KnoDAX

AWS for Beginners

SK Singh
AWS, Kafka, Hadoop, Unix, Oracle, Java Certified
Founder, Software, Cloud, Data Engineer

ISBN

Table of Contents

Take a chance! All life is a chance. The man who goes farthest is generally the one who is willing to do and dare. -- Dale Carnegie

Introduction

I hope you will find this book helpful in understanding AWS. In this book, we will learn AWS's foundational concepts and hands-on with many popular AWS Services. In addition, the book provides an overview of many AWS Services, which you may find helpful in many AWS certification exams, and a broad general understanding of different service offerings from AWS. There are exercises, true/false, and multiple-choice types of questions that will help you review the learning and help increase the degree of retention.

Chapter 1: Cloud Computing This chapter will develop a basic foundational understanding of cloud computing. Basic knowledge of cloud computing is a must to learn AWS.

Chapter 2: What is AWS? In this chapter, we will understand what AWS is and its overview. Then, we will know how it compares with its competitors, AWS Cloud history, AWS use cases, and some big AWS customers.

Chapter 3: AWS Account The next chapter is about the AWS account. In this chapter, we will learn how to sign up for an AWS account, how to set up a budget alert, best practices for an AWS root account, how to secure an AWS account using MFA, AWS free tier, AWS Billing & Cost Management Dashboard, and how to access AWS platform.

Chapter 4: AWS Cloud Infrastructure The following chapter is about AWS cloud infrastructure, which is the foundation of the AWS cloud platform concerning its infrastructure. We will also learn about AWS Regions and Availability Zones, essentially AWS data center-related concepts.

Chapter 5: Elastic Compute Cloud (EC2) The next chapter is about EC2 (Elastic Compute Cloud), an AWS IaaS service to launch virtual servers on AWS. First, we will learn what EC2 is. Next, we will learn how to launch an EC2 instance and set up a web server on EC2.

Chapter 6: Identity and Access Management (IAM) The following chapter is about IAM (Identity and Access Management). IAM is used to create and manage AWS users. In this chapter, we will learn how to create an AWS user, attach an IAM policy, and generate the keys required to access AWS programmatically.

Chapter 7: Simple Storage Service (S3) The next chapter is about S3 (Simple Storage Service). It is a very popular AWS service. First, we will get an introduction to S3. Then, we will learn how to create a bucket on S3 and upload objects on S3.

Chapter 8: AWS Security and Compliance The following chapter is about an introduction to AWS security and compliance. This chapter will provide a high-level understanding of how AWS approaches the security and compliance of deployed applications.

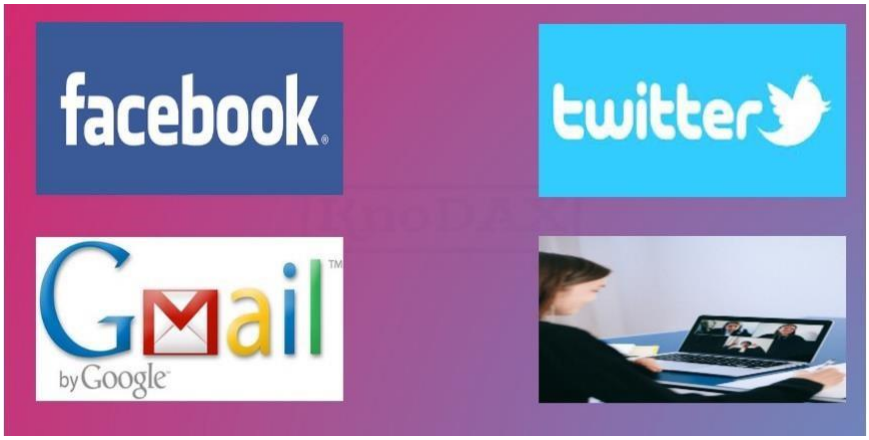
Chapter 9: AWS Cloud Computing Platform This chapter introduces the AWS cloud computing platform covering many popular AWS services.

Chapter 10: Cost-Benefit Analysis This chapter discusses the cost-benefit analysis of moving to a cloud platform. This chapter will also help synthesize your learning from previous chapters.

