EXAM CRAM

AWS Certified SysOps Administrator – Associate

(SOA-C02)

CramSheet

Flash Cards

Practice Tests

MARKO SLUGA
RICK CRISCI
WILLIAM "BO" ROTHWELL

EXAM CRAM

AWS Certified SysOps Administrator – Associate (SOA-C02) Exam Cram

Marko Sluga
Rick Crisci
William “Bo” Rothwell

Pearson

AWS Certified SysOps Administrator – Associate (SOA-C02) Exam Cram

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: 978-0-13-750958-4

ISBN-10: 0-13-750958-8

Library of Congress Cataloging-in-Publication Data: 2022902124

ScoutAutomatedPrintCode

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 authors 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.

Special Sales

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact intlcs@pearson.com.

Editor-in-Chief
Mark Taub

Director, ITP Product Management
Brett Bartow

Executive Editor
Nancy Davis

Managing Editor
Sandra Schroeder

Development Editor
Christopher A. Cleveland

Project Editor
Mandie Frank

Copy Editor
Chuck Hutchinson

Indexer
Ken Johnson

Proofreader
Donna E. Mulder

Technical Editors
Mark Wilkins
Ryan Dymek

Publishing Coordinator
Cindy Teeters

Designer
Chuti Prasertsith

Compositor
codeMantra

Pearson's Commitment to Diversity, Equity, and Inclusion

Pearson is dedicated to creating bias-free content that reflects the diversity of all learners. We embrace the many dimensions of diversity, including but not limited to race, ethnicity, gender, socioeconomic status, ability, age, sexual orientation, and religious or political beliefs.

Education is a powerful force for equity and change in our world. It has the potential to deliver opportunities that improve lives and enable economic mobility. As we work with authors to create content for every product and service, we acknowledge our responsibility to demonstrate inclusivity and incorporate diverse scholarship so that everyone can achieve their potential through learning. As the world's leading learning company, we have a duty to help drive change and live up to our purpose to help more people create a better life for themselves and to create a better world.

Our ambition is to purposefully contribute to a world where

- Everyone has an equitable and lifelong opportunity to succeed through learning
- Our educational products and services are inclusive and represent the rich diversity of learners
- Our educational content accurately reflects the histories and experiences of the learners we serve
- Our educational content prompts deeper discussions with learners and motivates them to expand their own learning (and worldview)

While we work hard to present unbiased content, we want to hear from you about any concerns or needs with this Pearson product so that we can investigate and address them.

- Please contact us with concerns about any potential bias at https://www.pearson.com/report-bias.html.

Credits

Figures	Credit
Figures 2.1–2.3, 4.2–4.7, 4.10–4.13, 5.1, 5.3, 5.4, 6.1–6.5, 7.1–7.3, 9.1–9.9, 10.1, 11.1–11.9, 11.11, 12.2, 12.3, 12.8–12.17, 13.1–13.7, 14.1–14.8, 15.3–15.5	Amazon Web Services, Inc
Figure 4.1	Medium
Figure 5.2	DeveloperCK
Figures 10.2, 10.3	Google LLC

Contents at a Glance

Table of Contents

CHAPTER 7

CHAPTER 8

CHAPTER 12

CHAPTER 13

CHAPTER 14

CHAPTER 15

About the Authors

Marko Sluga has more than 20 years of experience in IT and has had the benefit of witnessing the rise of cloud computing. Marko has worked on a variety of cloud-related projects, from the early stages of SOC, corporate virtualization, and open-source API offerings to modern, fully automated, intelligent, serverless, and cloud-native solutions. Marko has been working with Amazon Web Services (AWS) since the start of the 2010s and holds three associate, two professional, and three specialty AWS certifications. Marko performs training and advising on cloud technologies and strategies, DevOps, and IT system and process optimization to clients from a wide range of companies, including startups, SMBs, enterprise businesses, and Fortune 500 companies. Marko runs his own cloud training, coaching, and consulting practice under the markocloud.com brand. He is the author of the *AWS Certified Developer - Associate (DVA-C01) Cert Guide*.

Rick Crisci is the founder and principal trainer for Trainertests.com and is an experienced AWS and VMware instructor. Original content by Rick can be found on a variety of platforms including Pearson, LinkedIn Learning, and Udemy, with more than 30 courses, over 250,000 students, and exceptionally high course review scores. Rick also teaches live hands-on AWS courses regularly for Pearson on O'Reilly.

Prior to becoming an instructor, Rick had over 15 years of real-world experience. Some career highlights include designing high-speed Internet networks in the early 2000s and managing virtualization and networking teams for a financial institution.

In 2017, VMware recognized Rick as the first runner-up for Instructor of the Year for the Americas. This honor was largely due to the glowing feedback from his students. Rick seeks to help students learn by taking complex concepts and providing clear and simple explanations, diagrams, and analogies.

At the impressionable age of 14, **William "Bo" Rothwell** crossed paths with a TRS-80 Micro Computer System (affectionately known as a "Trash 80"). Soon after, the adults responsible for Bo made the mistake of leaving him alone with the TRS-80. He immediately dismantled it and held his first computer class, showing his friends what made this "computer thing" work. Since that experience, Bo's passion for understanding how computers work and sharing this knowledge with others has resulted in a rewarding career in IT training. His experience includes Cloud, Linux, UNIX, IT security, DevOps, and programming languages such as Perl, Python, Tcl, and BASH. He is the founder and lead instructor of One Course Source, an IT training organization.

Dedication

I would like to dedicate this book to my wife and children for their continued support of projects like this book.
—Marko Sluga

I would like to dedicate this book to my wife, Jessica, whose constant dedication to helping her students is always an inspiration, and to my children, who give me a purpose to work for. Also, to my mother for giving me confidence to pursue big goals, and to my father who showed me the value of hard work, dedication, and always putting his family before himself.
—Rick Crisci

Normally, I use this space to thank my wife, daughter, and parents. While they continue to be my biggest supporters, I feel compelled to dedicate this book to those who embrace lifelong learning. Thank you all for making this book a small part of your learning goals.

"There is no end to education. It is not that you read a book, pass an examination, and finish with education. The whole of life, from the moment you are born to the moment you die, is a process of learning."
—Jiddu Krishnamurti

—William "Bo" Rothwell

Acknowledgments

I need to thank Rick and Bo for helping make this book a reality. I would also like to thank Nancy Davis and Chris Cleveland for their generous support during the creation of this title.

—Marko Sluga

Thanks to Marko for giving me the opportunity to help with this book, and to Nancy Davis for keeping this thing going. I'm appreciative of the opportunity to work with such wonderful people.

—Rick Crisci

Thanks to the entire Pearson team for all of the gentle pushes, patience, and dedication to making this book a success.

—William "Bo" Rothwell

About the Technical Reviewers

Mark Wilkins is an electronic engineering technologist with a wealth of experience in designing, deploying, and supporting software and hardware technology in the corporate and small business world. Since 2013, Mark has focused on supporting and designing cloud service solutions with Amazon Web Services, Microsoft Azure, and the IBM Cloud. He is certified in Amazon Web Services (Architecture and SysOps). Mark is also a Microsoft Certified Trainer (MCT) and holds certifications in MCTS, MCSA, Server Virtualization with Windows Server Hyper-V, and Azure Cloud Services. Mark currently develops AWS curriculum on technical aspects of AWS architecture for O'Reilly Media, Pluralsight, and LinkedIn Learning. His published books include *Learning Amazon Web Services*, *AWS Certified Solutions Architect - Associate (SAA-C02) Cert Guide*, *Windows 2003 Registry for Dummies*, and *Administering SMS 3.0*, *Administering Active Directory*.

As a consultant, **Ryan Dymek** has been building, designing, and improving cloud solutions for some of the largest companies in the world for more than 10 years. Ryan began working with AWS cloud solutions in 2009 and in 2016 added GCP to his portfolio. Ryan transforms teams and organizations using DevOps principles and solid architectural approaches through a model of culture and "people first" philosophies. Ryan has trained more than 10,000 people over the past 6 years on topics such as DevOps engineering, security, cost optimization, performance tuning, and operational excellence.

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 or write 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: community@informit.com

Introduction

Welcome to *AWS Certified SysOps Administrator – Associate (SOA-C02) Exam Cram*. This book will help you get ready to take and pass the AWS Certified SysOps Administrator – Associate (SOA-C02) exam.

This book is designed to remind you of everything you need to know to pass the SOA-C02 certification exam. Each chapter includes a number of practice questions that should give you a reasonably accurate assessment of your knowledge, and, yes, we've provided the answers and their explanations for these questions. Read this book, understand the material, and you'll stand a very good chance of passing the real test.

Exam Cram books help you understand and appreciate the subjects and materials you need to know to pass AWS certification exams. *Exam Cram* books are aimed strictly at test preparation and review. They do not teach you everything you need to know about a subject. Instead, the authors streamline and highlight the pertinent information by presenting and dissecting the questions and problems they've discovered that you're likely to encounter on an AWS certification exam.

Let's begin by looking at preparation for the exam.

How to Prepare for the Exam

This text follows the official exam objectives closely to help ensure your success. The AWS Certified SysOps Administrator – Associate exam covers 6 domains and 14 objectives, and this book is aligned with those domains and objectives. These official objectives from AWS can be found here:

> https://d1.awsstatic.com/training-and-certification/docs-sysops-associate/AWS-Certified-SysOps-Administrator-Associate_Exam-Guide.pdf

As you examine the numerous exam topics now covered in the exam, resist the urge to panic! This book will provide you with the knowledge (and confidence) that you need to succeed. You just need to make sure you read it and follow the guidance it provides throughout your SysOps Administrator – Associate journey.

Practice Tests

This book is filled with practice exam questions to get you ready!

- **CramSaver questions at the beginning of each major topic in each chapter:** These difficult, open-ended questions ensure you really know the material. Some readers use these questions to “test out” of a particular topic.
- **CramQuizzes at the end of each chapter:** These quizzes provide another chance to demonstrate your knowledge after completing a chapter.

