**1**

**Understanding AWS Principles and Key Characteristics**

The last decade has revolutionized the IT infrastructure industry; cloud computing was introduced and now it is everywhere, from small start-ups to large enterprises. Nowadays, the cloud is the new normal. It all started with Amazon launching a cloud service called **Amazon Web Services** (**AWS**) in 2006 with a couple of services.

Netflix migrated to AWS in 2008 and became a market disrupter. After that, there was no looking back and there were many industry revolutions led by cloud-born start-ups like Airbnb in hospitality, Robinhood in finance, Lyft in transportation, and many more. The cloud rapidly gained the market share, and now big names like Capital One, JP Morgan Chase, Nasdaq, the NFL, and General Electric are all accelerating their digital journey with cloud adoption.

Even though the term ‘cloud’ is pervasive today, not everyone understands what the cloud is as it can be different things for different people, and it is continuously evolving. In this chapter, you will learn what the cloud is, and then what AWS is more specifically. You will learn about the vast and ever-growing influence and adoption of the cloud in general and of AWS in particular. After that, you will start getting introduced to some elementary cloud and AWS terms to get your feet wet with the lingo while gaining an understanding of why cloud computing is so popular. In this chapter, we will cover the following topics:

* What is cloud computing?
* What is **Amazon Web Services** (**AWS**)?
* The market share, influence, and adoption of AWS
* Basic cloud and AWS terminology
* Why is AWS so popular?

Let’s get started, shall we?

**What is cloud computing?**

What exactly is cloud computing? It is a term often thrown around by many people who don’t understand it and wonder what it means. Having your infrastructure in the cloud does not mean you have your servers up in the sky. Let’s try to define it plainly.

Essentially, **cloud computing** is outsourcing a company’s hardware and software infrastructure to a third party. At a high level, it is the on-demand availability of IT resources such as servers, storage, databases, and so on over the web, without the hassle of managing physical infrastructure. Instead of having their own data center, enterprises borrow someone else’s data center. Cloud computing has many advantages:

* Economies of scale are associated with buying in bulk.
* You only pay for the time you use the equipment in increments of minutes or seconds.
* Arguably one of the most important benefits is the ability to scale up, out, down, and in.

When using cloud computing, you are not buying the equipment; you are leasing it. Equipment leasing has been around for a long time, but not at the speeds that cloud computing provides. Cloud computing makes it possible to start a resource within minutes, use it for a few hours, minutes, or even seconds, and then shut it down. You will only pay for the time you use it. Furthermore, with the advent of *serverless* computing, such as AWS Lambda services, we don’t even need to provision servers, and we can call a Lambda function and pay by the function call. The idea of being able to *scale out* and *scale in* is often referred to as **elasticity** or **elastic computing**. This concept allows companies to treat their computing resources as just another utility bill and only pay for what they need at any given moment in time.

The best way to understand the cloud is to take the electricity supply analogy. To get light in your house, you just flip a switch on, and electric bulbs light up your home. In this case, you only pay for your electricity use when you need it; when you switch off electric appliances, you do not pay anything. Now, imagine if you needed to power a couple of appliances, and for that, you had to set up an entire powerhouse. It would be costly, right? It would involve the costs of maintaining the turbine and generator and building the whole infrastructure. Utility companies make your job easier by supplying electricity in the quantity you need.

They maintain the entire infrastructure to generate electricity and they can keep costs down by distributing electricity to millions of houses, which helps them benefit from mass utilization. Here, the utility companies represent cloud providers such as AWS, and the electricity represents the IT infrastructure available in the cloud.

While consuming cloud resources, you pay for IT infrastructure such as computing, storage, databases, networking, software, machine learning, and analytics in a pay-as-you-go model. Here, public clouds like AWS do the heavy lifting to maintain IT infrastructure and provide you with on-demand access over the internet. As you generally only pay for the time and services you use, most cloud providers can provide massive scalability, making it easy to scale services up and down. Where, traditionally, you would have to maintain your servers all by yourself on-premise to run your organization, now you can offload that to the public cloud and focus on your core business. For example, Capital One’s core business is banking and it does not run a large data center.

As much as we tried to nail it down, this is still a pretty broad definition. For example, we specified that the cloud can offer software, that’s a pretty general term. Does the term software in our definition include the following?

* Video conferencing
* Virtual desktops
* Email services
* Contact center
* Document management

These are just a few examples of what may or may not be included as available services in a cloud environment. When AWS started, it only offered a few core services, such as compute (Amazon EC2) and basic storage (Amazon S3). AWS has continually expanded its services to support virtually any cloud workload. As of 2022, it has more than 200 fully featured services for computing, storage, databases, networking, analytics, machine learning, artificial intelligence, Internet of Things, mobile, security, hybrid, virtual and augmented reality, media, application development, and deployment. As a fun fact, as of 2023, Amazon **Elastic Compute Cloud** (**EC2**) alone offers over 500 types of compute instances.

For the individual examples given here, AWS offers the following:

* Video conferencing – Amazon Chime
* Virtual desktops – AWS WorkSpaces
* Email services – Amazon WorkMail
* Contact Center – Amazon Connect
* Document Management – Amazon WorkDocs

Not all cloud services are highly intertwined with their cloud ecosystems. Take these scenarios, for example:

* Your firm may be using AWS services for many purposes, but they may be using WebEx, Microsoft Teams, Zoom, or Slack for their video conference needs instead of Amazon Chime. These services have little dependency on other underlying core infrastructure cloud services.
* You may be using Amazon SageMaker for artificial intelligence and machine learning projects, but you may be using the TensorFlow package in SageMaker as your development kernel, even though Google maintains TensorFlow.