Chapter 11: Serverless Computing The following chapter is about serverless computing. We will get a good understanding of serverless computing, which is getting popular nowadays**.**

Chapter 12: AWS Services High-Level Overview This is the final chapter. In this chapter, you will get a high-level understanding about understand many AWS Services in different categories. High-level knowledge of these services can help you in many AWS certification exams. You can further explore a particular service(s) based on the use case of your project or the job.
You will also find references to YouTube videos in many chapters where we have related YouTube videos. These videos will further help in making learning more accessible.

This was the introduction and high-level summary of what is covered in each chapter. Let's start with the first chapter: What is AWS? Happy learning!



Chapter 1. What is Cloud Computing?

"Every kid coming out of Harvard, every kid coming out of school now thinks he can be the next Mark Zuckerberg, and with these new technologies like cloud computing, he has a shot." -- Marc Andreessen

◆◆◆

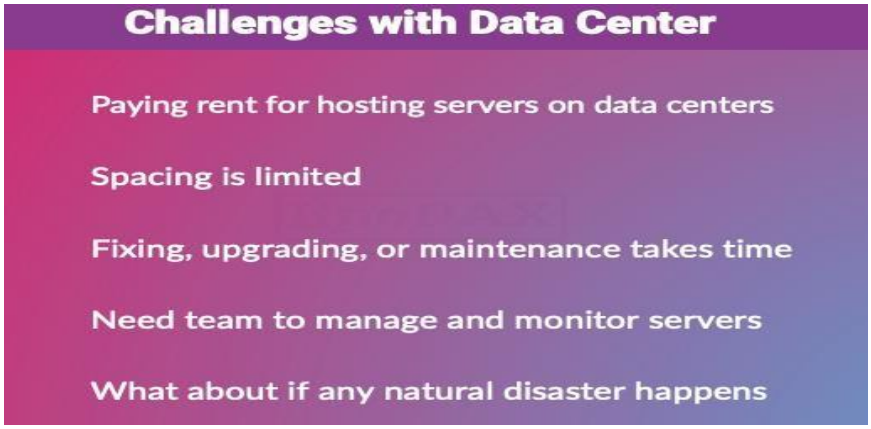
Traditional IT infrastructure

In the late 90s, with the dot com boom, we saw so many startups.
Some of them have become big names, such as Amazon and Google. Hand, however, most of those startups have ed from the so-called garage. First, they started with a few servers. Then, as their user base increased, they needed more machines to scale up their business.

Then, to handle the scalability issue, or in other words, to maintain system performance with the matching workload on the system, they moved their server infrastructure from the garage to the office, where they set up their servers in a so-called computer room or server room. That helped them overcome network bandwidth, power supply, and AC challenges when running the business with more servers.

When the user base increased further, they needed to scale further again. They moved their servers or IT infrastructure to data centers to manage the scalability issue. These data centers have more computing resources, power, air conditioning, security, and other related things that run 24x7 operations of 100s or 1000s servers.

But still, there are challenges and issues with data centers, and what are those? And is there a better solution for this?



challenges with data center

Let's talk about them. Depending on how much space you require for your servers, it costs a lot. And there are reasons for the cost as data centers provide 24 x 7 power supply, AC, maintenance, and security. So, it's obvious there will be a cost to all these services.

There is limited space – each data centers have some limited capacity. Even though data centers have a vast area, the space is limited. If you need to upgrade servers or do some maintenance, you will have to go to the data center (in many cases) to have the part replaced or do an upgrade, etc. You also need to manage and maintain servers 24x7. There is a single point of failure. What if any natural disaster happens?

So, the bigger general question is -- do we have a solution for all these challenges? Is there any other solution besides leveraging data centers for IT infrastructure? And the answer is Cloud Computing. So, let's start with cloud computing.

What is Cloud Computing?

Before understanding the term **cloud computing**, it is essential to know about the word **"cloud,"** as this is an interesting word in this term. Interestingly, the word **"cloud"** in the term cloud computing is not related to the literal cloud. **Instead, the word "cloud" in cloud computing is a metaphor for the Internet.** Thus, cloud (as a metaphor for Internet) computing refers to Internet-based computing in which IT resources are delivered on-demand with a pay-as-you-go pricing model.

Let's talk about a formal definition of cloud computing. According to Special Publication SP 800 – 145 [Sept 2011, Peter Mell (NIST), Tim Grance (NIST)] from The National Institute of Standards and Technology (NIST) of the United States.

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model comprises five essential characteristics, three service models, and four deployment models.