In addition, the book comes with two full practice tests in the Pearson Test Prep software available to you either online or as an offline Windows application. To access the practice exams, please see the instructions in the card inserted in the sleeve in the back of the book. This card includes a unique access code that enables you to activate your exams in the Pearson Test Prep software.

If you are interested in more practice exams than are provided with this book, Pearson IT Certification publishes a Premium Edition eBook and Practice Test product. In addition to providing you with three eBook files (EPUB, PDF, and Kindle), this product provides you with two additional exams’ worth of questions. The Premium Edition version also offers you a link to the specific section in the book that presents an overview of the topic covered in the question, allowing you to easily refresh your knowledge. The insert card in the back of the book includes a special offer for an 80 percent discount off this Premium Edition eBook and Practice Test product, which is an incredible deal.

Taking a Certification Exam

After you prepare for your exam, you need to register with a testing center. At the time of this writing, the cost to take the AWS Certified SysOps Administrator – Associate exam is $150 USD for individuals. The test is administered by Pearson VUE testing centers with locations globally or as an online proctored exam.

You will have 180 minutes to complete the exam. The exam consists of a maximum of 65 questions. If you have prepared, you should find that this is plenty of time to properly pace yourself and review the exam before submission.

Arriving at the Exam Location

As with any examination, arrive at the testing center early (at least 15 minutes). Be prepared! You need to bring two forms of identification (one with a picture). The testing center staff requires proof that you are who you say you are and that someone else is not taking the test for you. Arrive early, because if you are late, you will be barred from entry and will not receive a refund for the cost of the exam.

ExamAlert

You'll be spending a lot of time in the exam room. Plan on using the full 180 minutes allotted for your exam and surveys. Policies differ from location to location regarding bathroom breaks. Check with the testing center before beginning the exam.

In the Testing Center

You will not be allowed to take into the examination room study materials or anything else that could raise suspicion that you're cheating. This includes practice test material, books, exam prep guides, or other test aids. The Testing Center will provide you with scratch paper and a pen or pencil. These days, this often comes in the form of an erasable whiteboard.

Examination results are available after the exam. After submitting the exam, you will be notified whether you have passed or failed. The test administrator will also provide you with a printout of your results.

About This Book

The ideal reader for an *Exam Cram* book is someone seeking certification. However, it should be noted that an *Exam Cram* book is a very easily readable, rapid presentation of facts. Therefore, an *Exam Cram* book is also extremely useful as a quick reference manual.

The book can be read cover to cover, or you may jump across chapters as needed. Because the book chapters align with the exam objectives, some chapters may have slight overlap on topics. Where required, references to the other chapters are provided for you. If you need to brush up on a topic, you can use the index, table of contents, or Table I.1 to find the topics and go to the questions that you need to study. Beyond helping you prepare for the test, we think you'll find this book useful as a tightly focused reference on some of the most important aspects of the AWS Certified SysOps Administrator – Associate (SOA-C02) certification.

This book includes other helpful elements in addition to the actual logical, step-by-step learning progression of the chapters themselves. *Exam Cram* books use elements such as ExamAlerts, tips, notes, and practice questions to make information easier to read and absorb. This text also includes a very helpful glossary to assist you.

Note

Reading this book from start to finish is not necessary; this book is set up so that you can quickly jump back and forth to find sections you need to study.

Use the *CramSheet* to remember last-minute facts immediately before the exam. Use the practice questions to test your knowledge. You can always brush up on specific topics in detail by referring to the table of contents and the index. Even after you achieve certification, you can use this book as a rapid-access reference manual.

Exam Objectives

Table I.1 lists the skills the AWS Certified SysOps Administrator – Associate (SOA-C02) exam measures and the chapter in which the objective is discussed.

TABLE I.1

Exam Domain	Objective	Chapter in Book That Covers It
1.0 Monitoring, Logging, and Remediation	1.1 Implement metrics, alarms, and filters by using AWS monitoring and logging services	Chapters 1, 2, 3
1.0 Monitoring, Logging, and Remediation	1.2 Remediate issues based on monitoring and availability metrics	Chapters 1, 2, 3
2.0 Reliability and Business Continuity	2.1 Implement scalability and elasticity	Chapters 1, 4, 5, 6
2.0 Reliability and Business Continuity	2.2 Implement high availability and resilient environments	Chapters 1, 4, 5, 6
2.0 Reliability and Business Continuity	2.3 Implement backup and restore strategies	Chapters 1, 4, 5, 6
3.0 Deployment, Provisioning, and Automation	3.1 Provision and maintain cloud resources	Chapters 1, 4, 7, 8
3.0 Deployment, Provisioning, and Automation	3.2 Automate manual or repeatable processes	Chapters 1, 4

Exam Domain	Objective	Chapter in Book That Covers It
4.0 Security and Compliance	4.1 Implement and manage security and compliance policies	Chapters 1, 2, 3, 6, 9
4.0 Security and Compliance	4.2 Implement data and infrastructure protection strategies	Chapters 1, 2, 3, 6, 10
5.0 Networking and Content Delivery	5.1 Implement networking features and connectivity	Chapters 1, 4, 5, 11
5.0 Networking and Content Delivery	5.2 Configure domains, DNS services, and content delivery	Chapters 1, 4, 5, 12
5.0 Networking and Content Delivery	5.3 Troubleshoot network connectivity issues	Chapters 1, 4, 5, 13
6.0 Cost and Performance Optimization	6.1 Implement cost optimization strategies	Chapters 1, 14
6.0 Cost and Performance Optimization	6.2 Implement performance optimization strategies	Chapter 15

The Chapter Elements

Each *Exam Cram* book has chapters that follow a predefined structure. This structure makes *Exam Cram* books easy to read and provides a familiar format for all *Exam Cram* books. The following elements typically are used:

- Chapter topics
- CramSavers
- CramQuizzes
- ExamAlerts
- Notes
- Available exam preparation software practice questions and answers

Note

Bulleted lists, numbered lists, tables, and graphics are also used where appropriate. A picture can paint a thousand words sometimes, and tables can help to associate different elements with each other visually.

Now let's look at each of the elements in detail.

- **Chapter topics**—Each chapter contains details of all subject matter listed in the table of contents for that particular chapter. The objective of an

Exam Cram book is to cover all the important facts without giving too much detail; it is an exam cram.

- **CramSavers**—Each chapter kicks off with a short-answer quiz to help you assess your knowledge of the chapter topic. This chapter element is designed to help you determine whether you need to read the whole chapter in detail or merely skim the material and skip ahead to the CramQuiz at the end of the chapter.
- **CramQuizzes**—Each chapter concludes with a multiple-choice quiz to help ensure that you have gained familiarity with the chapter content.
- **ExamAlerts**—ExamAlerts address exam-specific, exam-related information. An ExamAlert addresses content that is particularly important, tricky, or likely to appear on the exam. An ExamAlert looks like this:

ExamAlert

Make sure you remember the different ways in which you can access a router remotely. Know which methods are secure and which are not.

- **Notes**—Notes typically contain useful information that is not directly related to the current topic under consideration. To avoid breaking up the flow of the text, they are set off from the regular text.

Note

This is a note. You have already seen several notes.

Other Book Elements

Most of this *Exam Cram* book on SysOps Administrator - Associate follows the consistent chapter structure already described. However, various important elements are not part of the standard chapter format. These elements apply to the book as a whole.

- **Glossary**—The glossary contains a listing of important terms used in this book with explanations.
- **CramSheet**—The CramSheet is a quick-reference, tear-out cardboard sheet of important facts useful for last-minute preparation. CramSheets often include a simple summary of the facts that are most difficult to remember.

- **Companion website**—The companion website for your book allows you to access several digital assets that come with your book, including
  - Pearson Test Prep software (both online and Windows desktop versions)
  - Key Terms Flash Cards application
  - A PDF version of the CramSheet

To access the book's companion website, simply follow these steps:

1. Register your book by going to: PearsonITCertification.com/ register and entering the ISBN: **9780137509584**.
2. Respond to the challenge questions.
3. Go to your account page and select the **Registered Products** tab.
4. Click the **Access Bonus Content** link under the product listing.

Pearson Test Prep Practice Test Software

As noted previously, this book comes complete with the Pearson Test Prep practice test software containing two full exams. These practice tests are available to you either online or as an offline Windows application. To access the practice exams that were developed with this book, please see the instructions in the card inserted in the sleeve in the back of the book. This card includes a unique access code that enables you to activate your exams in the Pearson Test Prep software.

Accessing the Pearson Test Prep Software Online

The online version of this software can be used on any device with a browser and connectivity to the Internet, including desktop machines, tablets, and smartphones. To start using your practice exams online, simply follow these steps:

1. Go to http://www.PearsonTestPrep.com.
2. Select **Pearson IT Certification** as your product group.
3. Enter your email/password for your account. If you don't have an account on PearsonITCertification.com or CiscoPress.com, you will need to establish one by going to PearsonITCertification.com/join.

4. In the **My Products** tab, click the **Activate New Product** button.
5. Enter the access code printed on the insert card in the back of your book to activate your product.
6. The product will now be listed in your My Products page. Click the **Exams** button to launch the exam settings screen and start your exam.

Accessing the Pearson Test Prep Software Offline

If you wish to study offline, you can download and install the Windows version of the Pearson Test Prep software. There is a download link for this software on the book's companion website, or you can just enter this link in your browser:

http://www.pearsonitcertification.com/content/downloads/pcpt/engine.zip

To access the book's companion website and the software, simply follow these steps:

1. Register your book by going to PearsonITCertification.com/register and entering the ISBN: **9780137509584**.
2. Respond to the challenge questions.
3. Go to your account page and select the **Registered Products** tab.
4. Click the **Access Bonus Content** link under the product listing.
5. Click the **Install Pearson Test Prep Desktop Version** link under the Practice Exams section of the page to download the software.
6. After the software finishes downloading, unzip all the files on your computer.
7. Double-click the application file to start the installation, and follow the on-screen instructions to complete the registration.
8. When the installation is complete, launch the application and select the **Activate Exam** button on the My Products tab.
9. Click the **Activate a Product** button in the Activate Product Wizard.
10. Enter the unique access code found on the card in the sleeve in the back of your book and click the **Activate** button.
11. Click **Next** and then the **Finish** button to download the exam data to your application.