If you are using Amazon RDS and choose MySQL as your database engine, you should not have too much trouble porting your data and schemas over to another cloud provider that supports MySQL if you decide to switch over. However, it will be a lot more difficult to switch to some other services. Here are some examples:

* Amazon DynamoDB is a NoSQL proprietary database only offered by AWS. If you want to switch to another NoSQL database, porting it may not be a simple exercise.
* Suppose you are using CloudFormation to define and create your infrastructure. In that case, it will be difficult, if not impossible, to use your CloudFormation templates to create infrastructure in other cloud provider environments. Suppose the portability of your infrastructure scripts is important to you, and you are planning on switching cloud providers. In that case, using Ansible, Chef, or Puppet may be a better alternative.
* Suppose you have a streaming data requirement and use Amazon Kinesis Data Streams. You may have difficulty porting out of Amazon Kinesis since the configuration and storing mechanism are quite dissimilar if you decide to use another streaming data service like Kafka.

As far as we have come in the last 15 years with cloud technologies, I think vendors realize that these are the beginning innings, and locking customers in right now while they are still deciding who their vendor should be will be a lot easier than trying to do so after they pick a competitor.

However, looking at a cloud-agnostic strategy has its pros and cons. You want to distribute your workload between cloud providers to have competitive pricing and keep your options open like in the old days. But each cloud has different networking needs, and connecting distributed workloads between clouds to communicate with each other is a complex task. Also, each major cloud provider, like AWS, Azure, and GCP, has a breadth of services, and building a workforce with all three skill sets is another challenge.

Finally, clouds like AWS provide economy of scale, which means the more you use, the more the price goes down, which may not benefit you if you choose multi-cloud. Again, it doesn’t mean you cannot choose a multi-cloud strategy, but you have to think about logical workload isolation. It would not be wise to run the application layer in one cloud and the database layer in other, but you can think about logical isolation like running the analytics workload and application workload in a separate cloud.

In this section, you learned about cloud computing at a very high level. Let’s learn about the difference between the public and private clouds.

**Private versus public clouds**

A private cloud is a service dedicated to a single customer—it is like your on-premise data center, which is accessible to one large enterprise. A private cloud is a fancy name for a data center managed by a trusted third party. This concept gained momentum to ensure security as, initially, enterprises were skeptical about public cloud security, which is multi-tenant. However, having your own infrastructure in this manner diminishes the value of the cloud as you have to pay for resources even if you are not running them.

Let’s use an analogy to understand the difference between private and public clouds further. The gig economy has great momentum. Everywhere you look, people are finding employment as contract workers. One of the reasons contract work is getting more popular is because it enables consumers to contract services that they may otherwise not be able to afford. Could you imagine how expensive it would be to have a private chauffeur? But with Uber or Lyft, you almost have a private chauffeur who can be at your beck and call within a few minutes of you summoning them.

A similar economy of scale happens with a public cloud. You can have access to infrastructure and services that would cost millions of dollars if you bought them on your own. Instead, you can access the same resources for a small fraction of the cost.

In general, private clouds are expensive to run and maintain in comparison to public clouds. For that reason, many of the resources and services offered by the major cloud providers are hosted in a shared tenancy model. In addition to that, you can run your workloads and applications on a public cloud securely: you can use security best practices and sleep well at night knowing that you use AWS’s state-of-the-art technologies to secure your sensitive data.

Additionally, most major cloud providers’ clients use public cloud configurations. That said, there are a few exceptions even in this case. For example, the United States government intelligence agencies are a big AWS customer. As you can imagine, they have deep pockets and are not afraid to spend. In many cases with these government agencies, AWS will set up the AWS infrastructure and dedicate it to the government workload. For example, AWS launched a Top Secret Region–AWS Top Secret-West–which is accredited to operate workloads at the Top-Secret U.S. security classification level. The other AWS GovCloud regions are:

* GovCloud (US-West) Region - Launched in 2011

Availability Zones: 3

* GovCloud (US-East) Region - Launched in 2018

Availability Zones: 3

AWS GovCloud (US) is a set of AWS Regions that have been purposely isolated to enable U.S. government entities and clients to transfer sensitive workloads to AWS. This platform caters to particular regulatory and compliance standards such as Department of Defense Security Requirements Guide (DoD SRG) Impact Levels 4 and 5, Federal Risk and Authorization Management Program (FedRAMP) High, and Criminal Justice Information Services (CJIS), among others.

Public cloud providers such as AWS provide you choices to adhere to compliance needs as required by government or industry regulations. For example, AWS offers Amazon EC2 dedicated instances, which are EC2 instances that ensure that you will be the only user for a given physical server. Further, AWS offers AWS Outpost, where you can order server racks and host workloads on-premise using the AWS control plane.

Dedicated instance and outpost costs are significantly higher than on-demand EC2 instances. On-demand instances are multi-tenant, which means the physical server is not dedicated to you and may be shared with other AWS users. However, just because the physical servers are multi-tenant doesn’t mean that anyone else can access your server as those will be dedicated virtual EC2 instances accessible to you only.