There are some keywords to notice in the NIST definition of cloud computing. These are on-demand network access, a shared pool of configurable computing resources rapidly provisioned and released. On the other hand, in the traditional classic on-premises data center, the computing, storage, and network resources are bought, set up, and permanently configured by the customers at maximum capacity, regardless of needing help. Depending on the business season, this resource allocation may be less. In that case, resources are wasted. However, there is also a possibility that the resources cannot meet demand. In that case, there is the chance of reducing service quality and the risk of losing customers because of quality concerns. There is no demand concept, sharing of the resource pool, and rapid on-demand provision in a classic on-premises data center. Another important point to keep in mind is that cloud computing is predicated upon the idea of purchasing "services" based on the needs of customers - on-demand -- and stopping, closing the service, or terminating when you are done with the usage.

Using cloud computing, organizations (cloud computing providers) offer services such as virtual machines (compute resource that uses software instead of a physical computer), virtual storage (storage pool formed by combining multiple network storage devices), and many other types of software applications (or services) over the Internet. So, for example, if you want to set up a Linux virtual machine, and if you have an account with a cloud provider, you can launch it within a few minutes – just by using the web browser. And start using the Linux VM as you would any regular physical Linux machine, for example, setting up a web server, database, or any other use of Linux machine you do.

In addition to virtual servers, cloud computing providers can also offer virtual storage. For example, if you need extra storage to store an extensive collection of media files, you can use a cloud computing provider's storage service to store them quickly. You just need an account with the cloud provider and a web browser -- no need to shop around to buy the storage and waste additional time to set up the device, such as installing a driver before using the storage.

On the other hand, cloud computing users such as organizations can develop and offer software applications using cloud computing.
(for example, Gmail, Office365, Facebook) or other related services.

What is cloud computing?

In the above discussion, we learned about the term **cloud computing, cloud computing providers,** and **cloud computing users.**

Based on the above discussion, we can see that to launch a virtual machine or get virtual storage, we only need an account with the cloud provider and a web browser. In other words, cloud computing offerings (the common term is services) are provided over the Internet. Nonetheless, in general, there is nothing special about hardware. Cloud computing's underpinning hardware is the same physical server, storage, and network used in on-prem datacenters.

What is cloud computing?

Then, the question comes of how cloud computing differs from classic (non-cloud) computing. The main difference is that cloud computing uses cloud architecture. The architecture combines technology components to help build a cloud that can perform resource pooling through virtualization – running virtual machines as an abstraction layer over a physical device. In other words, cloud architecture helps organize and consolidate massive hardware such as computing resources, storage, and network -- to form resource pooling – and make it available over the Internet.

You may be thinking about why there is so much talk about cloud computing. As you have noticed in the above discussion, cloud computing has many advantages. One aspect of the advantages is that though organizations have been developing, delivering, and managing software for many decades, cloud computing has made this process of developing, delivering, and managing software to end users -- globally --much faster and relatively cheaper (cheaper

may not always be true). The reason is that hardware infrastructure, software tools, and other whole hosts of things required for software development, testing, and deployment can be easily and quickly acquired and set up quickly. Additionally, it could be less expensive --- cloud providers nowadays offer various pricing models. Though cloud computing has many advantages, it may not be appropriate for all use cases. Therefore, you will still need to do your homework if cloud computing is advantageous for your use case or organization.

Cloud Computing Related Terms

Before discussing cloud computing features, services, and deployment models, let's talk about other cloud computing terms and roles, as these terms and roles may be used later. It's better to be equipped with the knowledge of these cloud computing terms and cloud computing roles as these terms are commonly used in cloud computing talks and books.

Cloud Computing Platform

cloud computing platform

The back-end system providing services is called a cloud computing platform.

Cloud Services (Web Services)

cloud services

Another related term is cloud services -- also commonly called web services. Services provided by the cloud computing platform are called cloud services, for example, Gmail and Office365.

Cloud Services definition, which is based on ISO/IEC 17788, "Cloud Computing – Overview and Vocabulary":

One or more capabilities are offered via cloud computing and are invoked using a defined interface.

Cloud Computing Platform Provider

cloud platform provider

We know the terms cloud computing, cloud computing platform, and cloud services (web services). Another related term to know is cloud computing platform provider. Cloud providers such as AWS, Google, Microsoft, IBM, Oracle, Salesforce, SAP, and others that provide cloud services from their cloud computing platform are called cloud computing platform providers (also commonly called cloud services providers or cloud providers). AWS, Google, and Microsoft are the leading cloud computing platform providers.