12. You can now start using the practice exams by selecting the product and clicking the **Open Exam** button to open the exam settings screen.

Note that the offline and online versions will synch together, so saved exams and grade results recorded on one version will be available to you on the other as well.

Customizing Your Exams

Once you are in the exam settings screen, you can choose to take exams in one of three modes:

- Study Mode
- Practice Exam Mode
- Flash Card Mode

Study Mode allows you to fully customize your exams and review answers as you are taking the exam. This is typically the mode you would use first to assess your knowledge and identify information gaps. Practice Exam Mode locks certain customization options because it is presenting a realistic exam experience. Use this mode when you are preparing to test your exam readiness. Flash Card Mode strips out the answers and presents you with only the question stem. This mode is great for late stage preparation when you really want to challenge yourself to provide answers without the benefit of seeing multiple-choice options. This mode will not provide the detailed score reports that the other two modes will, so it should not be used if you are trying to identify knowledge gaps.

In addition to these three modes, you will be able to select the source of your questions. You can choose to take exams that cover all of the chapters, or you can narrow your selection to just a single chapter or the chapters that make up specific parts in the book. All chapters are selected by default. If you want to narrow your focus to individual chapters, simply deselect all the chapters and then select only those on which you wish to focus in the Objectives area.

You can also select the exam banks on which to focus.

There are several other customizations you can make to your exam from the exam settings screen, such as the time of the exam, the number of questions served up, whether to randomize questions and answers, whether to show the number of correct answers for multiple-choice questions, or whether to serve up only specific types of questions. You can also create custom test banks by selecting only questions that you have marked or questions on which you have added notes.

Updating Your Exams

If you are using the online version of the Pearson Test Prep software, you should always have access to the latest version of the software as well as the exam data. If you are using the Windows desktop version, every time you launch the software, it will check to see if there are any updates to your exam data and automatically download any changes that were made since the last time you used the software. This requires that you are connected to the Internet at the time you launch the software.

Sometimes, due to many factors, the exam data may not fully download when you activate your exam. If you find that figures or exhibits are missing, you may need to manually update your exams.

To update a particular exam you have already activated and downloaded, simply select the **Tools** tab and select the **Update Products** button. Again, this is an issue only with the desktop Windows application.

If you wish to check for updates to the Pearson Test Prep exam engine software, Windows desktop version, simply select the **Tools** tab and select the **Update Application** button. This will ensure you are running the latest version of the software engine.

Contacting the Authors

Hopefully, this book provides you with the tools you need to pass the AWS SysOps Administrator - Associate exam. Feedback is appreciated. You can follow and contact the authors on LinkedIn:

Marko Sluga:

https://www.linkedin.com/in/markosluga/

Rick Crisci:

https://www.linkedin.com/in/rickcrisci/

William "Bo" Rothwell:

https://www.linkedin.com/in/bo-rothwell/

This page intentionally left blank

CHAPTER 4

Implementing Scalability and Elasticity

This chapter covers the following official AWS Certified SysOps Administrator - Associate (SOA-C02) exam domains:

- Domain 2: Reliability and Business Continuity
- Domain 3: Deployment, Provisioning, and Automation
- Domain 5: Networking and Content Delivery

(For more information on the official AWS Certified SysOps Administrator - Associate [SOA-C02] exam topics, see the Introduction.)

Ensuring your application's infrastructure is scalable and elastic delivers a double benefit for the application. First, by adding more resources dynamically when required, you can adapt to any amount of traffic to ensure you do not leave any request unanswered. Second, by removing resources, you ensure the application is cost-effective when little or no requests are being received. However, designing a scalable/elastic application is not always an easy task.

In this chapter, we examine scaling, request offloading, and loose coupling as strategies that can enable an application to meet demand while maintaining cost-effectiveness. Ensuring your application is scalable and elastic also builds a good underlying foundation to achieve high availability and resilience, which we discuss in Chapter 5, "High Availability and Resilience."

Scaling in the Cloud

This section covers the following official AWS Certified SysOps Administrator - Associate (SOA-C02) exam domains:

- Domain 2: Reliability and Business Continuity
- Domain 3: Deployment, Provisioning, and Automation

CramSaver

If you can correctly answer these questions before going through this section, save time by skimming the Exam Alerts in this section and then completing the Cram Quiz at the end of the section.

1. You are operating a forum application that consists of three layers: a web front end on EC2, an application on EC2, and a database layer RDS. The web front end delivers the static forum content and formatting, the application layer stores the session information for each user, and the database layer stores all the user preferences. Assess the scalability of this deployment, identify any issues, and propose a solution.
2. You have been tasked with troubleshooting an image-processing platform. The application resides on a single-layer ECS container deployment that accepts requests for image processing from an incoming S3 bucket and deposits the processed image in an output S3 bucket. Lately, a spike in usage has caused the ECS application to reach its maximum scale. Users using the paid platform have reported that bulk image uploads complete successfully to S3, but some images are never processed. As a result, users are left searching for unprocessed images and need to resubmit them for processing. How could you ensure that the application works as intended?
3. Your developers have updated the forum application as per your previous comments. Your application is now growing, and you have been tasked with ensuring the application maintains scalability to millions of users. To assess the scalability, you have been given more information on the deployment. The web front end uses Apache2 HTTPS servers on Ubuntu Linux on EC2. The application layer runs custom Python code on EC2 that connects to a DynamoDB table to store session data. The database layer is deployed on a Multi-AZ RDS MySQL cluster with two instances (primary and secondary). Assess the scalability of this deployment and identify any potential bottlenecks.
4. After you optimize the application, your clients report highly improved performance. After receiving the latest AWS bill, the CFO has questions about additional cost of the application. While examining the cost report, you find that the application layer seems to still be deployed with a static number of servers like it was before the session state was offloaded. How can you further optimize the application to reduce the cost?

Answers

1. Answer: Both the web and database layers are scalable. The application layer is limited in elasticity due to the persistence of the session data on the EC2 instances. Session data should be moved off the EC2 instances.
2. Answer: The bulk image uploads seem to exceed the capacity of the ECS cluster. The application needs to be redesigned with a buffer for the image-processing requests. Implementing a message queue service could offload the requests so that the back end can process them in a more predictable manner.
3. Answer: The only issue to identify is with the database layer. A Multi-AZ RDS deployment is only vertically scalable with an upper limit of the maximum size of the RDS instance. The maximum size of the instance could potentially bottleneck the whole forum application.
4. Answer: Find a good metric on which to scale the application layer and implement AWS Autoscaling. The application should shut down some of the instances when usage is low and power them on when traffic increases.

For any application to be made elastic and scalable, you need to consider the sum of the configurations of its components. Any weakness in any layer or service that the application depends on can cause the application scalability and elasticity to be reduced. Any reduction in scalability and elasticity can potentially introduce weaknesses in the high availability and resilience of your application. At the end of the day, any poorly scalable, rigid application with no guarantee of availability can have a tangible impact on the bottom line of any business.

The following factors need to be taken into account when designing a scalable and elastic application:

- **Compute layer:** The compute layer receives a request and responds to it. To make an application scalable and elastic, you need to consider how to scale the compute layer. Can you scale on a metric-defined amount of CPU, memory, and number of connections, or do you need to consider scaling at the smallest scale—per request? Also consider the best practice for keeping the compute layer disposable. There should never be any persistent data within the compute layer.
- **Persistent layer:** Where is the data being generated by the application stored? Is the storage layer decoupled from the instances? Is the same (synchronous) data available to all instances, or do you need to account for asynchronous platforms and eventual consistency? You should always ensure that the persistent layer is designed to be scalable and elastic and take into account any issues potentially caused by the replication configuration.

- **Decoupled components:** Are you scaling the whole application as one, or is each component or layer of the application able to scale independently? You need to always ensure each layer or section of the application can scale separately to achieve maximum operational excellence and lowest cost.
- **Asynchronous requests:** Does the compute platform need to process every request as soon as possible within the same session, or can you schedule the request to process it at a later time? When requests are allowed to process for a longer amount of time (many seconds, perhaps even minutes or hours), you should always decouple the application with a queue service to handle any requests asynchronously—meaning at a later time. Using a queue can enable you to buffer the requests, ensuring you receive all the requests on the incoming portion of the application and handle the processing with predictable performance on the back end. A well-designed, asynchronously decoupled application should almost never respond with a 500-type HTTP (service issue) error.

Assessing your application from these points of view should give you a rough idea of the scalability and elasticity of the platform. When you have a good idea of the scalability/elasticity, also consider any specific metrics within the defined service-layer agreement (SLA) of the application. After both are defined, assess whether the application will meet the SLA in its current configuration. Make a note if you need to take action to improve the scalability/elasticity and continuously reassess because both the application requirements and defined SLA of the application are likely to change over time.

After the application has been designed to meet the defined SLA of the application, you can make use of the cloud metrics provided in the platform at no additional cost to implement automated scaling and meet the demand in several different ways. We discuss how to implement automation in the AWS Autoscaling later in this section.

ExamAlert

Remember, one of the crucial factors that enables scalability and elasticity is ensuring your resources are disposable. That means any data is always written outside the processing layer. All databases, files, logs, and any kind of output the application generates should always be decoupled from the processing layer. In the exam, different services might be part of the proposed solution. When analyzing the proposed answer, always ensure the data being output by the application doesn't stay on the EC2 instance for a longer time and that the solution has the lowest cost. For example, one answer could propose that logs be sent to CloudWatch logs via the CloudWatch agent, whereas another answer might propose logs be written to S3 every hour. Although the second solution will probably be more cost-effective, you should consider the first solution if the defined SLA of the application requires that all logs need to be captured.