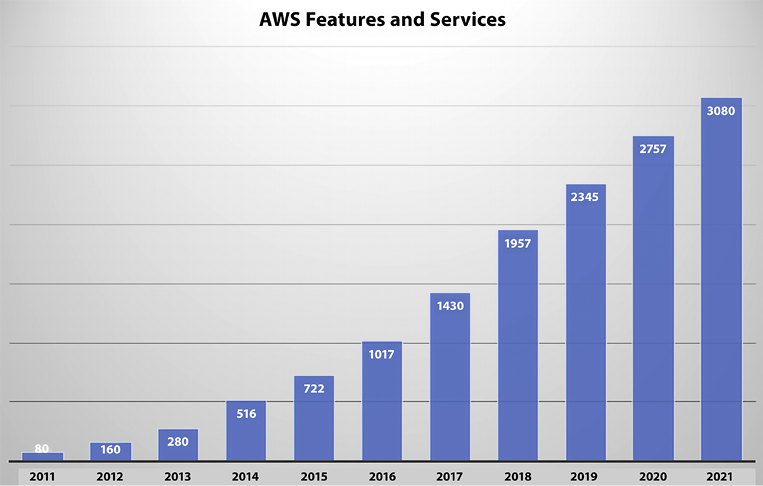
As we will discuss later in this chapter, you will never know the difference when using EC2 instances if they are hosted on a dedicated physical server compared to a multi-tenant server because of virtualization and hypervisor technology. One common use case for choosing dedicated instances is government regulations and compliance policies that require certain sensitive data to not be in the same physical server with other cloud users.

Now that we have gained a better understanding of cloud computing in general, let’s get more granular and learn about how AWS does cloud computing.

**What is AWS (Amazon Web Services)?**

With over 200 fully-featured services available across the world, **Amazon Web Services** (**AWS**) is the most widely used cloud platform globally. Even though there are a few worthy competitors, it doesn’t seem like anyone will push them off the podium for a while.

For example, keeping up with AWS’ pace of innovation can be challenging. The growth of AWS services and features has been tremendous each year, as demonstrated in the graph below. In 2011, AWS introduced over 80 significant services and features, followed by nearly 160 in 2012, 280 in 2013, 516 in 2014, 722 in 2015, 1,017 in 2016, 1,430 in 2017, 1,957 in 2018, 2,345 in 2019, 2,757 in 2020, and 3,084 in 2021.



*Figure 1.1: AWS – number of features released per year*

There is no doubt that the number of offerings will continue to grow at a similar rate for the foreseeable future. Gartner named AWS as a leader for the 12th year in a row in the 2022 Gartner Magic Quadrant for Cloud Infrastructure & Platform Services. AWS is innovating fast, especially in new areas such as machine learning and artificial intelligence, the **Internet of Things** (**IoT**), serverless computing, blockchain, and even quantum computing.

The following are some of the key differentiators for AWS in a nutshell:

|  |  |
| --- | --- |
| Oldest and most experienced cloud provider | AWS was the first major public cloud provider (started in 2006) and since then it has gained millions of customers across the globe. |
| Fast pace of innovation | AWS has 200+ fully featured services to support any cloud workload. They released 3000+ features in 2021 to meet customer demand. |
| Continuous price reduction | AWS has reduced its prices across various services 111 times since its inception in 2006 to improve the **Total Cost of Ownership** (**TCO**). |
| Community of partners to help accelerate the cloud journey | AWS has a large Partner Network of 100,000+ partners across 150+ countries. These partners include large consulting partners and software vendors. |
| Security and compliance | AWS provides security standards and compliance certifications to fulfill your local government and industry compliance needs. |
| Global infrastructure | As of January 2023, AWS has 99 Availability Zones within 31 geographic Regions, 32 Local Zones, 29 Wavelength Zones, 410+ Points of Presence (400+ Edge locations and 13 regional mid-tier caches) in 90+ cities across 47 countries. For the latest information, refer to <https://aws.amazon.com/about-aws/global-infrastructure/>. |

It’s not always possible to move all workloads into the cloud, and for that purpose, AWS provides a broad set of hybrid capabilities in the areas of networking, data, access, management, and application services. For example, VMware Cloud on AWS allows customers to seamlessly run existing VMware workloads on AWS with the skills and toolsets they already have without additional hardware investment. If you want to run your workload on-premise, then AWS Outposts enables you to utilize native AWS services, infrastructure, and operating models in almost any data center, co-location space, or on-premises facility if you prefer to run your workload on-premise. You will learn more details about hybrid cloud services later in this book.

This is just a small sample of the many AWS services that you will see throughout this book. Let’s delve a little deeper into how influential AWS currently is and how influential it has the potential to become.

**The market share, influence, and adoption of AWS**

For the first nine years of AWS’s existence, Amazon did not break down its AWS sales, and since 2015 Amazon started reporting AWS sales separately. As of April 2022, Microsoft does not fully break down its Azure revenue and profit in its quarterly reports. They disclosed their Azure revenue growth rate without reporting the actual revenue number, instead burying Azure revenues in a bucket called Commercial Cloud, which also includes items such as Office 365 revenue. Google has been cagey about breaking down its **Google Cloud Platform** (**GCP**) revenue for a long time. Google finally broke down its GCP revenue in February 2019, but GCP also combines its cloud and workplace (G-suite) tools in the same bucket.

AWS has a large market share with a $80.1 B run rate in 2022 and a 29% year-over-year growth. It is predicted to hit $100 B business by end of 2023, which is phenomenal for a business of its size. As of 2022, AWS is leading cloud IaaS with 34% of the market share as per TechRadar’s global cloud market report. AWS has done a great job of protecting its market share by adding more and more services, adding features to existing services, building higher-level functionality on top of the core services it already offers, and educating the masses on how to best use these services.