As a side note, sometimes you will notice that the word "computing" may be missing in some casual or informal cloud computing discussions. For example, you might hear cloud service(s) as opposed to cloud computing service(s), cloud provider(s) as opposed to cloud computing provider(s), or cloud platform(s) as opposed to cloud computing platform(s). But that doesn't change their semantics.

Below are some other cloud computing-related terms, and their definitions are given. These definitions are based on ISO/IEC 17788,
"Cloud Computing – Overview and Vocabulary."

Availability

Property of being accessible and usable upon demand by an authorized entity.

Confidentiality

Property that information is not made available or disclosed to unauthorized individuals, entities, or processes.

Integrity

Property of accuracy and completeness.

Information Security

Preservation of confidentiality, integrity, and availability of information. In addition, other properties, such as authenticity, accountability, non-repudiation, and reliability, can also be involved.

Service Level Agreement (SLA)

Documented agreement between the service provider and customer that identifies services and service targets. A service level agreement can also be established between the service provider and a supplier, an internal group, or a customer acting as a supplier. A service level agreement can be included in a contract or another documented agreement.

Cloud Application

An application that does not reside or run on a user's device is accessible via a network.

Cloud Service Provider

Party (Natural person or legal person, whether or not incorporated, or a group of either) makes cloud services available.

Cloud Service Customer

Party (Natural person or legal person, whether or not incorporated, or a group of either) which is in a business relationship to use cloud services.

Cloud Service User

A natural person, or entity acting on their behalf, is associated with a cloud service customer that uses cloud services. Examples of such entities include devices and applications.

Measured service

Metered delivery is of cloud services such that users can be monitored, controlled, reported, and billed.

Tenant

One or more cloud customers share access to a pool of resources.

Multi-tenancy

Allocation of physical or virtual resources means multiple tenants and their computations and data are isolated and inaccessible to one another. We will talk about multi-tenancy in another chapter in some detail.

On-demand Self-service

Feature where a cloud service customer can automatically provision computing capabilities as needed or with minimal interaction with the cloud service provider.

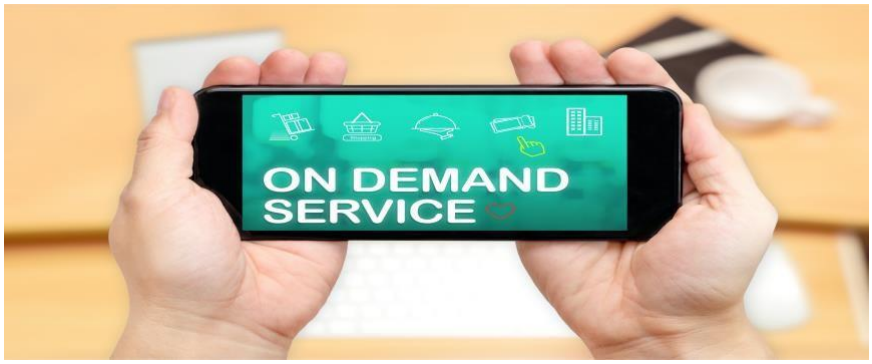
Resource pooling

Aggregation of a cloud service provider's physical or virtual resources to serve one or more cloud service customers.

Cloud Computing Key Features

Let's talk about key features or characteristics of a cloud computing platform. On-demand service, network access, resource pooling, elasticity, metered service, multitenancy, availability, and scalability are the key features of cloud computing. Let's discuss each of these features in a bit more detail.

On-demand Service



Offering cloud services on-demand by cloud providers to its customers is one of the key features of cloud computing. The most common mechanism to provide on-demand service is Web UI. This is not the only way; the other ways are APIs, Command Line Interface (CLI), and programmatic ways such as using SDK. If you take the example of AWS, AWS users can launch services using Web UI, AWS API, AWS CLI, and AWS SDK.

Network Access

Network access, more specifically Internet access, is another critical feature of cloud computing. In other words, cloud providers must offer services over the Internet to be called actual cloud providers.

Resource Pooling

Another critical feature of cloud computing is resource pooling. The main driver for the innovation of cloud computing was efficiently utilizing a vast pool of idle resources and generating a business model for it. Cloud providers create a resource pool of computing, storage, and network resource of different types, shapes, and sizes and provide suffering offerings from the resource pool based on what has been requested. The beauty of this is that when the resource is released from the customer, it returns to the resource pool and is ready to be served to another customer.