Horizontal vs. Vertical Scaling

The general consensus is that there are only two ways (with minor variance, depending on the service or platform) to scale a service or an application:

- Vertically, by adding more power (more CPU, memory, disk space, network bandwidth) to an application instance
- Horizontally, by adding more instances to an application layer, thus increasing the power by a factor of the size of the instance added

A great benefit to vertical scaling is that it can be deployed in any circumstance, even without any application support. Because you are maintaining one unit and increasing its size, you can vertically scale to the maximum size of the unit in question. The maximum scaling size of an instance is defined by the maximum size that the service supports. For example, at the time of writing, the maximum instance size supported on EC2 (u-24tb1.metal) offers 448 CPU cores and 24 TB (that's 24,576 GB!) of memory, while the smallest still-supported size (t2.nano) has only 1 CPU core and 0.5 GB of memory. Additionally, there are plenty of instance types and sizes to choose from, which means that you can horizontally scale an application the exact size you need at a certain moment in time. This same fact applies to other instance-based services such as EMR, RDS, and DocumentDB. Figure 4.1 illustrates vertical scaling of instances.

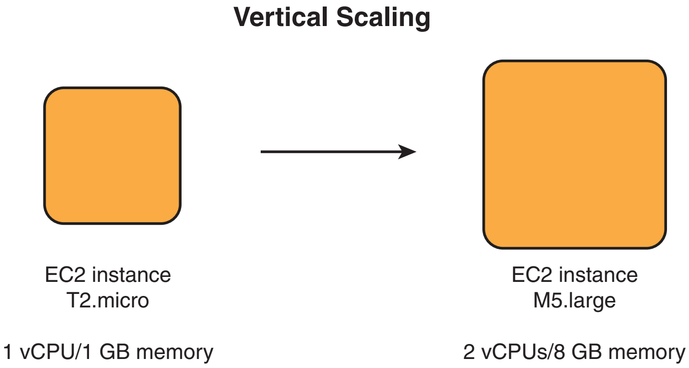


FIGURE 4.1 Vertical scaling

However, when thinking about scalability, you have to consider the drawbacks of vertical scaling. We mentioned maximum size, but one other major drawback is what makes horizontal scaling impractical—a single instance. Because all of AWS essentially operates on EC2 as the underlying design, you can take EC2 as a great example of why a single instance is not the way to go. Each instance you deploy is deployed on a hypervisor in a rack in a datacenter, and this datacenter can only ever be part of one availability zone. In Chapter 1, "Introduction to AWS," we defined an availability zone as a fault isolation environment—meaning any failure, whether it is due to a power or network outage or even an earthquake or flood, is always isolated to one availability zone. Although a single instance is vertically scalable, it is by no means highly available, nor does the vertical scaling make the application very elastic, because every time you scale to a different instance type, you need to reboot the instance.

This is where horizontal scaling steps in. With horizontal scaling, you add more instances (scale-out) when traffic to your application increases and remove instances (scale-in) when traffic to your application is reduced. You still need to select the appropriate scaling step, which is, of course, defined by the instance size. When selecting the size of the instances in a scaling environment, always ensure that they can do the job and don't waste resources. Figure 4.2 illustrates horizontal scaling.

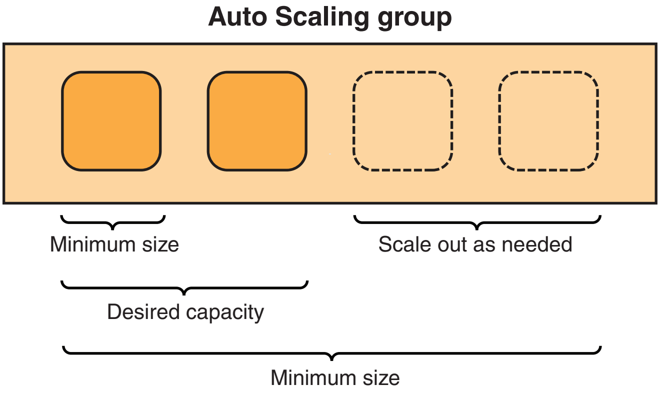
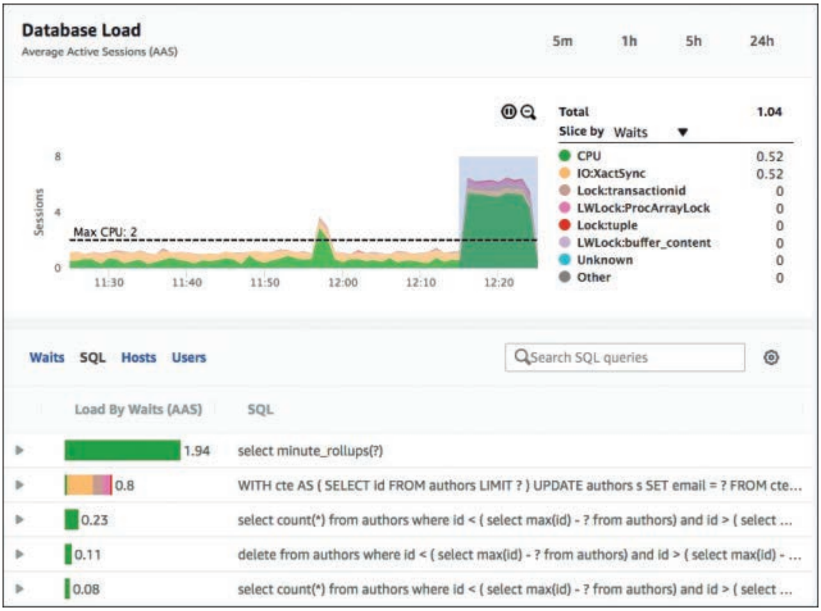


FIGURE 4.2 **Horizontal scaling**

In the ideal case, the application is stateless, meaning it does not store any data in the compute layer. In this case, you can easily scale the number of instances instead of scaling one instance up or down. The major benefit is that you can scale across multiple availability zones, thus inherently achieving high availability as well as elasticity. This is why the best practice on AWS is to create stateless, disposable instances and decouple the processing from the data and the layers of the application from each other. However, a potential drawback of horizontal scaling is when the application does not support it. The reality is that you will sooner or later come upon a case where you need to support an "enterprise" application, being migrated from some virtualized infrastructure, that simply does not support adding multiple instances to the mix. In this case you can still use the services within AWS to make the application highly available and recover it automatically in case of a failure. For such instances you can now utilize the AWS EC2 Auto Recovery service, for the instance types that support it, which automatically re-creates the instance in case of an underlying system impairment or failure.

Another potential issue that can prevent horizontal scalability is the requirement to store some data in a stateful manner. Thus far, we have said that you need to decouple the state of the application from the compute layer and store it in a back-end service—for example, a database or in-memory service. However, the scalability is ultimately limited by your ability to scale those back-end services that store the data. In the case of the Relational Database Service (RDS), you are always limited to one primary instance that handles all writes because both the data and metadata within a traditional database need to be consistent at all times. You can scale the primary instance vertically; however, there is a maximum limit the database service will support, as Figure 4.3 illustrates.



Total	1.04
Slice by Waits ▼	
● CPU	0.52
● IO:XactSync	0.52
● Lock:transactionid	0
● LWLock:ProcArrayLock	0
● Lock:tuple	0
● LWLock:buffer_content	0
● Unknown	0
● Other	0

	Load By Waits (AAS)	SQL
▶	1.94	select minute_rollups(?)
▶	0.8	WITH cte AS (SELECT id FROM authors LIMIT ?) UPDATE authors s SET email = ? FROM cte...
▶	0.23	select count(*) from authors where id < (select max(id) - ? from authors) and id > (select ...
▶	0.11	delete from authors where id < (select max(id) - ? from authors) and id > (select max(id) - ...
▶	0.08	select count(*) from authors where id < (select max(id) - ? from authors) and id > (select ...

FIGURE 4.3 **Database bottlenecking the application**

You can also create a Multi-AZ deployment, which creates a secondary, synchronous replica of the primary database in another availability zone; however, Multi-AZ does not make the application more scalable because the replica is inaccessible to any SQL operations and is provided for the sole purpose of high availability. Another option is adding read replicas to the primary instance to offload read requests. We delve into more details on read replicas later in this chapter and discuss database high availability in Chapter 5.

> **ExamAlert**
>
> If the exam question is ambiguous about scalability or elasticity, you should still consider these as an important requirement of the application. Unless specific wording indicates cost as the primary or only driver, always choose the answer with the solution that will scale and is designed with elastic best practices in mind.

AWS Autoscaling

After you design all the instance layers to be scalable, you should take advantage of the AWS Autoscaling service to automate the scale-in and scale-out operations for your application layers based on performance metrics—for example, EC2 CPU usage, network capacities, and other metrics captured in the CloudWatch service.

The AutoScaling service can scale the following AWS services:

- **EC2:** Add or remove instances from an EC2 AutoScaling group.
- **EC2 Spot Fleets:** Add or remove instances from a Spot Fleet request.
- **ECS:** Increase or decrease the number of containers in an ECS service.
- **DynamoDB:** Increase or decrease the provisioned read and write capacity.
- **RDS Aurora:** Add or remove Aurora read replicas from an Aurora DB cluster.

To create an autoscaling configuration on EC2, you need the following:

- **EC2 Launch template:** Specifies the instance type, AMI, key pair, block device mapping, and other features the instance should be created with.
- **Scaling policy:** Defines a trigger that specifies a metric ceiling (for scaling out) and floor (for scaling in). Any breach of the floor or ceiling for a certain period of time triggers autoscaling.
- **EC2 AutoScaling group:** Defines scaling limits and the minimum, maximum, and desired numbers of instances. You need to provide a launch configuration and a scaling policy to apply during a scaling event.

