around the three main steps of Prepare, Operate, and Evolve with AWS solutions. This domain makes up 6 percent of the exam. Topics include

5.1 Choose design features in solutions that enable operational excellence.

Steps to Becoming an AWS Certified Solutions Architect - Associate

To become an AWS Certified Solutions Architect - Associate, a test candidate should meet certain prerequisites (none of these are formal prerequisites) and follow specific procedures. Test candidates must qualify for the exam and sign up for the exam.

Recommended Experience

There are no prerequisites for the Solutions Architect - Associate certification. However, Amazon recommends that candidates possess the Certified Cloud Practitioner certification or equivalent knowledge.

NOTE Other certifications you might possess in related areas, such as Microsoft Azure, can also prove beneficial.

Signing Up for the Exam

The steps required to sign up for the Solutions Architect - Associate exam are as follows:

- **Step 1.** Create an AWS Certification account at https://www.aws.training/ Certification and schedule your exam.
- **Step 2.** Complete the Examination Agreement, attesting to the truth of your assertions regarding professional experience and legally committing to the adherence of the testing policies.
- **Step 3.** Submit the examination fee.

Facts About the Exam

The exam is a computer-based test. The exam consists of multiple-choice questions only. You must bring a government-issued identification card. No other forms of ID will be accepted.

TIP Refer to the AWS Certification site at https://aws.amazon.com/certification/ for more information regarding this, and other, AWS Certifications. I am also in the process of building a simple hub site for everything AWS Certification related at awscerthub.com. This site is made up of 100 percent AWS solutions. Of course!

About the Solutions Architect – Associate Cert Guide

This book maps directly to the topic areas of the exam and uses a number of features to help you understand the topics and prepare for the exam.

Objectives and Methods

This book uses several key methodologies to help you discover the exam topics on which you need more review, to help you fully understand and remember those details, and to help you prove to yourself that you have retained your knowledge of those topics. This book does not try to help you pass the exam only by memorization; it seeks to help you to truly learn and understand the topics. This book is designed to help you pass the AWS Certified Solutions Architect - Associate exam by using the following methods:

- Helping you discover which exam topics you have not mastered
- Providing explanations and information to fill in your knowledge gaps
- Supplying exercises that enhance your ability to recall and deduce the answers to test questions
- Providing practice exercises on the topics and the testing process via test questions on the companion website

Book Features

To help you customize your study time using this book, the core chapters have several features that help you make the best use of your time:

- Foundation Topics: These are the core sections of each chapter. They explain the concepts for the topics in that chapter.
- Exam Preparation Tasks: After the "Foundation Topics" section of each chapter, the "Exam Preparation Tasks" section lists a series of study activities that you should do at the end of the chapter:
 - Review All Key Topics: The Key Topic icon appears next to the most important items in the "Foundation Topics" section of the chapter. The "Review All Key Topics" activity lists the key topics from the chapter,

along with their page numbers. Although the contents of the entire chapter could be on the exam, you should definitely know the information listed in each key topic, so you should review these.

- Define Key Terms: Although the Solutions Architect exam may be unlikely to ask a question such as "Define this term," the exam does require that you learn and know a lot of new terminology. This section lists the most important terms from the chapter, asking you to write a short definition and compare your answer to the glossary at the end of the book.
- Review Questions: Confirm that you understand the content that you just covered by answering these questions and reading the answer explanations.
- Web-based Practice Exam: The companion website includes the Pearson Cert Practice test engine that allows you to take practice exam questions. Use it to prepare with a sample exam and to pinpoint topics where you need more study.

Figure Credits

Chapter 1, Figures 1-1 through 1-7, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 2, Figures 2-1 through 2-4, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 3, Figures 3-1 through 3-3, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 4, Figure 4-2, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 5, Figures 5-1 and 5-2, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 6, quote from Amazon in the note courtesy of Amazon Web Services, Inc.

Chapter 6, Figure 6-1, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 7, Figures 7-1 and 7-2, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 8, Figures 8-1 through 8-4, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 9, Figures 9-1 and 9-2, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 10, Figures 10-1 through 10-4, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 11, reference to NIST 800-88 courtesy of NIST 800-88 "Guidelines for Media Sanitization," Recommendations of the National Institute of Standards and Technology, Richard Kissel September, 2006.

Chapter 12, Figures 12-1 through 12-3, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 13, Figures 13-1 through 13-3, screenshot of AWS © 2018, Amazon Web Services, Inc.

Chapter 14, Figures 14-1 through 14-3, screenshot of AWS © 2018, Amazon Web Services, Inc.

Designing for Elasticity

We discussed elasticity briefly when we presented the various advantages that cloud technologies provide IT solutions in Chapter 1, "The Fundamentals of AWS." Remember, *elasticity* refers to the capability of your resources to scale up and down as well as in and out when resource requirements change due to changes in demand.

"Do I Know This Already?" Quiz

The "Do I Know This Already?" quiz allows you to assess whether you should read the entire chapter. Table 9-1 lists the major headings in this chapter and the "Do I Know This Already?" quiz questions covering the material in those headings so you can assess your knowledge of these specific areas. The answers to the "Do I Know This Already?" quiz appear in Appendix A.

 Table 9-1
 "Do I Know This Already?" Foundation Topics Section-to-Question Mapping

Foundation Topics Section	Question
Elastic Load Balancing	1–2
Auto Scaling	3-4

CAUTION The goal of self-assessment is to gauge your mastery of the topics in this chapter. If you do not know the answer to a question or are only partially sure of the answer, you should mark that question as wrong for purposes of the self-assessment. Giving yourself credit for an answer you correctly guess skews your self-assessment results and might provide you with a false sense of security.

- 1. Which is not a common service used with ELB?
 - **a.** EC2
 - **b.** CloudWatch
 - c. EBS
 - d. Route 53
- 2. Which of the following is not a valid type of Elastic Load Balancer in AWS?
 - a. Classic
 - **b.** Service-based
 - c. Network
 - d. Application
- 3. How many Auto Scaling policies can you set against an AWS resource?
 - **a.** 1
 - **b.** 10
 - **c.** 50
 - **d.** 100
- 4. What is the cooldown period used with Auto Scaling?
 - **a.** It is the mandatory time that a service must run before sampling can begin.
 - **b.** It is the mandatory sample time of metrics for policy-based Auto Scaling.
 - **c.** It is the mandatory time that Auto Scaling must wait after a service reboots before scale-out operations.
 - **d.** It is the mandatory time that Auto Scaling must wait before it takes additional scaling actions.

Foundation Topics

Elastic Load Balancing

Elastic Load Balancing (ELB) distributes incoming application or network traffic across multiple targets. These targets are AWS resources such as Amazon EC2 instances, containers, and IP addresses. To foster high availability, ensure the resources are in multiple Availability Zones.