Elasticity

In cloud computing, you will come across the term elastic a lot. For example, AWS has many services that include the time elastic in its name, for example, Elastic Compute Cloud, Elastic Load Balancer, Elastic Block Storage, and Elastic MapReduce.

The term elastic in cloud computing is analogous to an elastic band. You can stretch an elastic band size beyond its rest state, and when you let it go, it will return to its resting length.
This elastic concept -- going back to its resting size when we let it go from its stretched state -- is instrumental in cloud computing.

Let's take an example to understand the term elastic as it relates to cloud computing on AWS. The hypothetical use case is related to setting up a scalable web server. We set up the web server with a minimum of 3 and a maximum of 6 EC2 instances. Each EC2 instance will be launched using a custom AMI to launch the Apache webserver. We have also configured AWS Elastic Load Balancer to launch additional EC2 instances if CPU utilization for an EC2 instance reaches above 70% on AWS Cloud Watch – a maximum of up to six instances. And terminate the EC2 instance when the CPU utilization comes down to less than 70% -- minimum of up to three instances. As you can see, we have set up a scalable and -- elastic -- web server.

Metered Service

Metered Service is an essential feature of cloud computing. This concept is very similar to what we have already experienced, such as paying for gas utility bills based on several units of energy consumed or parking meter bills based on vehicle parking time. There are many examples of metered service in our daily life. The metered service of cloud computing is one of the key drivers that are so popular today in technology – you pay what you use, how much you use.

Let's try to understand with a concrete example of metered service in cloud computing. Say you need around 3TB storage to store some videos for a month – maybe to share with your friends or relatives or may by trying out some new business model. Let's first see what we can do without leveraging a cloud computing solution. We would try to find out some machine with 3TB space to make sure to have this machine be available on the Internet. Sounds easy, technically – correct?

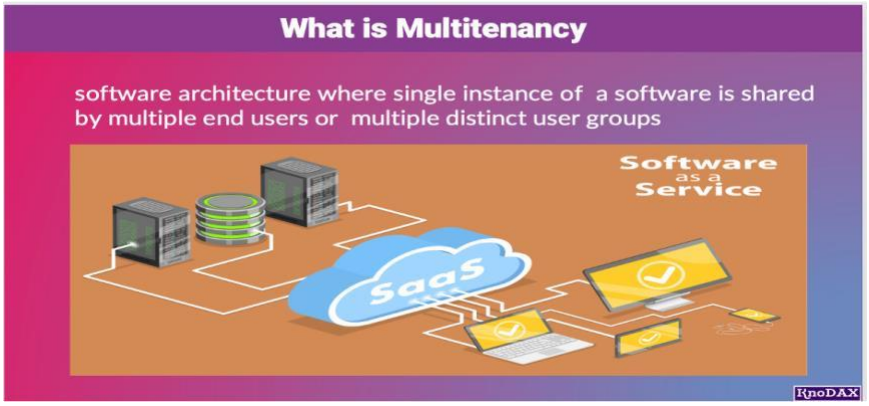
How much would it cost? It depends how much on your rate and billable hour. Of course, in real life, you will not calculate it – but there is a hidden cost to doing all this setup. In some cases, you may not have the hardware or have the hardware but not have the storage space. There is also the possibility that you may have an Internet connectivity issue. You get the idea that getting just 3TB storage space to share your videos with your friends and relatives sounds like busy work.

Now let's see how easy to set up this use case on the cloud. We will consider AWS storage service S3 (you will learn later about S3). So, the question is how you would set up this use

case. You register with AWS if you have. Once you have an account with AWS, you find the S3 service, create a bucket and upload all your videos to the bucket. You will get an URL that you can share with your friends. You pay based on how long you have videos on the S3 and what the size and AWS pricing were – usually, it is per GB / month.

With the metered service feature, you pay for the cloud services based on your usage and pricing. This is a beautiful feature of the cloud as it saves many costs in many use cases for almost all types of organizations and individuals trying out or learning the cloud.

Multitenancy



Multitenancy is another critical feature of cloud computing. Let's understand this feature in a bit more detail.

What is Multitenancy? Multitenancy is a software architecture where an instance of a single software can be used by multiple end users or multiple distinct user groups. SaaS software, such as Salesforce, Google Gmail, Microsoft Office 365, and TurboTax, are typical examples of multi-tenancy.