Dynamic Scaling

Traditionally, scaling policies have been designed with dynamic scaling in mind. For example, a common setup would include

- An AutoScaling group with a minimum of 1 and a maximum of 10 instances
- A CPU % ceiling of 70 percent for scale-out
- A CPU % floor of 30 percent for scale-in
- A breach duration of 10 minutes
- A scaling definition of +/– 33 percent capacity on each scaling event

The application is now designed to operate at a particular scale between 30 and 70 percent aggregate CPU usage of the AutoScaling group. After the ceiling is breached for 10 minutes, the Autoscaling service adds a third more instances to the AutoScaling group. If you are running one instance, it adds another because it needs to meet 33 percent or more of the capacity. If you are running two instances, it also adds one more; however, at three instances, it needs to add two more instances to meet the rules set out in the scaling policy. When the application aggregate CPU usage falls below 30 percent for 10 minutes, the AutoScaling group is reduced by 33 percent, and the appropriate number of instances is removed each time the floor threshold is breached. Figure 4.4 illustrates dynamic scaling.

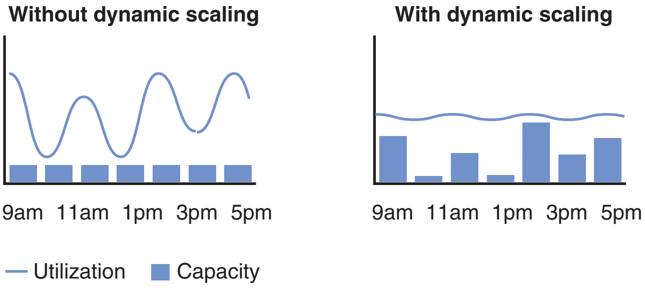


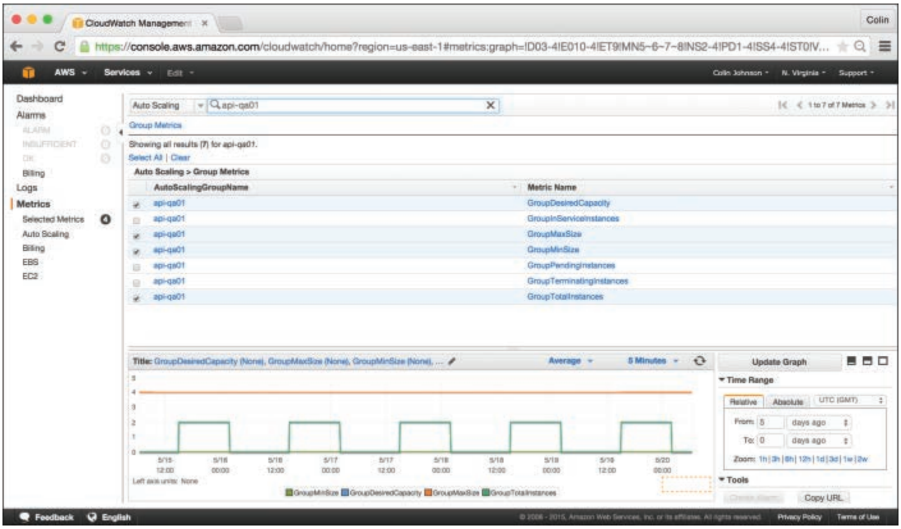
FIGURE 4.4 **Dynamic scaling**

Manual and Scheduled Scaling

The AutoScaling configuration also has a desired instance count. This feature enables you to scale manually and override the configuration as per the scaling policy. You can set the desired count to any size at any time and resize the

AutoScaling group accordingly. This capability is useful if you have knowledge of an upcoming event that will result in an increase of traffic to your site. You can prepare your environment to meet the demand in a much better way because you can increase the AutoScaling group preemptively in anticipation of the traffic.

You can also set up a schedule to scale if you have a very predictable application. Perhaps it is a service being used only from 9 a.m. to 5 p.m. each day. You simply set the scale-out to happen at 8 a.m. in anticipation of the application being used and then set a scale-in scheduled action at 6 p.m. after the application is not being used anymore. This way you can easily reduce the cost of operating intermittently used applications by over 50 percent. Figure 4.5 illustrates scheduled scaling.



Showing all results (7) for api-qa01.

Auto Scaling > Group Metrics

	AutoScalingGroupName	Metric Name
☑	api-qa01	GroupDesiredCapacity
□	api-qa01	GroupInServiceInstances
☑	api-qa01	GroupMaxSize
☑	api-qa01	GroupMinSize
□	api-qa01	GroupPendingInstances
□	api-qa01	GroupTerminatingInstances
☑	api-qa01	GroupTotalInstances

Title: GroupDesiredCapacity (None), GroupMaxSize (None), GroupMinSize (None), ...

Left axis units: None

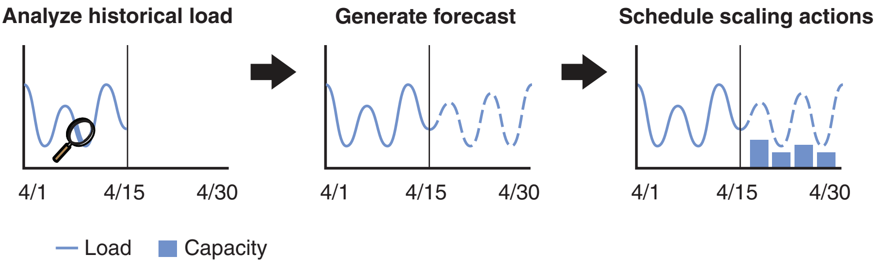
From: 5 days ago

To: 0 days ago

FIGURE 4.5 **Scheduled scaling**

Predictive Scaling

Another AutoScaling feature is predictive scaling, which uses machine learning to learn the scaling pattern of your application based on the minimum amount of historical data. The machine learning component then predicts the scaling after reviewing CW data from the previous 14 days to account for daily and weekly spikes as it learns the patterns on a longer time scale. Figure 4.6 illustrates predictive scaling.



Analyze historical load

Generate forecast

Schedule scaling actions

FIGURE 4.6 **Predictive scaling**

Cram Quiz

Answer these questions. The answers follow the last question. If you cannot answer these questions correctly, consider reading this section again until you can.

1. Which of the following are not characteristics of a scalable/elastic application?

○ **A.** Synchronous request handling in the compute layer

○ **B.** Session persistence in an external database

○ **C.** Session persistence in the compute layer

○ **D.** Asynchronous request offloading to a message queue

2. Which of the following are required to enable the application to scale automatically with AWS AutoScaling? (Choose three.)

○ **A.** EC2 Launch Configuration

○ **B.** Scaling Policy

○ **C.** EC2 User Data

○ **D.** DynamoDB

○ **E.** CloudWatch Alarm

○ **F.** AutoScaling Group

3. True or False: After you assess that your application is fully scalable and elastic, you only need to maintain the application as is in the cloud.

4. True or False: AutoScaling supports only dynamic, scheduled, and predictive scaling.

Cram Quiz Answers

1. Answer: C is correct. The compute layer should be made stateless. Any persistence in the compute layer hinders scalability and elasticity and potentially causes disruption in the application operation. If an instance in a cluster is lost, all the sessions on the instances are lost with it, meaning all the users connected to that particular instance have to log in and start working with the application from scratch.
2. Answer: A, B, and F are correct. To create an autoscaling configuration on EC2, you need an EC2 Launch Configuration that defines how to configure the EC2 instances that are launched; a scaling policy that determines the scaling thresholds; and an autoscaling group that determines the minimum, maximum, and desired numbers of instances.
3. Answer: False. The application should periodically be reassessed for scalability and elasticity because both the application requirements and the SLA might have changed.
4. Answer: False. Autoscaling also supports manual scaling by setting the desired number of instances in the autoscaling group.

Caching

This section covers the following official AWS Certified SysOps Administrator - Associate (SOA-C02) exam domains:

- Domain 2: Reliability and Business Continuity
- Domain 5: Networking and Content Delivery

CramSaver

If you can correctly answer these questions before going through this section, save time by skimming the Exam Alerts in this section and then completing the Cram Quiz at the end of the section.

1. You have implemented autoscaling on both the web and app tier of your three-tier application, but in times of high read requests, the application seems to be performing slowly or even times out. What could you do to make the application more responsive?
2. You have been tasked with deploying a reliable caching solution that can handle multiple different data types and deliver microsecond to millisecond response performance. Which AWS service would you recommend?
3. True or False: To deliver static content to the user in the fastest possible manner, use a web server with lots of memory and utilize server-side caching.

Answers

1. Answer: Implement the read cache to offload the database that seems to be bottlenecking the read requests.
2. Answer: ElastiCache Redis would support all the required features.
3. Answer: False. Static content should be delivered via a content delivery network (CDN). In AWS, you can use CloudFront to deliver static content through more than 200 geographically distributed locations across the globe.

Now that you have seen how to create an application that is highly scalable and elastic, you need to also consider the scaling impact on the persistent layer. As mentioned in the previous section, you need to consider the scalability of the persistent layer and include it in the overall assessment of the elasticity of the application. It serves no purpose to make an application highly elastic when the database back end is rigid and introduces a bottleneck for the whole application.

This is where caching comes in as a good way to ensure the data in the application is accessible in the fastest manner possible. When delivering an application from the cloud, you can use several different caching strategies to deliver and reuse frequently used data and offload the need to request the same data over and over again.

A simple analogy to caching is your refrigerator. It takes a minute for you to walk to the fridge (the cache) and grab some milk (cache hit). However, when there is no milk (cache miss) in the fridge, you need to go to the store (the origin), grab a few cartons, and take them home to put them in the fridge. The journey to the store and back is many times longer, so it makes sense to buy items that you frequently use and put them in the fridge. The fridge does have a limited capacity, just like cache usually does, and you will typically find a much greater variety of items at the store.

Types of Caching

There are several different types of caching that you can use in your application.

Client-Side Caching

When a client requests the contents of the application from a server, you should ensure that components that are static or change infrequently are reused with client-side caching. Modern browsers have this capability built in, and you can use it by specifying cache control headers within your web server or the service that delivers the content, such as S3.

Edge Caching

When content is delivered frequently to multiple users, you can employ edge caching or what is more commonly referred to as a content delivery network. In AWS, you can use the Amazon CloudFront service to deliver frequently used content in a highly efficient manner to millions of users around the globe while at the same time offloading multiple same requests off the application or back end.