Elastic Load Balancing scales your load balancer as traffic to your application changes over time and can scale to a majority of workloads automatically. You can add and remove compute resources from your load balancer as your needs change, without disrupting the overall flow of requests to your applications.

You can configure health checks, which are used to monitor the health of the computing resources so that the load balancer can send requests only to the healthy ones. You can also offload the work of SSL/TLS encryption and decryption to your load balancer so that your compute resources can focus on their main work.

Elastic Load Balancing supports three types of load balancers: Application Load Balancers, Network Load Balancers, and Classic Load Balancers. You can select a load balancer based on your application needs. Figure 9-1 shows the configuration of a Network Load Balancer in AWS.

EC2 Management	Cons × + ~		
- → O ଲ	යි https://sa-east-1.console.aws.amazon.com/ec2/v2/home?region=sa-east-14V2Creat 🔲 🙀 🧯	L	6 🖬
aws Servic	ces 🗸 Resource Groups 👻 🚯 🗘 asequeira 🕫 Ostraugg 🍷 São Par	do = 5	Support 🔹
1. Configure Load Balancer	2. Configure Routing 1. Register Targets 4. Review		
Step 1: Configure	e Load Balancer		
Basic Configuration	n		
configure your load baland	cer, provide a name, select a scheme, specify one or more listeners, and select a network. The default of	onfiguratio	on is an
ternet-facing load balancer	in the selected network with a listener that receives TCP traffic on port 80.		
Name (j)	Only a-z, A-Z, 0-9 and hyphens are allowed		
Name (j) Scheme (j)	Only a-z, A-Z, 0-9 and hyphens are allowed Only a-z, A-Z, 0-9 and		
Name () Scheme ()	Only a-z, A-Z, 0-9 and hyphens are allowed Imternet-tacing Internal		
Name (j) Scheme (j)	Only a-z, A-Z, 0-9 and hyphens are allowed internet facing O internal		
Name () Scheme () Listeners	Only a-z, A-Z, 6-9 and hyphens are allowed Internet facing Internal		
Name () Scheme () .isteners	Only a-Z, A-Z, 0-9 and tryphens are allowed I internet-facing I internal Hecks for connection requests, using the protocol and port that you configured.		
Name (j) Scheme (j) Listeners Listener is a process that ch Load Balancer Protocol	Only a-2, A-2, 0-9 and hyphens are allowed Immediating Immediation Internal mecks for connection requests, using the protocol and port that you configured. Load Balancer Port		
Name () Scheme () isteners Sistener is a process that of Load Balancer Protocol TCP V	Only a-z, A-Z, 0-9 and hyphens are allowed Internet-facing Internal Internal Internal Load Balancer Port 00		6
Name () Scheme () Isteners Vistener is a process that ch Load Balancer Protocol (TCP) Add listener	Only a-z, A-Z, 0-9 and hyphens are allowed internet-facing internal necks for connection requests, using the protocol and port that you configured. Load Balancer Port 00		6
Name () Scheme () Scheme () Sisteners Sisteners that of Load Balancer Protocol (TCP) Add listener	Only a-2, A-2, 0-9 and hyphens are allowed Immediating Immediation Internal Immediation Immediate for the protocol and port that you configured. Load Balancer Port B0		G
Name () Scheme () Sisteners Sistener is a process that of Load Balancer Protocol TCP V Add listener	Only a-2, A-2, 0-9 and hyphens are allowed Immediating Internal mecks for connection requests, using the protocol and port that you configured. Load Balancer Port 00		G

FIGURE 9-1 Configuring a Network Load Balancer in AWS

Although you can monitor and configure your Elastic Load Balancer with the traditional options of AWS (such as the Management Console or SDKs), you can also use the Query API. This interface provides low-level API actions that you call using HTTPS requests.

Elastic Load Balancing works with the following services to improve the availability and scalability of your applications:

- **EC2:** These virtual servers run your applications in the cloud. You can configure your load balancer to route traffic to your EC2 instances.
- **ECS:** This service enables you to run, stop, and manage Docker containers on a cluster of EC2 instances. You can configure your load balancer to route traffic to your containers.
- Auto Scaling: This service ensures that you are running your desired number of instances, even if an instance fails, and enables you to automatically increase or decrease the number of instances as the demand on your instances changes. If you enable Auto Scaling with Elastic Load Balancing, instances that are launched by Auto Scaling are automatically registered with the load balancer, and instances that are terminated by Auto Scaling are automatically deregistered from the load balancer.
- CloudWatch: This service enables you to monitor your load balancer and take action as needed.
- Route 53: This service provides a reliable and cost-effective way to route visitors to websites by translating domain names (such as www.ajsnetworking. com) into the numeric IP addresses (such as 69.163.163.123) that computers use to connect to each other. AWS assigns URLs to your resources, such as load balancers. However, you might want a URL that is easy for users to remember. For example, you can map your domain name to a load balancer.

Auto Scaling

Auto Scaling enables you to quickly discover the scalable AWS resources that are part of your application and configure dynamic scaling in a matter of minutes. The Auto Scaling console provides a single user interface to use the automatic scaling features of multiple AWS services. It also offers recommendations to configure scaling for the scalable resources in your application.

Use Auto Scaling to automatically scale the following resources that support your application:

- EC2 Auto Scaling groups
- Aurora DB clusters





- DynamoDB global secondary indexes
- DynamoDB tables
- ECS services
- Spot Fleet requests

With Auto Scaling, you create a scaling plan with a set of instructions used to configure dynamic scaling for the scalable resources in your application. Auto Scaling creates target tracking scaling policies for the scalable resources in your scaling plan. Target tracking scaling policies adjust the capacity of your scalable resource as required to maintain resource utilization at the target value that you specified.

You can create one scaling plan per application source (an AWS CloudFormation stack or a set of tags). You can add each scalable resource to one scaling plan. If you have already configured scaling policies for a scalable resource in your application, Auto Scaling keeps the existing scaling policies instead of creating additional scaling policies for the resource.

Auto Scaling involves the creation of a Launch Configuration and an Auto Scaling group. Figure 9-2 shows the configuration of Auto Scaling in AWS.

	Services - R	esource	Groups 🗸	*	Δ	asequeira @ ct	tnugg 👻	São Paul		Suppo	et. *
1. Choose AMI	2. Choose Instance Type	3. Co	onfigure details	4. Add Storage	5. Configu	re Security Group	6. Review				
create La	aunch Configur	ation	6								
	Name	1	[
	Purchasing option	1	Request S	pot Instances							
	IAM role	1	None			~					
	Monitoring	()	Enable Clo Learn more	oudWatch detailed	monitoring						
Advance	d Details	~	The state of								
	RAM Disk ID	0	Use default			~					
	User data	0	As text O	As file 🗌 Input is	aiready bas	e64 encoded					
			(Optional)								
			Cashi araja	n a public (P addr	ess to instar	ices launched in	the default \	PC and s	subnet		

FIGURE 9-2 Configuring Auto Scaling in AWS

Target Tracking Scaling Policies

With target tracking scaling policies, you select a predefined metric or configure a customized metric and set a target value. Application Auto Scaling creates and manages the CloudWatch alarms that trigger the scaling policy and calculates the scaling adjustment based on the metric and the target value. The scaling policy adds or removes capacity as required to keep the metric at, or close to, the specified target value. In addition to keeping the metric close to the target value, a target tracking scaling policy also adjusts to changes in the metric due to a changing load pattern and minimizes changes to the capacity of the scalable target.