Let's see what single tenancy is. This will further solidify your understanding of multitenancy.

As the name suggests, it is the opposite of multitenancy. In a single tenant, each end-user or each group of users uses its software instance. There are plenty of examples. Take an example of tax software; for instance, TurboTax has its SaaS version. They also have their old classic desktop version, which you can buy, install, and use as a single user or tenant. The typing software can be another example. There are many SaaS software to practice typing, but you can also purchase a single-tenant desktop version of typing software.

The multitenancy concept is not new in the mainframe era, which was around the 1960s. To share mainframe computing resources among multiple users, timeshare software was used. Cloud computing now uses the same multitenancy idea to allow sharing of computing resources – particularly in the public cloud computing deployment model. The pool of

computing resources – processing power and memory – is divided among multiple users or multitenant in the public cloud. This multitenancy is at the server level.

Multitenancy saves costs and enables flexibility. Concerning protecting cost advantage, the reason is apparent. Since the computing resources are consolidated and shared among multiple users or clients, this sharing helps lower costs for individual users in a multitenant environment. For example, if you use the TurboTax SaaS version instead of TurboTax Desktop or the single-tenant version, using the SaaS version software for tax filing is much cheaper than buying the TurboTax desktop single-tenant software. Another advantage of multitenancy is that it enables flexibility. As we know, doing estimation is a challenging exercise. If you over-provision, the cost will go high.
On the other hand, if you under-provision, then your output would suffer. But in a multitenant environment, you only pay for what you use. Also, you would be free to manage the resources, such as applying a patch and securing them as the provider takes care of resource management.

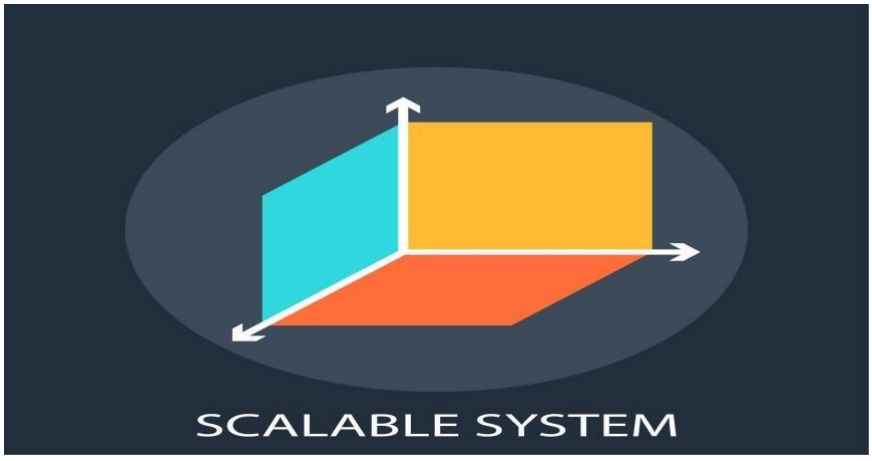
Availability

First, let's understand availability and reliability. Availability measures the percentage of time the system is in operable condition. On the other hand, reliability measures how long the system performs its intended function without breaking it down.

Say you have a machine that shuts down once in one hour, and it's down for 6 minutes before becoming operable again. In that case, the machine availability is (60 - 6) / 60 = 90%; and it has the reliability of 1 hour because it can go down once within an hour. It can be considered highly available if the machine or software goes down but comes back immediately by rebooting or restarting itself.

Just keep this in mind. A reliable system has high availability, but an available system may not be reliable.

Scalability



Scalability is the other important attribute of cloud computing. Scalability is a crucial quality attribute to pay attention to about software, particularly in cloud computing, where request load is often uneven.

Having said this, what is scalability in software? Scalability in software refers to the quality attribute of the software that measures the system's ability to perform as the load or the number of requests increases. If the software performs without degradation of its performance as stated in its service level agreement, then it is considered scalable. On the other hand, if the system performance is impacted negatively, it is called a non-scalable system.

If we didn't manage scalability, the performance of the software would degrade if the load on the system increased. For example, suppose we deployed a simple e-commerce web application on a machine with 2x CPUs with 8 GB RAM. Say the application can handle 100 concurrent requests per second. Suppose due to some deal, the traffic on the system has increased. If we didn't manage the increase in the request, the application's performance would degrade – which means the system will start to behave poorly as it needs to process more than 100 concurrent requests per second. This could cause the CPU / RAM/ IO utilization to reach its maximum limit. As a result, the system would stop processing any further requests. This type of system can be called non-scalable.