We are in an exciting period when it comes to cloud adoption. Until a few years ago, many C-suite executives were leery of adopting cloud technologies to run their mission-critical and core services. A common concern was that they felt having on-premises implementations was more secure than running their workloads on the cloud.

It has become apparent to most of them that running workloads on the cloud can be just as secure as running them on-premises. There is no perfectly secure environment, and it seems that almost every other day, we hear about sensitive information being left exposed on the internet by yet another company. But having an army of security experts on your side, as is the case with the major cloud providers, will often beat any security team that most companies can procure on their own.

The current state of the cloud market for most enterprises is a state of **Fear Of Missing Out** (**FOMO**). Chief executives are watching their competitors jump in on cloud technology, and they are concerned that they will be left behind if they don’t leap.

Additionally, we see an unprecedented level of disruption in many industries propelled by the power of the cloud. Let’s take the example of Lyft and Uber. Both companies rely heavily on cloud services to power their infrastructure, and old-guard companies in the space, such as Hertz and Avis, that depend on older on-premises technology are getting left behind. Part of the problem is the convenience that Uber and Lyft offer by being able to summon a car on demand. But the inability to upgrade their systems to leverage cloud technologies undoubtedly played a role in their diminishing share of the car rental market.

Let’s continue learning some of the basic cloud terminologies and AWS terminology.

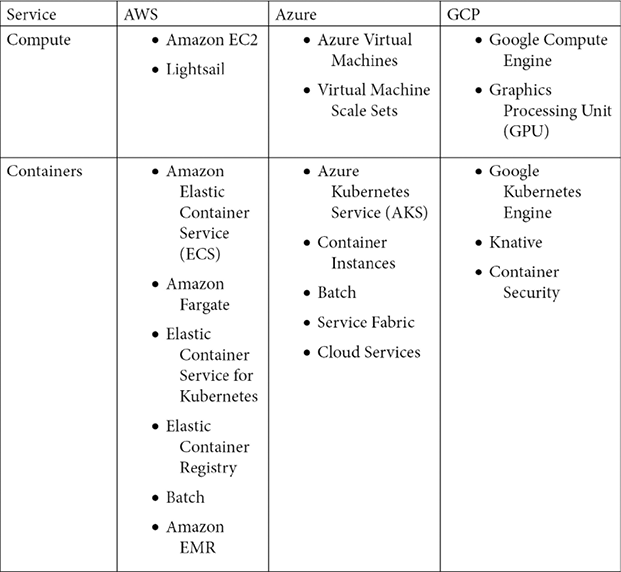
**Basic cloud and AWS terminology**

There is a constant effort by technology companies to offer common standards for certain technologies while providing exclusive and proprietary technology that no one else offers. An example of this can be seen in the database market. The **Standard Query Language** (**SQL**) and the ANSI-SQL standard have been around for a long time. The **American National Standards Institute** (**ANSI**) adopted SQL as the SQL-86 standard in 1986. Since then, database vendors have continuously supported this standard while offering various extensions to make their products stand out and lock in customers to their technology.

Cloud providers provide the same core functionality for a wide variety of customer needs, but they all feel compelled to name these services differently, no doubt in part to try to separate themselves from the rest of the pack. As an example, every major cloud provider offers compute services. In other words, it is simple to spin up a server with any provider, but they all refer to this compute service differently:

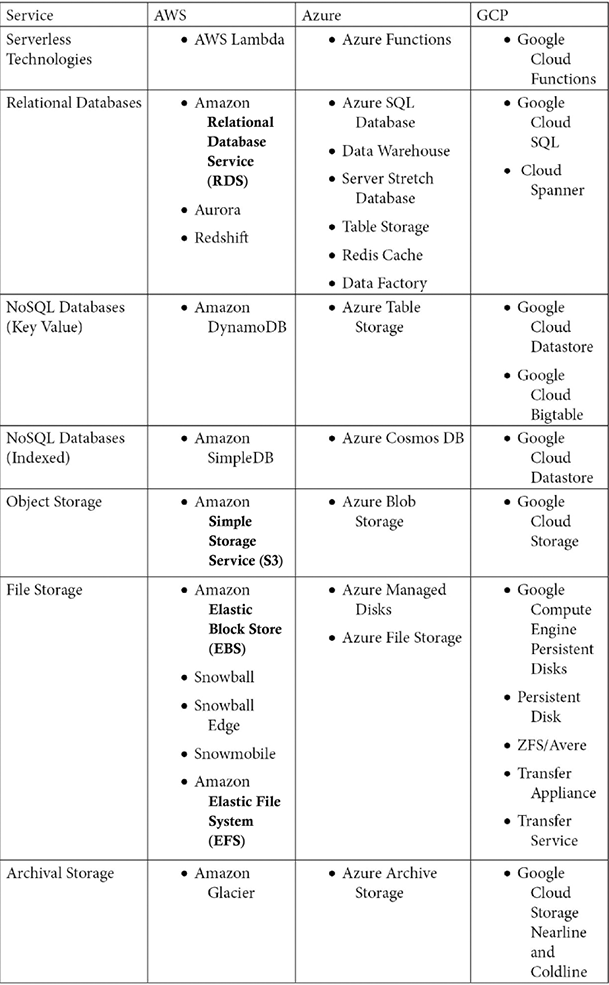
* AWS uses **Elastic Compute Cloud** (**EC2**) instances.
* Azure uses **Azure Virtual Machines**.
* GCP uses **Google Compute Engine**.