When specifying a customized metric, be aware that not all metrics work for target tracking. The metric must be a valid utilization metric and describe how busy a scalable target is. The metric value must increase or decrease proportionally to the capacity of the scalable target so that the metric data can be used to proportionally scale the scalable target.

You can have multiple target tracking scaling policies for a scalable target, provided that each of them uses a different metric. Application Auto Scaling scales based on the policy that provides the largest capacity for both scale-in and scale-out. This provides greater flexibility to cover multiple scenarios and ensures that there is always enough capacity to process your application workloads.

NOTE When discussing scaling the resources of a service, we are scaling those resources horizontally (out and in with elasticity), while the service made up of those resources is being scaled up and down (vertically because the single service is getting bigger or smaller). A single service scales both up and down and out and in, depending on the context.

You can also optionally disable the scale-in portion of a target tracking scaling policy. This feature provides the flexibility to use a different method for scale-in than you use for scale-out.

Keep the following in mind for Auto Scaling:

- You cannot create target tracking scaling policies for Amazon EMR clusters or AppStream 2.0 fleets.
- You can create 50 scaling policies per scalable target. This includes both step scaling policies and target tracking policies.
- A target tracking scaling policy assumes that it should perform scale-out when the specified metric is above the target value. You cannot use a target tracking scaling policy to scale out when the specified metric is below the target value.

- A target tracking scaling policy does not perform scaling when the specified metric has insufficient data. It does not perform scale-in because it does not interpret insufficient data as low utilization. To scale in when a metric has insufficient data, create a step scaling policy and have an alarm invoke the scaling policy when it changes to the INSUFFICIENT_DATA state.
- You may see gaps between the target value and the actual metric data points. The reason is that Application Auto Scaling always acts conservatively by rounding up or down when it determines how much capacity to add or remove. This prevents it from adding insufficient capacity or removing too much capacity. However, for a scalable target with small capacity, the actual metric data points might seem far from the target value. For a scalable target with larger capacity, adding or removing capacity causes less of a gap between the target value and the actual metric data points.
- We recommend that you scale based on metrics with a 1-minute frequency because that ensures a faster response to utilization changes. Scaling on metrics with a 5-minute frequency can result in slower response time and scaling on stale metric data.
- To ensure application availability, Application Auto Scaling scales out proportionally to the metric as fast as it can but scales in more gradually.
- Do not edit or delete the CloudWatch alarms that Application Auto Scaling manages for a target tracking scaling policy. Application Auto Scaling deletes the alarms automatically when you delete the Auto Scaling policy.

The Cooldown Period

The scale-*out* cooldown period is the amount of time, in seconds, after a scale-out activity completes before another scale-out activity can start. While this cooldown period is in effect, the capacity that has been added by the previous scale-out event that initiated the cooldown is calculated as part of the desired capacity for the next scale-out event. The intention is to continuously scale out.

The scale-*in* cooldown period is the amount of time, in seconds, after a scale-in activity completes before another scale-in activity can start. This cooldown period is used to block subsequent scale-in events until it has expired. The intention is to scale in conservatively to protect your application's availability. However, if another alarm triggers a scale-out policy during the cooldown period after a scale-in event, Application Auto Scaling scales out your scalable target immediately.

Exam Preparation Tasks

As mentioned in the section "How to Use This Book" in the Introduction, you have a couple of choices for exam preparation: the exercises here, Chapter 16, "Final Preparation," and the exam simulation questions in the Pearson Test Prep practice test software.

Review All Key Topics

Review the most important topics in this chapter, noted with the Key Topics icon in the margin of the page. Table 9-2 lists a reference of these key topics and the page numbers on which each is found.

Kov	
rey.	
I OPIC	

Table 9-2 Key Topics for Chapter 9

Key Topic Element	Description	Page Number		
List	Resources you can use with Elastic Load Balancing	160		
List	Resources you can use with Auto Scaling	160		

Complete Tables and Lists from Memory

There are no memory tables in this chapter.

Define Key Terms

Define the following key term from this chapter and check your answer in the glossary:

Cooldown period

Q&A

The answers to these questions appear in Appendix A. For more practice with exam format questions, use the Pearson Test Prep practice test software.

- 1. What are the three types of load balancers in AWS?
- 2. Do you need to delete the individual CloudWatch alarms that you use for Auto Scaling when you delete the Auto Scaling policy ?

Index

A

access ACLs (access control lists), 173, 175 authorization, 181-182 IAM (Identity and Access Management) capabilities of, 169-171 groups, 171 policies, 172-173 resource access authorization, 181 - 182roles, 172 users, 171 Pearson Test Prep practice test engine, 249-251 ACLs (access control lists), 173, 175 Active Directory, 33 adaptive capacity, DynamoDB, 122 addresses IP (Internet Protocol), 85, 193, 203-204 NAT (network address translation), 194, 205-206 Advisor, Amazon Redshift, 134–135 AKP (Asynchronous Key Prefetch), 116 alarms, creating, 230-231 alias records, 149-150 allocated memory utilization, 232 Amazon Aurora. See Aurora Amazon CloudFront, 18, 99, 100, 147-149 Amazon CloudWatch. See CloudWatch

Amazon Cognito, 77 Amazon DynamoDB. See DynamoDB Amazon Elastic Block Store. See EBS (Elastic Block Store) Amazon Elastic Compute Cloud. See EC2 (Elastic Compute Cloud) services Amazon Elastic Container Service. See ECS (Elastic Container Service) Amazon Elastic File System. See EFS (Elastic File System) Amazon Glacier. See Glacier services Amazon Kinesis, 19 Amazon Machine Images (AMIs), 47 Amazon Redshift. See Redshift Amazon Route 53. See Route 53 service Amazon Simple Storage Service. See S3 (Simple Storage Service) Amazon Virtual Private Cloud. See VPC (Virtual Private Cloud) AMIs (Amazon Machine Images), 47 APIs (application programming interfaces), support for, 7 application security IAM (Identity and Access Management) capabilities of, 169–171 groups, 171 policies, 172-173 resource access authorization, 181 - 182roles, 172