Server-Side Caching

When a feature, a module, or certain content stored within the web service is requested frequently, you typically use server-side caching to reduce the need for the server to look for the feature on disk. The first time the feature is requested and the response assembled, the server caches the response in memory so it can be delivered with much lower latency than if it were read from disk and reassembled each time. There is a limitation to the amount of memory the server has, and of course, server-side caching is traditionally limited to each instance. However, in AWS, you can use the ElastiCache service to provide a shared, network-attached, in-memory datastore that can fulfill the needs of caching any kind of content you would usually cache in memory.

Database Caching

The last layer of caching is database caching. This approach lets you cache database contents or database responses into a caching service. There are two approaches to database caching:

- **In-line caching:** This approach utilizes a service that manages the reads and writes to and from the database.
- **Side-loaded caching:** This approach is performed by an application that is aware of the cache and database as two distinct entities. All reads and writes to and from the cache and the database are managed within the application because both the cache and database are two distinct entities.

An example of an in-line caching solution is the DynamoDB Accelerator (DAX) service. With DAX, you can simply address all reads and writes to the DAX cluster, which is connected to the DynamoDB table in the back end. DAX automatically forwards any writes to DynamoDB, and all reads deliver the data straight from the cache in case of a cache hit or forward the read request to the DynamoDB back end transparently. Any responses and items received from DynamoDB are thus cached in the response or item cache. In this case the application is not required to be aware of the cache because all in-line cache operations are identical to the operations performed against the table itself. Figure 4.7 illustrates DAX in-line caching.

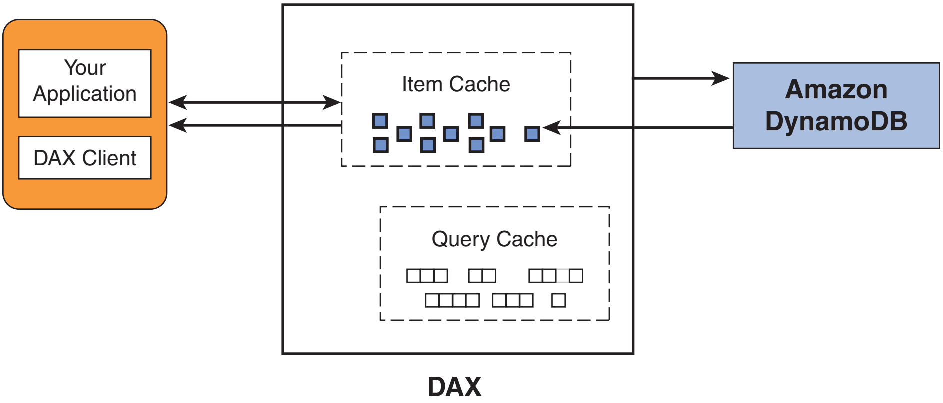


FIGURE 4.7 DAX in-line caching

An example of a sideloaded caching solution is ElastiCache. First, you set up the caching cluster with ElastiCache and a database. The database can be DynamoDB, RDS, or any other database because ElastiCache is not a purpose-built solution like DAX. Second, you have to configure the application to look for any content in the cache first. If the cache contains the content, you get a cache hit, and the content is returned to the application with very low latency. If you get a cache miss because the content is not present in the cache, the application needs to look for the content in the database, read it, and also perform a write to the cache so that any subsequent reads are all cache hits.

There are two traditional approaches to implement sideloaded caching:

- **Lazy loading:** Data is always written only to the database. When data is requested, it is read from the database and cached. Every subsequent read of the data is read from the cache until the item expires. This is a lean approach, ensuring only frequently read data is in the cache. However, every first read inherently suffers from a cache miss. You can also warm up the cache by issuing the reads you expect to be frequent. Figure 4.8 illustrates lazy loading.

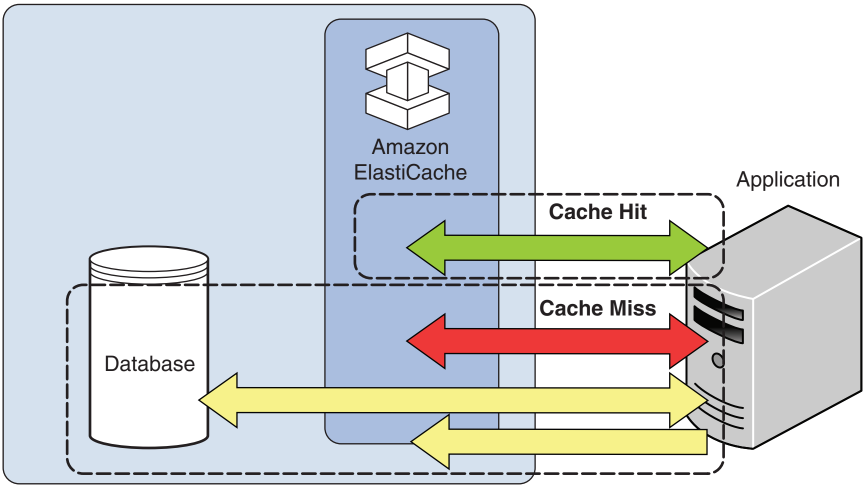


FIGURE 4.8 Lazy loading

- **Write through:** Data is always written to both the database and the cache. This avoids cache misses; however, this approach is highly intensive on the cache. Figure 4.9 illustrates write-through caching.

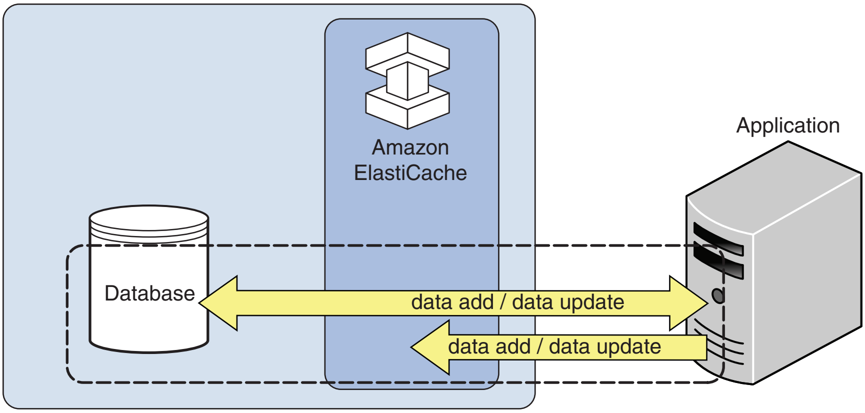


FIGURE 4.9 **Write-through caching**

Because the application controls the caching, you can implement some items to be lazy loaded but others to be written through. The approaches are in no way mutually exclusive, and you can write the application to perform caching based on the types of data being stored in the database and how frequently those items are expected to be requested.

ExamAlert

When choosing a caching strategy, always consider the rate of data change and choose the correct time-to-live (TTL) of the data in the cache, to match the rate of change of the data. Data on an e-commerce site such as item descriptions, reviews, and images are unlikely to change frequently, but data such as item stock and price might not be suitable for caching at all. Always keep this in mind when selecting an answer on the exam.

ElastiCache

ElastiCache is a managed service that can deploy clusters of in-memory data stores. They can be used to perform server-side and database caching.

A typical use case is database offloading with an application-managed side-loaded cache, as described in the previous section. Most applications that work with a database have a high read-to-write ratio—somewhere in the range of 80–90 percent reads to 10–20 percent writes. This means that offloading the reads from the database could save as much as 60–80 percent of the load on the database. For example, if the read-to-write ratio is 90–10 percent, each write (10 percent resources) requires a subsequent read (at least 10 percent resources), and if the rest of the reads are cached, up to 80 percent of the read load can be offloaded to the cache.

An additional benefit of caching with ElastiCache is that the two engines used, Memcached and Redis, are both in-memory datastores. In comparison with databases where response latencies are measured in milliseconds to seconds, the response times from in-memory databases are measured in microseconds to milliseconds. That can mean that any cached data can be delivered 10, 100, or potentially even 1000 times faster than it would be if the request were read from the database. Figure 4.10 contrasts latency of Redis versus S3.

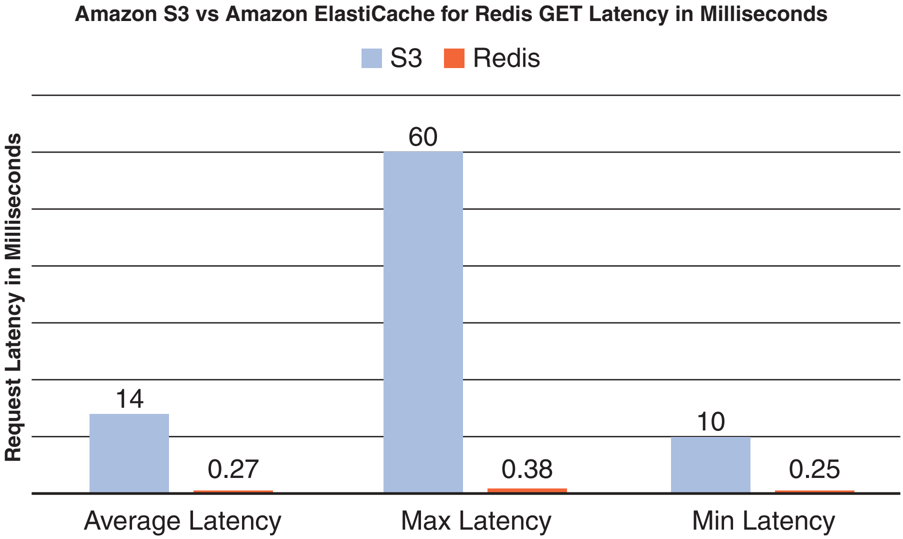


FIGURE 4.10 **Redis vs. S3 GET latencies**

Memcached

One of the engines supported by ElastiCache is Memcached, a high-performance, distributed, in-memory key-value store. The service has a simple operational principle in which one key can have one or many nested values. All the data is served out of memory; thus, there is no indexing or data organization in the platform. When using Memcached, you can design either a single instance cache or a multi-instance cache where data is distributed into partitions. These partitions can be discovered by addressing the service and requesting a partition map. There is no resilience or high availability within the cluster because there is no replication of partitions.