The following tables give a non-comprehensive list of the different core services offered by AWS, Azure, and GCP and the names used by each of them. However, if you are confused by all the terms in the tables, don’t fret. We will learn about many of these services throughout the book and when to use them.



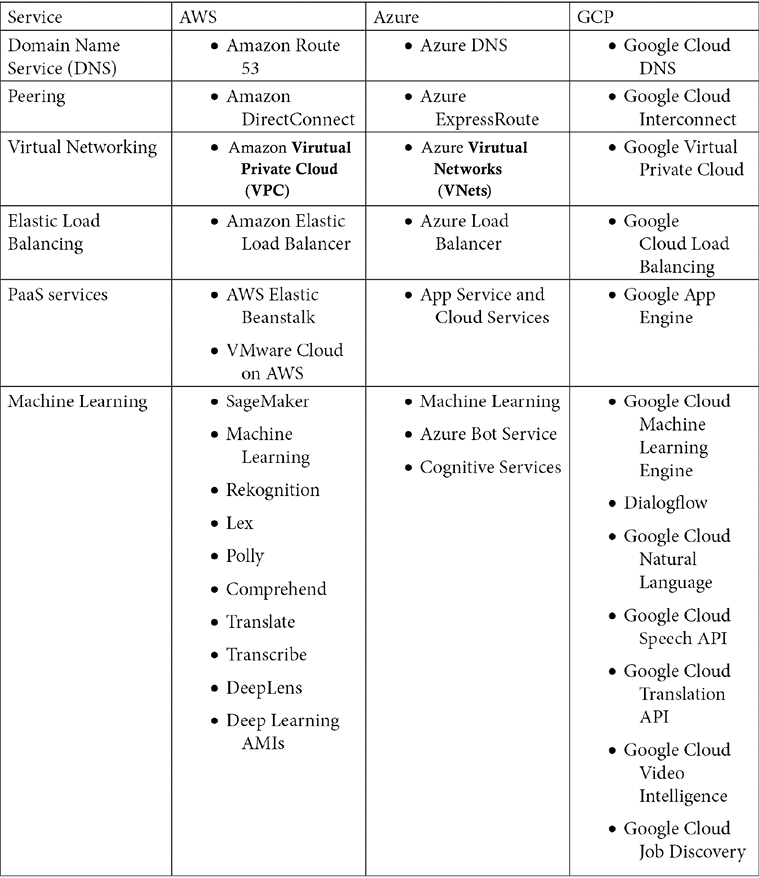
*Figure 1.2: Cloud provider terminology and comparison (part 1)*

These are some of the other services, including serverless technology services and database services:



*Figure 1.3: Cloud provider terminology and comparison (part 2)*

These are additional services:



*Figure 1.4: Cloud provider terminology and comparison (part 3)*

The next section will explain why cloud services are becoming popular and why AWS adoption is prevalent.

**Why is AWS so popular?**

Depending on who you ask, some estimates peg the global cloud computing market at around 545.8 billion USD in 2022, growing to about 1.24 trillion USD by 2027. This implies a **Compound Annual Growth Rate** (**CAGR**) of around 17.9% for the period.

There are multiple reasons why the cloud market is growing so fast. Some of them are listed here:

* Elasticity and scalability
* Security
* Availability
* Faster hardware cycles
* System administration staff

In addition to the above, AWS provides access to emerging technologies and faster time to market. Let’s look at the most important reason behind the popularity of cloud computing (and, in particular, AWS) first.

**Elasticity and scalability**

The concepts of *elasticity* and *scalability* are closely tied. Let’s start by understanding scalability. In the context of computer science, *scalability* can be used in two ways:

* An application can continue to function correctly when the volume of users and/or transactions it handles increases. The increased volume is typically handled by using bigger and more powerful resources (scaling up) or adding more similar resources (scaling out).
* A system can function well when it rescales and can take full advantage of the new scale. For example, a program is scalable if it can be reinstalled on an operating system with a bigger footprint. It can take full advantage of the more robust operating system, achieving greater performance, processing transactions faster, and handling more users.

Scalability can be tracked over multiple dimensions, for example:

* **Administrative scalability** – Increasing the number of users of the system
* **Functional scalability** – Adding new functionality without altering or disrupting existing functionality
* **Heterogeneous scalability** – Adding disparate components and services from a variety of vendors
* **Load scalability** – Expanding capacity to accommodate more traffic and/or transactions
* **Generation scalability** – Scaling by installing new versions of software and hardware
* **Geographic scalability** – Maintaining existing functionality and SLAs while expanding the user base to a larger geographic region.

Scalability challenges are encountered by IT organizations all over daily. It is difficult to predict demand and traffic for many applications, especially internet-facing applications. Therefore, it is difficult to predict how much storage capacity, compute power, and bandwidth will be needed.

Say you finally launch a site you’ve been working on for months and within a few days you begin to realize that too many people are signing up and using your service. While this is an excellent problem, you better act fast, or the site will start throttling, and the user experience will go down or be non-existent. But the question now is, how do you scale? When you reach the limits of your deployment, how do you increase capacity? If the environment is on-premises, the answer is *very painful*. You will need approval from the company leadership. New hardware will need to be ordered. Delays will be inevitable. In the meantime, the opportunity in the marketplace will likely disappear because your potential customers will bail to competitors that can meet their needs. Being able to deliver quickly may not just mean getting there first. It may be the difference between getting there first and not getting there in time.