users, 171

network ACLs (access control lists), 175 security groups, 174 application services CloudFront, 18, 99, 100, 147–149 HA (high availability), 88 Kinesis, 19 OpsWorks, 18 SQS (Simple Queue Service), 19, 63-66 architecture HA (high availability), 81 application services, 88 Availability Zones, 85-87 database services, 88–90 EC2 instances, 85-88 FT (fault tolerance), 84 metrics, 84 monitoring services, 93 networking services, 91-92 overview of, 84-85 security services, 92-93 multitier advantages of, 74-75 components of, 75–76 three-tier architecture, 76–77 single-tier building with EC2, 72–74 example of, 71–72 SOA (service-oriented architecture), 74 asynchronous decoupling overview of, 62-63 SQS (Simple Queue Service), 63–66 Asynchronous Key Prefetch (AKP), 116 Aurora AKP (Asynchronous Key Prefetch), 116 Aurora Serverless, 114 DB connections, showing, 115 hash joins, 117 multithreaded replication, 116

overview of, 20, 114–115 read scaling capabilities, 117 T2 instances, 115–116 TCP keepalive parameters, 117–118 AuroraReplicaLag, 116 authoritative name servers, 150 authorization, 181–182 Auto Scaling, 16–17, 88, 160–163 Availability Zones. See HA (high availability) architecture AWS Batch, 14 AWS Cloud Compliance, 239 AWS CloudFormation, 17 AWS CloudTrail. See CloudTrail AWS Config, 240 AWS Cost Explorer, 228–230 AWS Database Migration Service, 23 AWS Direct Connect, 25, 91–92, 185 AWS Directory Service, 33 AWS Elastic Beanstalk, 14–15 AWS Elastic MapReduce, 184 AWS Identity and Access Management. See IAM (Identity and Access Management) AWS Key Management Service, 33, 99, 182 - 183AWS Lambda services, 230–231, 240 AWS Management Console Lifecycle Management, 216–218 SQS (Simple Queue Service) configuration, 63-66 AWS network infrastructure. See network infrastructure AWS OpsWorks, 18 AWS Security Token Service (AWS STS), 170AWS Serverless Application Repository, 13 AWS Simple Monthly calculator, 214–216 AWS Snowball, 30 AWS Storage Gateway, 31

AWS Systems Manager Patch Manager, 175–176 Run Command, 240 AWS Trusted Advisor, 33–34, 239 AWS Web Application Firewall, 32 AWS Well-Architected Framework. *See* Well-Architected Framework AWS X-Ray, 240

В

background write process (ElastiCache), 129-130 balancing load. See ELB (Elastic Load Balancing) Batch, 14 Billed Duration value, 232 billing. See also cost optimization alarms, creating, 230-231 billed duration utilization, 232 metrics, 230 BitBucket, 242 black boxes, 61 **BRPOPLPUSH** command, 132 buckets (S3) creating, 44-46, 147, 216 deleting, 47, 149, 218 bulk retrieval (Glacier), 108, 219 burst capacity (DynamoDB), 121 BurstCreditBalance metric, 106 Bursting Throughput mode (EFS), 105-106, 219

С

caching CloudFront, 18, 99, 100, 147–149 DAX (DynamoDB Accelerator), 145–146 ElastiCache background write process, 129–130 lazy loading, 127–128 online cluster resizing, 131–132

overview of, 22, 77, 126–127, 142-145 reserved memory, 130-131 TTL (time to live), 129 write through, 128–129 Greengrass, 149–150 Route 53 service, 26, 150-153, 160 calendars, Simple Monthly, 214–216 CapEx, 6 CASE expression, 134 CI/CD (Continuous Integration/ Continuous Deployment) pipeline, 239 classes, S3 (Simple Storage Service), 42 - 44classic three-tier architecture, 76-77 ClassicLink, 194, 206 clearing data, 184 Cloud Compliance, 239 CloudFormation, 17 CloudFront, 18, 99, 100, 147–149 CloudTrail API activity tracking, 242 overview of, 34 storing encryption keys in, 182–183 CloudWatch baselines, 241 billing alarms, 230–231 BurstCreditBalance metric, 106 capabilities of, 160 data aggregation in, 93 HA (high availability), 93 operational health monitoring, 241 overview of, 34 clusters DAX (DynamoDB Accelerator), 146 Redis ElastiCache, 131–132, 142-145 Cognito service, 77 Cold HDD, 49–51, 102

compute services Auto Scaling capabilities of, 16-17, 88, 160-161 target tracking scaling policies, 162 - 163Batch, 14 CloudFormation, 17 EC2 (Elastic Compute Cloud) services Auto Scaling, 88 cost optimization, 227-231 HA (high availability) architecture, 85-88 instances, 72-74, 85-88 overview of, 8-10 purpose of, 160 resource access authorization, 181–182 single-tier architecture, 72–74 ECR (Elastic Container Registry), 12 - 13ECS (Elastic Container Service), 11-12, 160 EKS (Elastic Container Service for Kubernetes), 12–13 Elastic Beanstalk, 14–15 ELB (Elastic Load Balancing), 16, 62, 159 - 160Fargate, 13 Lambda services cost optimization, 230-231 metrics, 232 operational readiness and, 240 overview of, 10-11 Lightsail, 13–14 Serverless Application Repository, 13 Config, 240 configuration CloudFront, 147–149 EFS (Elastic File System), 51–53 Greengrass, 149–150 Lifecycle Management, 216–218 Redis ElastiCache clusters, 131–132, 142 - 145

Route 53 service, 150–153 SQS (Simple Queue Service), 63–66 Configure Queue command, 64 Connect to Redis Server command, 145 container management ECR (Elastic Container Registry), 12–13 ECS (Elastic Container Service), 11-12, 160 EKS (Elastic Container Service for Kubernetes), 12–13 Continuous Integration/Continuous Deployment (CI/CD) pipeline, 239 contractual commitments, lack of, 6 convertible RIs (reserved instances), 227 cooldown periods, 163 COPY command, 133 Cost Explorer, 228–230 cost optimization EC2 (Elastic Compute Cloud) services billing alarms, 230–231 considerations for, 227-228 Cost Explorer, 228–230 Lambda services, 231–232 storage EBS (Elastic Block Store), 218 EFS (Elastic File System), 218–219 Glacier services, 219–221 S3 (Simple Storage Service), 214–218 cp command, 100 CPUCreditBalance (Aurora), 115 Create Bucket command, 44, 147, 216 Create Distribution command, 148 Create File System command, 51 Create Hosted Zone command, 152 Create New Queue command, 64 Create Queue command, 66 Create Record Set command, 152 Create Vault command, 54, 220 Create Volume command, 49