Memcached is a great solution for offloading frequent identical database responses. Another use for Memcached is as a shared session store for multiple web instances. However, the lack of resilience within the Memcached design means that any failure of a node requires the application to rebuild the node from persistent database data or for the sessions with the users to be re-established. The benefit of Memcached is that it is linearly read and write scalable because all you need to do is add nodes to the cluster and remap the partitions.

Redis

The other engine supported by ElastiCache is Redis, a fully-fledged in-memory database. Redis supports much more complex datasets such as tables, lists, hashes, and geospatial data. Redis also has a built-in push messaging feature that can be used for high-performance messaging between services and chat. Redis also has three operational modes that give you advanced resilience, scalability, and elasticity:

- **Single node:** A single Redis node, not replicated, nonscalable, and nonresilient. It can be deployed only in a single availability zone; however, it does support backup and restore.
- **Multinode, cluster mode disabled:** A primary read-write instance with up to five read replicas. The read replicas are near synchronous and offer resilience and read offloading for scalability. Multinode clusters with cluster mode disabled can also be deployed in one or more availability zones.
- **Multinode, cluster mode enabled:** One or more shards of multinode deployments, each with one primary and up to five read replicas. By enabling cluster mode, you retain the resilience and scalability but also add elasticity because you can always reshard the cluster and horizontally add more write capacity. Cluster mode always requires the database nodes to be deployed across multiple availability zones. However, sharding means that multiple primary nodes are responsible for multiple sets of data. Just like with database sharding, this increases write performance in ideal circumstances, but it is not always guaranteed and can add additional complexity to an already-complex solution. Figure 4.11 illustrates Redis multinode cluster mode, both disabled and enabled.

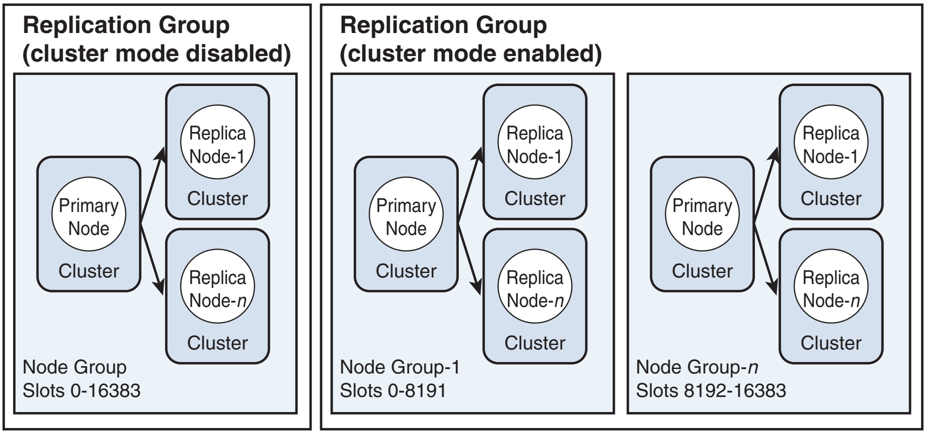


FIGURE 4.11 **Redis multinode, cluster mode disabled vs. enabled**

The features of Redis give you much more than just a caching service. Many businesses out there use ElastiCache Redis as a fully functional in-memory database because the solution is highly resilient, scalable, elastic, and can even be backed up and restored, just like a traditional database.

ExamAlert

If an exam question indicates the database is overloaded with reads, caching should be the main strategy to enable scalability of reads. The benefit of caching is that the data in the cache, when configured correctly, is highly likely to be current and synchronous with the database. However, always be careful when scaling the caching cluster because too many caching instances can add additional unnecessary cost to the application.

Amazon CloudFront

Now that the database and server-side caching are sorted, you need to focus on edge caching. CloudFront is a global content delivery network that can offload static content from the data origin and deliver any cached content from a location that is geographically much closer to the user. This reduces the response latency and makes the application feel as if it were deployed locally, no matter which region or which physical location on the globe the origin resides in.

CloudFront also can terminate all types of HTTP and HTTPS connections at the edge, thus offloading the load to servers or services. CloudFront also works in combination with the Amazon Certificate Manager (ACM) service, which can issue free HTTPS certificates for your domain that are also automatically renewed and replaced each year.

CloudFront enables you to configure it to accept and terminate the following groups of HTTP methods:

- **GET and HEAD:** Enables standard caching for documents and headers. Useful for static websites.
- **GET, HEAD, and OPTIONS:** Enables you to cache OPTIONS responses from an origin server.
- **GET, HEAD, OPTIONS, PUT, PATCH, POST, and DELETE:** Terminates all HTTP(S) sessions at the CloudFront Edge Location and can increase the performance of both the read and write requests.

One of the better features is the ability to rewrite the caching settings for both the cache held within the CloudFront CDN as well as the client-side caching headers defined within servers. This means that you can customize the time-to-live of both the edge and client-side cache. CloudFront distributions can be configured with the following options for setting TTL:

- **Min TTL:** Required setting when all HTTP headers are forwarded from the origin server. It defines the minimum cache TTL and also determines the shortest interval for CloudFront to refresh the data from the origin.
- **Max TTL:** Optional setting. It defines the longest possible cache TTL. It is used to override any cache-control headers defined at the origin server.
- **Default TTL:** Optional setting. It allows you to define cache behavior for any content where no TTL is defined at the origin server.

CloudFront Security

CloudFront is also inherently secure against distributed denial-of-service (DDoS) attacks because the content is distributed to more than 200 locations around the globe. An attacker would need to have a massive, globally distributed botnet to be able to attack your application. On top of the benefit of the distributed architecture, CloudFront is also resilient to L3 and L4 DDoS attacks with the use of AWS Shield Standard service. Any CloudFront distribution can also be upgraded to Shield Advanced, a subscription-based service that provides a detailed overview of the state of any DDoS attacks as well as a dedicated 24/7 response team, which can help with custom DDoS mitigation at any OSI layer.

CloudFront also supports integration with the AWS Web Application Firewall (WAF) service that can filter any attempts at web address manipulations, SQL injections, CSS attacks, and common web server vulnerabilities as well as filter traffic based on IP, geography patterns, regular expressions, and methods.

CloudFront also enables you to limit access to content by

- Restricting access to your application content with signed URLs or cookies
- Restricting access to content based on geolocation
- Restricting access to S3 buckets using Origin Access Identity (OAI)

Cram Quiz

Answer these questions. The answers follow the last question. If you cannot answer these questions correctly, consider reading this section again until you can.

1. Which AWS service enables you to easily deploy a horizontally scalable in-memory caching cluster?

   ❍ **A.** ElastiCache Memcached

   ❍ **B.** ElastiCache Redis, Cluster mode enabled

   ❍ **C.** ElastiCache Redis, Cluster mode disabled

   ❍ **D.** CloudFront

2. You have configured a CloudFront distribution to cache static content from an Apache2 web server. The content on the web server is refreshed every 15 minutes when the application is updated. However, the users are complaining that they seem to see updates only every 2 hours or so. What is most likely the problem, and how would you resolve this issue?

   ❍ **A.** CloudFront TTL is too long. Set the Min TTL to 15 minutes. This will ensure the content is refreshed every 15 minutes.

   ❍ **B.** Origin TTL is too long. Set the Max TTL to 15 minutes. This will ensure the content is refreshed every 15 minutes.

   ❍ **C.** Origin TTL does not exist. Set the Default TTL to 15 minutes. This will ensure the content is refreshed every 15 minutes.

   ❍ **D.** CloudFront TTL does not exist. Set the TTL to enabled and Default TTL to 15 minutes. This will ensure the content is refreshed every 15 minutes.

Cram Quiz Answers

1. Answer: B is correct. ElastiCache Memcached is a high-performance, distributed, in-memory key-value store that can scale horizontally.
2. Answer: B is correct. Max TTL defines the longest possible cache TTL and is used to override any cache-control headers defined at the origin server that are likely to be misconfigured.

Read Replicas

This section covers the following official AWS Certified SysOps Administrator - Associate (SOA-C02) exam domain:

- Domain 2: Reliability and Business Continuity

CramSaver

If you can correctly answer these questions before going through this section, save time by skimming the Exam Alerts in this section and then completing the Cram Quiz at the end of the section.

1. Your three-tier application has been connected to a business intelligence (BI) forecasting platform. While the forecasts are improving business practices, the users of your application are reporting the performance has decreased. The web and app tier are scaling appropriately, and the caching cluster is at about 40 percent capacity. What could be the cause of the slowdown seen by the users, and how could you resolve it?
2. True or False: Aurora natively supports both MySQL and PostgreSQL.

Answers

1. Answer: The BI platform has introduced additional load on the database. Because the BI forecasting requires access to most or all of the dataset, the cache cannot be used to offload the required reads. To mitigate, implement a database read replica.
2. Answer: True. The two engines are fully supported at the time of writing. Other database engines might be supported in the future.

There are five general approaches to scaling database performance:

- **Vertical scaling:** You can add more CPU and RAM to the primary instance.
- **Horizontal scaling:** You can add more instances to a database cluster to increase the available CPU and RAM, but this approach is not always supported.

- **Sharding:** When horizontal scaling is not supported, you can distribute the dataset across multiple primary database engines, thus achieving higher write performance.
- **Caching:** You can add a caching cluster to offload reads, which can be expensive.
- **Read replicas:** You can add read replicas to offload read traffic, possibly asynchronous.

Read replicas can potentially offer the same benefit of read offloading that caching can have for your application. One big benefit of read replicas is that the whole database is replicated to the read replica, not just frequently read items. This means that read replicas can be used where the reads are frequently distributed across the majority of the data in the database. As you can imagine, this would be very expensive and resource intensive to achieve in the cache. You would need to provision enough in-memory capacity (potentially terabytes), read the entire database, and write the data into the cache before you could perform any complex operation on the data in the cache. Sometimes complex operations can't even be performed on the caching server; for example, joins of multiple tables for analytics, business intelligence, or data mining purposes are just not possible within the cache.