If your environment is on the cloud, things become much simpler. You can simply spin up an instance that can handle the new workload (correcting the size of a server can even be as simple as shutting down the server for a few minutes, changing a drop-down box value, and restarting the server again). You can *scale* your resources to meet increasing user demand.

The scalability that the cloud provides exponentially improves the time to market by accelerating the time it takes for resources to be provisioned.

**NOTE**

There are two different methods for scaling resources: scaling up (vertical scaling) and scaling out (horizontal scaling).

Scaling up is achieved by getting a bigger boat. For example, AWS offers a range of different-sized instances, including; nano; micro; small; medium; large; xlarge; 2x, 4x, 8x, 16x, and 32x large. So, if you are running a job on a medium instance and the job starts hitting the performance ceiling for that size, you could swap your work to a large or xlarge instance. This could happen because a database needs additional capacity to perform at a prescribed level. The new instance would have a better CPU, more memory, more storage, and faster network throughput. Scaling up can also be achieved using software – for example, allocating more memory or overclocking the CPU.

While scaling up is achieved by using more powerful nodes, scaling out is achieved by adding more nodes. Scaling out can be achieved in the following ways:

* Adding infrastructure capacity by adding new instances or nodes on an application-by-application basis
* Adding additional instances independently of the applications
* Adding more processes, connections, or shards with software

Scaling out is particularly valuable for multi-tiered architectures where each tier has a well-defined responsibility as it allows you to modify just one resource where a bottleneck exists and leave the other resources alone. For example, if you are running a multi-tiered architecture and discover that an application server is running at 95% CPU, you can add additional application servers to help balance the load without having to modify your web server or database server.

These scaling options can also be used simultaneously to improve an application. For example, in addition to adding more instances to handle traffic, more significant and capable instances can be added to the cluster.

As well as making it easy to scale resources, AWS and other cloud operators allow you to quickly adapt to shifting workloads due to their elasticity. Elasticity is defined as the ability of a computing environment to adapt to changes in workload by *automatically* provisioning or shutting down computing resources to match the capacity needed by the current workload.

These resources could be a single instance of a database or a thousand copies of the application and web servers used to handle your web traffic. These servers can be provisioned within minutes. In AWS and the other main cloud providers, resources can be shut down without having to terminate them completely, and the billing for resources will stop if the resources are shut down.

The ability to quickly shut down resources and, significantly, not be charged for that resource while it is down is a very powerful characteristic of cloud environments. If your system is on-premises, once a server is purchased, it is a sunk cost for the duration of the server’s useful life. In contrast, whenever we shut down a server in a cloud environment. The cloud provider can quickly detect that and put that server back into the pool of available servers for other cloud customers to use that newly unused capacity.

This distinction cannot be emphasized enough. The only time absolute on-premises costs may be lower than cloud costs is when workloads are extremely predictable and consistent. Computing costs in a cloud environment on a per-unit basis may be higher than on-premises prices, but the ability to shut resources down and stop getting charged for them makes cloud architectures cheaper in the long run, often in a quite significant way.

The following examples highlight how useful elasticity can be in different scenarios:

* **Web storefront** – A famous use case for cloud services is to use them to run an online storefront. Website traffic in this scenario will be highly variable depending on the day of the week, whether it’s a holiday, the time of day, and other factors—almost every retail store in the USA experiences more than a 10x user workload during Thanksgiving week. The same goes for Boxing Day in the UK, Diwali in India, Singles’ Day in China, and almost every country has a shopping festival. This kind of scenario is ideally suited for a cloud deployment. In this case, we can set up resource auto-scaling that automatically scales up and down compute resources as needed. Additionally, we can set up policies that allow database storage to grow as needed.
* **Big data workloads** – As data volumes are increasing exponentially, the popularity of Apache Spark and Hadoop continues to increase to analyze GBs and TBs of data. Many Spark clusters don’t necessarily need to run consistently. They perform heavy batch computing for a period and then can be idle until the next batch of input data comes in. A specific example would be a cluster that runs every night for 3 or 4 hours and only during the working week. In this instance, you need decoupled compute and data storage where you can shut down resources that may be best managed on a schedule rather than by using demand thresholds.

Or, we could set up triggers that automatically shut down resources once the batch jobs are completed. AWS provides that flexibility where you can store your data in Amazon Simple Storage Service (S3) and spin up an Amazon Elastic MapReduce (EMR) cluster to run Spark jobs and shut them down after storing results back in decoupled Amazon S3.

* **Employee workspace** – In an on-premise setting, you provide a high configuration desktop/laptop to your development team and pay for it for 24 hours a day, including weekends. However, they are using one-fourth of the capacity considering an eight-hour workday. AWS provides workspaces accessible by low configuration laptops, and you can schedule them to stop during off-hours and weekends, saving almost 70% of the cost.

Another common use case in technology is file and object storage. Some storage services may grow organically and consistently. The traffic patterns can also be consistent. This may be one example where using an on-premises architecture may make sense economically. In this case, the usage pattern is consistent and predictable.

Elasticity is by no means the only reason that the cloud is growing in leaps and bounds. The ability to easily enable world-class security for even the simplest applications is another reason why the cloud is becoming pervasive.

**Security**

The perception of *on-premises* environments being more secure than cloud environments was a common reason companies big and small would not migrate to the cloud. More and more enterprises now realize that it is tough and expensive to replicate the security features provided by cloud providers such as AWS. Let’s look at a few of the measures that AWS takes to ensure the security of its systems.