D

Data Requests charge (S3), 216 Data Retrieval charge (S3), 216 data security data at rest, 183–184 data in transit, 185 decommissioning process, 184 encryption keys, storing in cloud, 182 - 183resource access authorization, 181–182 SSL (Secure Sockets Layer), 185 VPC (Virtual Private Cloud), 185 Data Security Standard (DSS), 170 data stores IAM-enabled options, 77 VPC-hosted options, 77 data tier, 77 Database Migration Service, 23 database services Aurora AKP (Asynchronous Key Prefetch), 116 Aurora Serverless, 114 DB connections, showing, 115 hash joins, 117 multithreaded replication, 116 overview of, 20, 114-115 read scaling capabilities, 117 T2 instances, 115–116 TCP keepalive parameters, 117–118 Database Migration Service, 23 DynamoDB adaptive capacity, 122 burst capacity, 121 overview of, 21–22, 77 query and design patterns, 119–121 query operations, 124-126 scan operations, 124-126 secondary indexes, 122–124 ElastiCache background write process, 129–130 configuration, 142–145 lazy loading, 127–128 online cluster resizing, 131–132 overview of, 22, 77, 126–127

reserved memory, 130–131 TTL (time to live), 129 write through, 128–129 HA (high availability), 88–90 RDS (Relational Database Services) best practices, 118-119 HA (high availability) architecture, 88-90 overview of, 20-21, 77 Redshift Amazon Redshift Advisor, 134–135 best practices, 132–134 overview of, 22, 77 DAX (DynamoDB Accelerator), 145–146 decommissioning data, 184 decoupling advantages of, 62 asynchronous, 62–63 definition of, 61 SQS (Simple Queue Service), 63–66 synchronous, 62 delegation sets, 151 Delete Bucket command, 47, 149, 218 Delete File System command, 53 Delete Hosted Zone command, 153 Delete Vault command, 55, 221 Delete Volume command, 49 deleting EBS (Elastic Block Store) volumes, 49 EFS (Elastic File System) file systems, 53 Glacier vaults, 55, 221 hosted zones, 153 queues, 66 S3 (Simple Storage Service) buckets, 47, 149, 218 On-Demand requests (Glacier), 108 DescribeDBInstances action, 89 describe-db-instances command, 89 design patterns, DynamoDB, 119–121 design principles, 237–238, 239–240 destroying data, 184

DHCP (Dynamic Host Control Protocol) option sets, 193, 202 Direct Connect, 25, 91–92, 185 Directory Service, 33 distribution, CloudFront, 147-149 DNS (Domain Name System) overview of, 193, 202-203 private, 151 records, 151, 152–153 resolver, 150–151 DSS (Data Security Standard), 170 Duration value, 232 Dynamic Host Control Protocol. See DHCP (Dynamic Host Control Protocol) option sets DynamoDB adaptive capacity, 122 burst capacity, 121 DAX (DynamoDB Accelerator), 145-146 overview of, 21-22, 77 query and design patterns, 119–121 query operations, 124-126 scan operations, 124-126 secondary indexes, 122–124 DynamoDB Accelerator (DAX), 145–146

Ε

EBS (Elastic Block Store) cost optimization, 218 EBS-optimized instances, 101–103 instance stores compared to, 47 overview of, 28–29 performance of, 101–103 resiliency, 47 volume creation, 48–49 volume deletion, 49 volume types, 49–51 volumes creating, 48–49 deleting, 49 types of, 49–51 EC2 (Elastic Compute Cloud) services Auto Scaling, 88 cost optimization billing alarms, 230-231 considerations for, 227-228 Cost Explorer, 228–230 HA (high availability) architecture, 85 - 88instances launching, 72–73 provisioning in different Availability Zones, 85–87 terminating, 74, 88 viewing, 73 overview of, 8-10 purpose of, 160 resource access authorization, 181 - 182single-tier architecture, building, 72 - 74ECR (Elastic Container Registry), 12 - 13ECS (Elastic Container Service), 11-12, 160edge locations, 147 EFS (Elastic File System) basic configuration, 51–53 cost optimization, 218-219 File Sync, 219 file systems creating, 51–53 deleting, 53 overview of, 29, 51 performance of, 103–107 Bursting Throughput mode, 105-106 General Purpose mode, 105 high-level characteristics, 103–104 performance tips, 107 Provisioned Throughput mode, 105-106 resiliency, 51-53

egress-only Internet gateways, 193, 201-202 EKS (Elastic Container Service for Kubernetes), 12–13 Elastic Beanstalk, 14–15 Elastic Block Store. See EBS (Elastic Block Store) Elastic Compute Cloud. See EC2 (Elastic Compute Cloud) services Elastic Container Registry (ECR), 12–13 Elastic Container Service (ECS), 11-12, 160Elastic Container Service for Kubernetes (EKS), 12–13 Elastic File System. See EFS (Elastic File System) elastic IP addresses, 85 Elastic Load Balancing (ELB), 62, 159-160 Elastic MapReduce, 184 ElastiCache background write process, 129–130 lazy loading, 127–128 online cluster resizing, 131–132 overview of, 22, 77, 126–127, 142–145 reserved memory, 130-131 TTL (time to live), 129 write through, 128–129 elasticity Auto Scaling capabilities of, 16-17, 88, 160-161 cooldown periods, 163 target tracking scaling policies, 162 - 163definition of, 6–7, 157 EBS (Elastic Block Store) cost optimization, 218 EBS-optimized instances, 101-103 instance stores compared to, 47 overview of, 28–29 performance of, 101–103

resiliency, 47–51 volumes, 48–51 EC2 (Elastic Compute Cloud) services Auto Scaling, 88 cost optimization, 227–231 HA (high availability) architecture, 85-88 instances, 72–74, 85–88 overview of, 8-10 purpose of, 160 resource access authorization, 181–182 single-tier architecture, 72–74 single-tier architecture, building, 72 - 74ECR (Elastic Container Registry), 12 - 13ECS (Elastic Container Service), 11-12, 160 EFS (Elastic File System) basic configuration, 51–53 cost optimization, 218-219 File Sync, 219 file system creation/deletion, 51-53 overview of, 29, 51 performance of, 103–107 resiliency, 51–53 EKS (Elastic Container Service for Kubernetes), 12–13 Elastic Beanstalk, 14–15 elastic IP addresses, 85, 193, 203-204 Elastic MapReduce, 184 ElastiCache background write process, 129–130 configuration, 142-145 lazy loading, 127–128 online cluster resizing, 131-132 overview of, 22, 77, 126-127 reserved memory, 130–131 TTL (time to live), 129 write through, 128-129