This is where read replicas excel. You can introduce read replicas in most instance-based AWS database services, including RDS, Aurora, DocumentDB, and Neptune.

In the RDS service, up to five read replicas can be deployed in any MySQL, MariaDB, PostgreSQL, or Oracle database. Because the data resides on the volume attached to the instance, built-in database replication tools are used to replicate the data across the network to the read replica. This means that read replicas introduce additional load on the primary instance and that the replication is always asynchronous with a potential lag of a few seconds to potentially a few minutes in extreme cases. Figure 4.12 illustrates RDS read replicas.

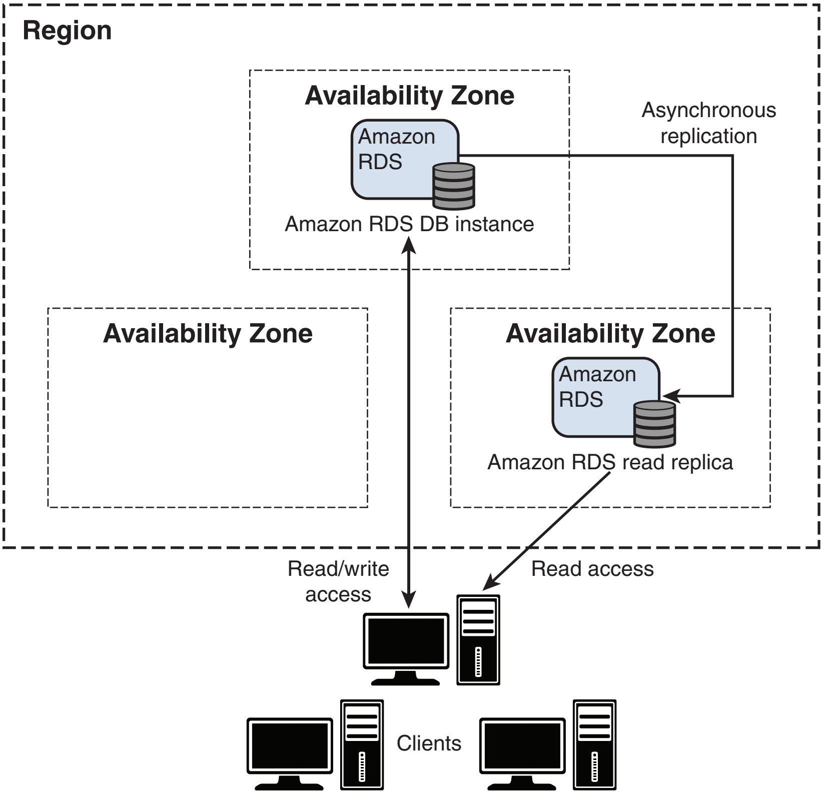


FIGURE 4.12 **RDS read replicas**

The replica can be placed in any availability zone in the same region as the primary database or can be deployed in a cross-region deployment. A read replica can also be promoted to a primary instance, which means that establishing a read replica can be an easy way to clone a database. After the replica is promoted to primary only, a sync of all the missing data is required.

ExamAlert

Traditional RDS database read replicas are a very cost-efficient way to provide data for small (terabyte)-scale analytics and should be selected as a preferred option when the entire dataset needs to be read offloaded. Always evaluate whether the question requires you to offload a certain portion of the data (with caching) or the entire dataset is required (with read replicas).

Other AWS database services employ a decoupled compute–datastore approach. For example, the AWS Aurora service stores all the data on a cluster volume that is replicated to six copies across (at least) three availability zones. The primary instance has read and write access. All writes are sent as changelogs directly to the six nodes of the storage volume, and the commit to the database page is done at the storage cluster. This means that the replication is near synchronous with potentially no more than a few milliseconds of replication lag due to the network distance between the cluster volume nodes in other availability zones and several hundred milliseconds if the replication is done across regions. Additionally, Aurora supports up to 15 read replicas per region and can be deployed to multiple regions with an additional 16 read replicas in other regions. All of the read replicas read from the same cluster volume, meaning they can deliver near synchronous data for each read request. The read replicas can also be seamlessly scaled because there is no requirement to replicate the data on to the replica. When a replica is started, it is simply connected to the cluster volume. This allows the cluster to avoid initial replication lags that you would see in traditional databases. This feature can be used both for elasticity and vertical as well as horizontal scaling of the cluster. Figure 4.13 illustrates an Aurora cluster design.

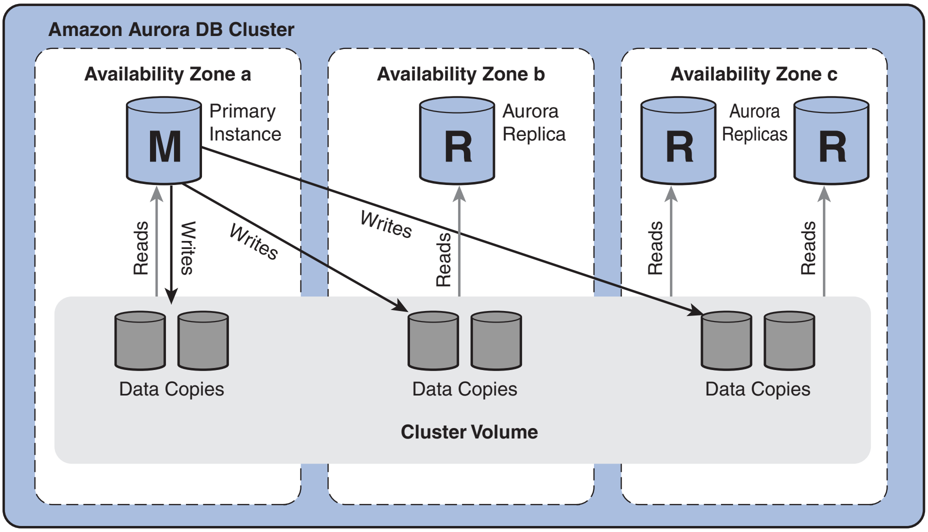


FIGURE 4.13 **Aurora cluster design**

A similar approach to decoupling the storage volume and the database nodes is used in DocumentDB, Neptune, and so on.

On top of this capability, Aurora can forward some read-heavy analytical operations to the cluster volume nodes and offload the primary from having to perform the reads for any JOIN, GROUP BY, UNION, and such operations across approximately a million or more rows. This capability extends the lightweight analytics capabilities of RDS read replicas to deliver much more power to your analytics queries.

Aurora natively supports MySQL and PostgreSQL databases and is slightly more expensive than RDS MySQL and RDS PostgreSQL.

> ExamAlert
>
> When an exam question indicates that high performance at high scale is required for MySQL or PostgreSQL, always evaluate whether the answer includes Aurora as an option. We recommend you select RDS MySQL or RDS PostgreSQL only if any cost considerations are explicitly stated in the question. Also, consider reducing the number of read replicas by powering them off or terminating them if they are not in use.

Cram Quiz

Answer these questions. The answers follow the last question. If you cannot answer these questions correctly, consider reading this section again until you can.

1. You have been instructed to eliminate any inefficiencies in the following deployment:

Web tier: EC2 autoscaling group scaling on CPU usage, 30% floor, 70% ceiling, minimum 1, maximum 10.

App tier: EC2 autoscaling group scaling on CPU usage, 20% floor, 80% ceiling, minimum 1, maximum 6.

Cache tier: 1 ElastiCache Memcached cluster with 2 partitions

Database tier: Multi-AZ MySQL RDS with 3 read replicas

Which of the following would allow you to cost optimize the cluster? (Choose all that apply.)

❍ **A.** Evaluate whether all RDS read replicas are required.

❍ **B.** Evaluate the maximum on the App and Web EC2 autoscaling groups.

❍ **C.** Evaluate the partition configuration of the ElastiCache cluster.

❍ **D.** Evaluate the network performance if the app tier matches the network performance of the database tier.

2. You have been asked to propose a solution to scale the read portion of an application database back end in the most cost-effective manner possible. Your application runs Linux LAMP stack with a 30 GB MySQL RDS Multi-AZ back end. The read-heavy operations are very predictable and occur during a particularly heavy four-hour operation each week and require the whole dataset. Which option would you recommend?

   ❍ **A.** Replace the RDS MySQL with Aurora MySQL and let Aurora add read replicas automatically as needed.

   ❍ **B.** Point the read-heavy operation at the RDS Multi-AZ replica in the other availability zone.

   ❍ **C.** Deploy a read replica in RDS. Point the application at the RDS read replica. Terminate the read replica after the read-heavy operation is complete. Repeat the process with a script next week.

   ❍ **D.** Deploy a read replica in RDS. Point the application at the RDS read replica. Create a script that powers off the read replica after the read-heavy operation is complete and powers it on before the next operation with enough time to synchronize the changes.

Cram Quiz Answers

1. Answer: A and C are correct. You should always evaluate whether the read replicas and caching clusters you deployed are required and properly scaled.
2. Answer: C is correct. Although A and D are possible solutions, deploying a read replica on a weekly basis would be the most cost-effective way to achieve this goal because the operation is sparse in nature and lasts for only four hours. Deploying a read replica is performed from a snapshot and can be just as fast or even faster than replicating a week's worth of data to the powered-off replica. Although the charges of powered-off instances are reduced, they are not zero. Remember also that you can stop a DB instance for up to seven days. If you don't manually start your DB instance after seven days, your DB instance is automatically started so that it doesn't fall behind any required maintenance updates. B is impossible because a Multi-AZ replica is not accessible for any SQL operations.

What Next?

If you want more practice on this chapter's exam objectives before you move on, remember that you can access all of the Cram Quiz questions on the Pearson Test Prep software online. You can also create a custom exam by objective with the Online Practice Test. Note any objective you struggle with and go to that objective's material in this chapter.

This page intentionally left blank

Index