**Physical security**

AWS data centers are highly secured and continuously upgraded with the latest surveillance technology. Amazon has had decades to perfect its data centers’ design, construction, and operation.

AWS has been providing cloud services for over 15 years, and they have an army of technologists, solution architects, and some of the brightest minds in the business. They are leveraging this experience and expertise to create *state-of-the-art* data centers. These centers are in nondescript facilities. You could drive by one and never know what it is. It will be extremely difficult to get in if you find out where one is. Perimeter access is heavily guarded. Visitor access is strictly limited, and they always must be accompanied by an Amazon employee.

Every corner of the facility is monitored by video surveillance, motion detectors, intrusion detection systems, and other electronic equipment. Amazon employees with access to the building must authenticate themselves four times to step on the data center floor.

Only Amazon employees and contractors that have a legitimate right to be in a data center can enter. Any other employee is restricted. Whenever an employee does not have a business need to enter a data center, their access is immediately revoked, even if they are only moved to another Amazon department and stay with the company. Lastly, audits are routinely performed and are part of the normal business process.

**Encryption**

AWS makes it extremely simple to encrypt data at rest and data in transit. It also offers a variety of options for encryption. For example, for encryption at rest, data can be encrypted on the server side, or it can be encrypted on the client side. Additionally, the encryption keys can be managed by AWS, or you can use keys that are managed by you using tamper-proof appliances like a **Hardware Security Module** (**HSM**). AWS provides you with a dedicated cloud HSM to secure your encryption key if you want one. You will learn more about AWS security in *Chapter 8*, *Best Practices for Application Security, Identity, and Compliance*.

**AWS supports compliance standards**

AWS has robust controls to allow users to maintain security and data protection. We’ll discuss how AWS shares security responsibilities with its customers, but the same is true of how AWS supports compliance. AWS provides many attributes and features that enable compliance with many standards established in different countries and organizations. By providing these features, AWS simplifies compliance audits. AWS enables the implementation of security best practices and many security standards, such as these:

* STAR
* SOC 1/SSAE 16/ISAE 3402 (formerly SAS 70)
* SOC 2
* SOC 3
* FISMA, DIACAP, and FedRAMP
* PCI DSS Level 1
* DOD CSM Levels 1-5
* ISO 9001 / ISO 27001 / ISO 27017 / ISO 27018
* MTCS Level 3
* FIPS 140-2
* I TRUST

In addition, AWS enables the implementation of solutions that can meet many industry-specific standards, such as these:

* **Criminal Justice Information Services** (**CJIS**)
* **Family Educational Rights and Privacy Act** (**FERPA**)
* **Cloud Security Alliance** (**CSA**)
* **Motion Picture Association of America** (**MPAA**)
* **Health Insurance Portability and Accountability Act** (**HIPAA**)

The above is not a full list of compliance standards; there are many more compliance standards met by AWS according to industries and local authorities across the world.

Another important thing that can explain the meteoric rise of the cloud is how you can stand up high-availability applications without paying for the additional infrastructure needed to provide these applications. Architectures can be crafted to start additional resources when other resources fail. This ensures that we only bring additional resources when necessary, keeping costs down. Let’s analyze this important property of the cloud in a deeper fashion.

**Availability**

Intuitively and generically, the word *“availability”* conveys that something is available or can be used. In order to be used, it needs to be up and running and in a functional condition. For example, if your car is in the driveway, it is working, and is ready to be used then it meets some of the conditions of availability. However, to meet the technical definition of *“availability,”* it must be turned on. A server that is otherwise working correctly but is shut down will not help run your website.

**NOTE**

Often high availability is confused with fault tolerance. A system can be 100% available but 50% fault tolerant. For example, suppose you need four servers to handle your application load and provide the required performance. You have built redundancy by putting two servers in two different data centers. In that case, your system is 100% available and 100% fault tolerant. But for some reason, one of the data centers has gone down. Your system is still 100% available but running at half capacity, which may impact system performance and user experience, which means fault tolerance is reduced to 50%. To achieve 100% fault tolerance, you must put eight servers, positioning four in each data center.

In mathematical terms, the formula for availability is simple:

https://learning.oreilly.com/api/v2/epubs/urn:orm:book:9781803238951/files/Images/B18475_01_001.png

For example, let’s say you’re trying to calculate the availability of a production system in your company. That asset ran for 732 hours in a single month. The system had 4 hours of unplanned downtime because of a disk failure and 8 hours of downtime for weekly maintenance. So, a total of 12 hours of downtime.

Using the preceding formula, we can calculate the following:

Availability = 732 / (732 + 12)

Availability = 732 / 744

Availability = 0.9838

Availability = 98.38%

It does not matter if your computing environment is on your premises or using the cloud – availability is paramount and critical to your business.

When we deploy infrastructure in an on-premises environment, we have two choices. We can purchase just enough hardware to service the current workload or ensure that there is enough excess capacity to account for any failures. This extra capacity and eliminating single points of failure is not as simple as it may seem. There are many places where single points of failure may exist and need to be eliminated:

* Compute instances can go down, so we need a few on standby.
* Databases can get corrupted.
* Network connections can be broken.
* Data centers can flood or be hit by earthquakes.

In addition to eliminating single points of failure, you want your system to be resilient enough to automatically identify when any resource in the system fails and automatically replace it with an equivalent resource. Say, for example, you are running a Hadoop cluster with 20 nodes, and one of the nodes fails. In that case, a recommended setup is immediately and automatically replacing the failed node with another well-functioning node. The only way this can be achieved on a pure *“on-prem”* solution is to have excess capacity servers sitting ready to replace any failing server nodes.