Elastisearch, 77, 241 ELB (Elastic Load Balancing), 16, 62, 159 - 160TCP window scaling, 99 Elastisearch, 77, 241 ELB (Elastic Load Balancing), 16, 62, 159 - 160ENCODE parameter, 135 EFS (Elastic File System), 107 keys KMI (key management infrastructure), 183–184 KMS (Key Management Service), 33, 99, 182–183 storing in cloud, 182–183 endpoints, VPC (Virtual Private Cloud), 194, 204–205 Enterprise Support, 239 ephemeral stores, 47 events, responding to, 241 Everything as a Service (XaaS), 7 Evolve area of operational health, 242 exam preparation chapter-ending review tools, 253 exam modes, 251 exam objectives and structure, 245–247 exam updates, 252, 273–274 memory tables how to use, 252–253 Pearson Test Prep practice test engine offline access, 250–251 online access, 249-250 Premium Edition, 252 updates, 252 suggested study plan, 253 tips and suggestions, 248-249 --exclude filter, 100 expedited retrieval (Glacier), 108, 219 experience, learning from, 242 EXPLAIN statement, 117

F

failovers, 90, 151 Fargate, 13 fault tolerance (FT), 84 File Sync (EFS), 219 file systems EFS (Elastic File System) basic configuration, 51–53 cost optimization, 218–219 File Sync, 219 file system creation/deletion, 51–53 overview of, 29, 51 performance of, 103-107 resiliency, 51–53 NFS (Network File System), 38, 51 firewalls, WAF (Web Application Firewall), 32 Flash Card mode, 251 flexibility. See elasticity FLUSHALL command, 132 FLUSHDB command, 132 FreeableMemory (Aurora), 116 FT (fault tolerance), 84

G

gateways definition of, 193 egress-only, 193, 201–202 gateway endpoints, 205 overview of, 201 Storage Gateway service, 31 General Purpose mode (EFS), 105 General Purpose SSD, 49–51, 103 geolocation routing policy, 151 GET requests, 99 GitHub, 242 Glacier services cost optimization, 219–221 overview of, 30 performance of, 108

resiliency, 53–55 retrieval policies, 55 vault creation, 54–55 vault deletion, 55 retrieval policies, 55 vaults creating and mounting, 54–55 deleting, 55 global infrastructure, 7, 23–24 global secondary indexes, 121, 122–124 Grant Public Read Access to This Object(s) command, 148 Greengrass, 149–150 GROUP BY clause, 134 groups IAM (Identity and Access Management), 171 security, 174 "Guidelines for Media Sanitization", 184

Η

HA (high availability) architecture application services, 88 Availability Zones, 85–87 database services, 88-90 EC2 instances, 85-88 FT (fault tolerance), 84 metrics, 84 monitoring services, 93 networking services, 91–92 overview of, 84-85 S3 (Simple Storage Service), 42 security services, 92-93 Hadoop, 184 hash joins, 117 health, operational, 241 high availability. See HA (high availability) architecture hosted zones, 151 creating, 152-153 deleting, 153 HTTPS, 185

L

IaaS (infrastructure as a service), 7 IAM (Identity and Access Management) capabilities of, 169–171 data store options, 77 groups, 171 HA (high availability) architecture, 92–93 overview of, 31–32 policies, 172-173 resource access authorization, 181–182 roles, 172 users, 171 identities (IAM) groups, 171 policies, 172-173 roles, 172 users, 171 Identity and Access Management. See IAM (Identity and Access Management) identity-based policies, 173 IIS (Internet Information Server), 71 images, AMIs (Amazon Machine Images), 47 indexes, 122-124 infrastructure as a service (IaaS), 7 initialization, EBS (Elastic Block Store), 102innodb_read_only global variable, 115 input/output operations per second (IOPS), 101 instance stores, 47 instances EBS-optimized, 101–103 EC2 (Elastic Compute Cloud) services HA (high availability) architecture, 85 - 88launching, 72-73 provisioning in different Availability Zones, 85–87 terminating, 74, 88 viewing, 73

RIs (reserved instances), 227–228 T2, 115–116 interdependencies, reducing. *See* decoupling interface endpoints, 204–205 interfaces, 193, 198–199 Internet gateways. *See* gateways Internet Information Server (IIS), 71 Internet of Things (IoT), 149–150 I/O requests, 102 IOPS (input/output operations per second), 101 IoT (Internet of Things), 149–150 IP addresses, 85, 193, 203–204

J-K

Key Management Service. See KMS (Key Management Service) kevs AKP (Asynchronous Key Prefetch), 116 KMI (key management infrastructure), 183-184 KMS (Key Management Service), 33, 99, 182-183 storing in cloud, 182-183 KEYS command, 132 Kinesis, 19 KMI (key management infrastructure), 183 - 184KMS (Key Management Service), 33, 99, 182 - 183Kubernetes, EKS (Elastic Container Service for Kubernetes), 12–13

L

Lambda services cost optimization, 230–231 metrics, 232 operational readiness and, 240 overview of, 10–11 latency. See also performant databases; performant storage EBS (Elastic Block Store), 102–103 EFS (Elastic File System), 104–107 S3 (Simple Storage Service), 99–100 latency routing policy, 151 Launch Cost Explorer link, 228 Launch Instance command, 72 lazy loading, 127–128 learning from experience, 242 sharing, 242 Lifecycle Management Glacier services, 220–221 S3 (Simple Storage Service), 216–218 Lightsail, 13–14 LIKE operators, 134 lists, access control, 175 load balancing, ELB (Elastic Load Balancing), 16, 62, 159–160 loading, lazy, 127-128 local secondary indexes, 122–124 Loggly, 241 logic tier, 77 loose coupling, 62, 63 Lua, 132

Μ

Management Console Lifecycle Management, 216–218 SQS (Simple Queue Service) configuration, 63–66 management services CloudTrail API activity tracking, 242 overview of, 34 storing encryption keys in, 182–183 CloudWatch baselines, 241 billing alarms, 230–231

billing alarms, creating, 230–231 BurstCreditBalance metric, 106 capabilities of, 160 data aggregation in, 93 HA (high availability), 93 operational health monitoring, 241 overview of, 34 Trusted Advisor, 33–34, 239 MariaDB, multi-AZ deployments for, 88 master keys storing in cloud, 182–183 Max Memory Used value, 232 media sanitization, 184 memcached, 142 memory allocation, 232. See also caching Memory Size value, 232 memory tables how to use, 252–253 metrics billing, 230 HA (high availability), 84 Lambda, 232 target tracking scaling policies, 162-163 MFA (multifactor authentication), 169-170 Microsoft Active Directory, 33 migration, Database Migration Service, 23 monitoring services, high availability, 93 mounting Glacier vaults, 54–55 Multi-AZ deployments, 88–90 multifactor authentication (MFA), 169 - 170multithreaded replication, 116 multitier architecture advantages of, 74–75 components of, 75-76 three-tier architecture, 76–77 multivalue answer routing policy, 151 mv command, 100

Ν

name servers, authoritative, 150 NAT (network address translation), 194, 205 - 206native Direct Connect, 91-92 negotiations, reduction in, 6 network ACLs (access control lists), 173, 175 network address translation (NAT), 194, 205-206 Network Data Transferred In charge (S3), 216 Network Data Transferred Out charge (S3), 216 Network File System (NFS), 38, 51, 107 network infrastructure ClassicLink, 194, 206 default components, checking, 194–198 DHCP (Dynamic Host Control Protocol) option sets, 193, 202 DNS (Domain Name System), 193, 202-203 elastic IP addresses, 193, 203-204 gateways definition of, 193 egress-only, 193, 201-202 gateway endpoints, 205 overview of, 201 Storage Gateway service, 31 NAT (network address translation), 194, 205–206 network interfaces, 193, 198-199 overview of, 193–194 route tables, 193, 199–200 VPC (Virtual Private Cloud), 185 data store options, 77 endpoints, 194, 204-205 overview of, 24 peering, 194, 206 network interfaces, 193, 198-199

networking services AWS global infrastructure, 23–24 Direct Connect, 25, 91–92, 185 HA (high availability), 91–92 Route 53 service, 26, 150–153, 160 VPC (Virtual Private Cloud), 185 data store options, 77 endpoints, 194, 204-205 overview of, 24 peering, 194, 206 New Relic, 241 NFS (Network File System), 38, 51, 107 NIST "Guidelines for Media Sanitization", 184 nodes, DAX (DynamoDB Accelerator), 146 n-tier architecture. See multitier architecture

0

object storage. See storage services Object-Level Logging, 46 one-tier architecture. See single-tier architecture online cluster resizing (ElastiCache), 131 - 132online resources book companion website, 273–274 Pearson Test Prep practice test engine, 249-251 Operate area of operational excellence, 241 operational excellence design principles, 238, 239–240 Evolve area, 242 Operate area, 241 Prepare area, 239–240 operational health, 241 operational priorities, 239 operational readiness, 240 operators, LIKE, 134

OpEx, 6 OpsWorks, 18 optimization, cost EC2 (Elastic Compute Cloud) services, 227–231 Lambda services, 231–232 Oracle, multi-AZ deployments for, 88

Ρ

PaaS (platform as a service), 7 Patch Manager, 175–176 "pay as you go" model, 6 Payment Card Industry (PCI), 170 PCI (Payment Card Industry), 170 Pearson Test Prep practice test engine offline access, 250-251 online access, 249–250 Premium Edition, 252 updates, 252 peering, VPC (Virtual Private Cloud), 194,206 performance optimization. See caching performant databases Aurora AKP (Asynchronous Key Prefetch), 116 Aurora Serverless, 114 DB connections, showing, 115 hash joins, 117 multithreaded replication, 116 overview of, 20, 114-115 read scaling capabilities, 117 T2 instances, 115–116 TCP keepalive parameters, 117–118 DynamoDB adaptive capacity, 122 burst capacity, 121 query and design patterns, 119-121 query operations, 124-126 scan operations, 124-126 secondary indexes, 122-124

ElastiCache background write process, 129–130 lazy loading, 127–128 online cluster resizing, 131–132 overview of, 126–127 reserved memory, 130-131 TTL (time to live), 129 write through, 128–129 RDS (Relational Database Services), 118 - 119Redshift Amazon Redshift Advisor, 132–135 best practices, 132–134 overview of, 22, 77 performant storage EBS (Elastic Block Store) cost optimization, 218 EBS-optimized instances, 101–103 instance stores compared to, 47 overview of, 28-29 performance of, 101–103 resiliency, 47–51 volumes, 48–51 EFS (Elastic File System) basic configuration, 51–53 Bursting Throughput mode, 105 - 106cost optimization, 218-219 File Sync, 219 file system creation/deletion, 51–53 General Purpose mode, 105 high-level characteristics, 103–104 overview of, 29, 51 performance of, 103–107 performance tips, 107 Provisioned Throughput mode, 105 - 106resiliency, 51-53 Glacier services cost optimization, 219-221 overview of, 30 performance of, 108

resiliency, 53–55 retrieval policies, 55 vault creation, 54–55 vault deletion, 55 S3 (Simple Storage Service) advantages of, 42 bucket creation, 44-46, 147, 216 bucket deletion, 47, 149, 218 cost optimization, 214-218 overview of, 26–28, 77 performance of, 99–101 resiliency, 42-47 permissions, 46, 170 platform as a service (PaaS), 7 policies IAM (Identity and Access Management), 172–173 routing, 151 scaling, 162–163 PostgreSQL, multi-AZ deployments for, 88 Practice Exam mode, 251 practice test engine (Pearson Test Prep) offline access, 250-251 online access, 249–250 Premium Edition, 252 updates, 252 Premium Edition (Pearson Test Prep), 252 preparation, operational excellence, 239-240 Prepare area of operational excellence, 239-240 presentation tier, 76 pre-warming, 102 pricing EBS (Elastic Block Store), 218 EC2 (Elastic Compute Cloud) services billing alarms, 230–231 considerations for, 227–228 Cost Explorer, 228–230 EFS (Elastic File System), 218–219

Glacier services, 219–221 Lambda services, 231–232 S3 (Simple Storage Service), 214–218 priorities, operational, 239 private DNS, 151 Provisioned Capacity (Glacier), 108 Provisioned IOPS SSD, 49–51, 103 Provisioned Throughput mode (EFS), 105–106 purging data, 184

Q

queries, DynamoDB, 124–126 query patterns, DynamoDB, 119–121 queues creating, 63–65 deleting, 66 SQS (Simple Queue Service) configuration, 63–66 overview of, 19

R

RDS (Relational Database Services) best practices, 118-119 HA (high availability) architecture, 88-90 overview of, 20-21, 77 read scaling, 117 readiness, operational, 240 records alias, 149–150 DNS (Domain Name System), 151, 152 - 153Recovery Point Objective (RPO), 84 Recovery Time Objective (RTO), 84 Redis, ElastiCache for background write process, 129–130 lazy loading, 127-128 online cluster resizing, 131–132 overview of, 22, 77, 126–127, 142–145 reserved memory, 130–131

TTL (time to live), 129 write through, 128–129 Redshift Amazon Redshift Advisor, 132–135 best practices, 132-134 overview of, 22, 77 registry, ECR (Elastic Container Registry), 12–13 Relational Database Services. See RDS (Relational Database Services) replication, multithreaded, 116 requests EBS (Elastic Block Store), 102 EFS (Elastic File System), 107 Glacier services, 108 S3 (Simple Storage Service), 99 reserved instances (RIs), 227-228 reserved memory (ElastiCache), 130–131 resharding (ElastiCache), 131–132 resiliency, 39 EBS (Elastic Block Store), 47 instance stores compared to, 47 volume creation, 48–49 volume deletion, 49 volume types, 49-51 EFS (Elastic File System) basic configuration, 51–53 file systems, 51-53 overview of, 51 Glacier services, 53–55 retrieval policies, 55 vault creation, 54-55 vault deletion, 55 S3 (Simple Storage Service) advantages of, 42 bucket creation, 44-46 bucket deletion, 47 capabilities of, 42 classes, 42-44 storage classes, 42-44 resolver (DNS), 150-151

resource access authorization, 181–182 resource-based policies, 173 responses to events, 241 rest, protecting data at, 183-184 retrieval policies, Glacier, 55, 108 reusable delegation sets, 151 RIs (reserved instances), 227–228 roles, IAM (Identity and Access Management), 172 root users, 169 Route 53 service, 26, 150–153, 160 route tables, 193, 199-200 routing policy, 151 RPO (Recovery Point Objective), 84 RTO (Recovery Time Objective), 84