In most cases, the only way this can be achieved is by purchasing additional servers that may never be used. As the saying goes, *it’s better to have and not need than to need and not have*. The price that could be paid if we don’t have these resources when needed could be orders of magnitude greater than the hardware price, depending on how critical the system is to your business operations.

Using the cloud simplifies the *“single point of failure”* problem and makes it easy to provision resources. We have already determined that provisioning software in an on-premises data center can be long and arduous. However, cloud services like AWS allow you to start up resources and services automatically and immediately when you need them and you only get charged when you start using these newly launched resources. So, we can configure minimal environments knowing that additional resources are a click away.

AWS data centers are built in different regions across the world. All data centers are *always-on* and deliver services to customers. AWS does not have *“cold”* data centers. Their systems are extremely sophisticated and automatically route traffic to other resources if a failure occurs. Core services are always installed in an N+1 configuration. In the case of a complete data center failure, there should be the capacity to handle traffic using the remaining available data centers without disruption.

AWS enables customers to deploy instances and persist data in more than one geographic region and across various data centers within a region. Data centers are deployed in fully independent zones. Data centers are constructed with enough separation between them such that the likelihood of a natural disaster affecting two of them simultaneously is very low. Additionally, data centers are not built in flood zones.

Data centers have discrete **Uninterruptable Power Supplies** (**UPSes**) and onsite backup generators to increase resilience. They are also connected to multiple electric grids from multiple independent utility providers. Data centers are connected redundantly to multiple tier-1 transit providers. Doing all this minimizes single points of failure, and improves availability. You will learn more details about AWS global infrastructure in *Chapter 4*, *Networking in AWS*.

**Faster hardware cycles**

When hardware is provisioned on-premises, it starts becoming obsolete from the instant that it is purchased. Hardware prices have been on an exponential downtrend since the first computer was invented, so the server you bought a few months ago may now be cheaper, or a new version of the server may be out that’s faster and still costs the same. However, waiting until hardware improves or becomes cheaper is not an option. A decision needs to be made at some point to purchase it.

Using a cloud provider instead eliminates all these problems. For example, whenever AWS offers new and more powerful processor types, using them is as simple as stopping an instance, changing the processor type, and starting the instance again. In many cases, AWS may keep the price the same or even cheaper when better and faster processors and technology become available, especially with their own proprietary technology like the Graviton chip.

The cloud optimizes costs by building virtualization at scale. **Virtualization** is running multiple virtual instances on top of a physical computer system using an abstract layer sitting on top of actual hardware. More commonly, virtualization refers to the practice of running multiple operating systems on a single computer at the same time. Applications running on virtual machines are unaware that they are not running on a dedicated machine and share resources with other applications on the same physical machine.

A **hypervisor** is a computing layer that enables multiple operating systems to execute in the same physical compute resource. The operating systems running on top of these hypervisors are **Virtual Machines** (**VMs**) – a component that can emulate a complete computing environment using only software but as if it was running on bare metal. Hypervisors, also known as **Virtual Machine Monitors** (**VMMs**), manage these VMs while running side by side. A hypervisor creates a logical separation between VMs.

It provides each of them with a slice of the available compute, memory, and storage resources. It allows VMs not to clash and interfere with each other. If one VM crashes and goes down, it will not make other VMs go down with it. Also, if there is an intrusion in one VM, it is fully isolated from the rest.

AWS uses its own proprietary Nitro hypervisor. AWS’s next-generation EC2 instances are built on the AWS Nitro System, a foundational platform that improves performance and reduces costs. Typically, hypervisors secure the physical hardware, while the BIOS virtualizes the CPU, storage, and networking, providing advanced management features. The AWS Nitro System enables the segregation of these functions, transferring them to dedicated hardware and software, and delivering almost all server resources to EC2 instances.

**System administration staff**

An on-premises implementation may require a full-time system administration staff and a process to ensure that the team remains fully staffed. Cloud providers can handle many of these tasks by using cloud services, allowing you to focus on core application maintenance and functionality and not have to worry about infrastructure upgrades, patches, and maintenance.

By offloading this task to the cloud provider, costs can come down because the administrative duties can be shared with other cloud customers instead of having a dedicated staff. You will learn more details about system administration in *Chapter 9*, *Driving Efficiency with CloudOps*.

This ends the first chapter of the book, which provided a foundation on the cloud and AWS. As you move forward with your learning journey, in subsequent chapters, you will dive deeper and deeper into AWS services, architecture, and best practices.

**Summary**

This chapter pieced together many of the technologies, best practices, and AWS services we cover in the book. As fully featured as AWS has become, it will certainly continue to provide more and more services to help enterprises, large and small, simplify the information technology infrastructure.

In this chapter, you learned about cloud computing and the key differences between the public and private cloud. This lead into learning more about the largest public cloud provider, AWS, and about its market share and adoption.

We also covered some reasons that the cloud in general and AWS, in particular, are so popular. As we learned, one of the main reasons for the cloud’s popularity is the concept of elasticity, which we explored in detail. You learned about AWS services growth over the year along with it’s key differentiators from other cloud providers. Further, you explored AWS terminology compared to other key players like Azure and GCP. Finally, you learned about the benefits of AWS and the reasons behind its popularity.

AWS provides some of the industry’s best architecture practices under their Well-Architected Framework. Let’s learn more about it. In the next chapter, you will learn about AWS’s Well-Architected Tool and how you can build credibility by getting AWS certified.