S

S3 (Simple Storage Service). See also Glacier services advantages of, 42 buckets creating, 44-46, 147, 216 deleting, 47, 149, 218 cost optimization, 214 estimated costs, 214-216 Lifecycle Management, 216–218 HA (high availability) architecture, 42 overview of, 26-28, 77 performance of, 99-101 resiliency bucket creation, 44–46 bucket deletion, 47 capabilities of, 42 storage classes, 42-44 S3 Glacier class, 43–44 S3 One Zone-Infrequent Access class, 43-44 S3 Standard class, 42-44 S3 Standard-Infrequent Access class, 43 - 44

SaaS (software as a service), 7 sanitization, media, 184 scalability, 157 Auto Scaling capabilities of, 16-17, 88, 160-161 cooldown periods, 163 target tracking scaling policies, 162–163 ELB (Elastic Load Balancing), 159–160 S3 (Simple Storage Service), 42 TCP windows, 99 scale-in cooldown period, 163 scale-out cooldown period, 163 SCAN command, 132 scans, DynamoDB, 124–126 scheduled RIs (reserved instances), 227 SCPs (Service Control Policies), 173 secondary indexes, 122–124 Secure Sockets Layer (SSL), 185 security groups, 174 security services. See also encryption ACLs (access control lists), 173 data security data at rest, 183-184 data in transit, 185 decommissioning process, 184 encryption keys, 182–183 resource access authorization, 181-182 SSL (Secure Sockets Layer), 185 VPC (Virtual Private Cloud), 185 Directory Services, 33 IAM (Identity and Access Management) capabilities of, 169-171 data store options, 77 groups, 171 HA (high availability) architecture, 92-93 overview of, 31-32 policies, 172-173

resource access authorization, 181-182 roles, 172 users, 171 KMS (Key Management Service), 33, 99, 182–183 network ACLs (access control lists), 175 security groups, 174 Systems Manager Patch Manager, 175 - 176WAF (Web Application Firewall), 32 Security Token Service (STS), 170–171 selective acknowledgement (TCP), 99 Send a Message command, 65 server access logging, 46 Serverless Application Repository, 13 Service Control Policies (SCPs), 173 Service Health dashboard, 241 service-oriented architecture (SOA), 74 sharing learning, 242 Simple Monthly calculator, 214–216 Simple Queue Service (SQS), 19, 63–66 simple routing policy, 151 Simple Storage Service. See S3 (Simple Storage Service) single-tier architecture building with EC2, 72–74 example of, 71-72SMEMBERS command, 132 Snowball, 30 SOA (service-oriented architecture), 74 software as a service (SaaS), 7 SourceForge, 242 Splunk, 241 SQL Server Mirroring, 88 SQS (Simple Queue Service), 19, 63–66 SSCAN command, 132 SSL (Secure Sockets Layer), 185 standard retrieval (Glacier), 108, 219 standard RIs (reserved instances), 227 stdin command, 100-101

stdout command, 100-101 Storage Gateway, 31 storage services EBS (Elastic Block Store) cost optimization, 218 EBS-optimized instances, 101–103 instance stores compared to, 47 overview of, 28-29 performance of, 101–103 resiliency, 47–51 volume creation, 48–49 volume deletion, 49 volume types, 49–51 volumes, 48–51 EFS (Elastic File System) basic configuration, 51-53 Bursting Throughput mode, 105 - 106cost optimization, 218-219 File Sync, 219 file system creation/deletion, 51-53 file systems, 51–53 General Purpose mode, 105 high-level characteristics, 103–104 overview of, 29, 51 performance of, 103–107 performance tips, 107 Provisioned Throughput mode, 105 - 106resiliency, 51–53 Glacier services cost optimization, 219–221 overview of, 30 performance of, 108 resiliency, 53–55 retrieval policies, 55 vault creation, 54–55 vault deletion, 55 S3 (Simple Storage Service) advantages of, 42 bucket creation, 44-46, 147, 216

bucket deletion, 47, 149, 218 classes, 42-44 cost optimization, 214–218 overview of, 26-28, 77 performance of, 99–101 resiliency, 42-47 Snowball, 30 Storage Gateway, 31 Storage Used charged (S3), 215 stores ephemeral, 47 instance, 47 STS (Security Token Services), 171 Study mode, 251 study plan for exam, 253 SumoLogic, 241 sync command, 100 synchronous decoupling, 62 Systems Manager Patch Manager, 175–176 Run Command, 240 Systems Manager Automation, 240

Т

T2 instances, 115–116 tables, route, 193, 199-200 TAMs (Technical Account Managers), 239 target tracking scaling policies, 162–163 TCP (Transmission Control Protocol) keepalive parameters, 117–118 selective acknowledgement, 99 window scaling, 99 Technical Account Managers (TAMs), 239 terminating EC2 instances, 74, 88 three-tier architecture, 76–77 throughput, EFS (Elastic File System), 105 - 106Throughput Optimized HDD, 49-51, 102 time to live (TTL), 129, 151 transit, data protection during, 185

Transmission Control Protocol. See TCP (Transmission Control Protocol) Trusted Advisor, 33–34, 239 TTL (time to live), 129, 151

U

updates, exam, 252, 273–274 updates, software, 175–176 users IAM (Identity and Access Management), 171 root, 169

V

vaults, Glacier creating and mounting, 54–55 deleting, 55 retrieval rate for, 220-221 View Instances command, 73 Virtual Private Cloud. See VPC (Virtual Private Cloud) virtual private networks (VPNs), 91-92, 185 volume queue length, 103 volumes (EBS) creating, 48-49 deleting, 49 types of, 49–51 VPC (Virtual Private Cloud) data security, 185 data store options, 77 endpoints, 194, 204-205 overview of, 24 peering, 194, 206 VPNs (virtual private networks), 91-92, 185

W

WAF (Web Application Firewall), 32 weighted round robin, 151

Well-Architected Framework operational excellence design principles, 237–238, 239–240 Evolve area, 242 Operate area, 241 Prepare area, 239–240 overview of, 237–238 WHERE clause, 134 write through, 128–129 WSCALE factor, 99

X-Y-Z

XaaS (Everything as a Service), 7 X-Ray, 240 zone apex, 150 zones, hosted, 151 creating, 152–153 deleting, 153