There are two approaches to managing software scalability: horizontal scalability and vertical scalability.

To make a system vertical scalable, we replace the existing system with a higher configuration system or increase the existing system RAM, CPU, and HDD. To manage the scalability of the software using horizontal scalability, instead of increasing the system resources or migrating the software to a server that is more powerful in terms of RAM, storage, or clock speed, we distribute the requests on multiple servers to maintain the software performance.

Cloud Computing Deployment Models

The cloud deployment model describes how a cloud computing platform is implemented and hosted and who has access to it. For instance, what is the accessibility of computing resources? Are all computing resources available to the public? Or not all -- only some are available to the public, and some computing resources are protected. How are computing resources accessed, managed, audited, or governed? How is the security of computing resources managed? Understanding computing resources in terms of these aspects of cloud computing comes under the term cloud computing deployment model (or cloud deployment model).

There are mainly four types of cloud deployment models: public cloud, private cloud, hybrid cloud, and multi-cloud.

In **Public Cloud Deployment Model**, all the components of an application are fully deployed on a cloud platform. The application's components may have been developed in an on-premises data center and migrated to the cloud next, or the application's components have been developed on the cloud platform and then deployed on the cloud platform.

Private Cloud Deployment Model refers to a cloud environment built on-premises data center using virtualization technology such as virtual machine Docker. Usually, private clouds are made to meet specific security and regulatory needs along with extremely critical network latency. These requirements may be difficult to fulfill on the public cloud.

In **Hybrid Cloud Deployment Model**, an organization deploys some part of the application's component on the public cloud, and the rest of the application, which needs more security, regulation, compliance, or tight control, stays in the organization's private cloud. Hybrid can also be used where there is a requirement of burst capacity in which workloads are "spilled over" to a public cloud to meet capacity demands for a short period. Because of unpredictability, it may not be a good idea to purchase the capacity in advance as the resources will be underutilized once the requirement for the capacity completes. Another use case of the Hybrid Cloud Deployment Model is a highly available or disaster recovery environment where the public cloud offers much flexibility in getting the resources in the event of data center failure. Essentially, the client may not need upfront investment for alternate site options in the event of data center failure; the client can leverage the on-demand feature of the public cloud.

In Multi-Cloud **Deployment Model**, an organization utilizes more than one cloud provider to deploy an application's components; for example, an organization may use some services of AWS and some benefits of Google based on pricing, and availability of services, forming a multi-cloud deployment model.

Cloud Computing Service Categories

There are many names for this term -- cloud computing service categories, cloud service categories, cloud computing delivery models, cloud delivery models, cloud computing platform types, cloud platform types, cloud computing types -- and it's difficult to say which

is more or most common. You may find any of these terms in this book, but they all mean the same. With this note, let's start understanding this topic.

A cloud computing platform is a back-end system that provides services over the Internet. The question is what kind of services the platform offers. For example, does the cloud computing platform provide Gmail or Office 365? Does it provide virtual servers and virtual storage? Or does the cloud computing platform offer database services over the internet? Depending on the cloud computing platform's kind of service, it has been categorized into a type. This categorization is called cloud computing platform types or cloud computing types.

Continuing our discussion about cloud computing platforms: cloud computing platform, the back-end system providing services, is a general term. To be more specific, there are three main types of cloud computing platforms: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-service (SaaS).

In addition to these main ones, other modern cloud computing platform types have emerged recently, such as Data-as-a-Service (DaaS), Desktop-as-a-Service, and Function-as-a-Service (FaaS). These modern cloud computing platform types provide more fine-grained services and are getting popular quickly.

Infrastructure-as-a-Service (IaaS)

Infrastructure-as-a-Service

One of the main types of cloud computing platforms is Infrastructure-as-a-Service, which is also called IaaS, is short. IaaS provides foundational services called technology infrastructure, which can be provisioned, managed, and maintained over the Internet. In other words, IaaS provides technology infrastructure components.

For example, IaaS offers virtual servers, virtual storage, and a virtual network as a service. As you can notice in the given picture, in the cloud computing pyramid diagram, IaaS is at the foundation. What it means is that IaaS acts as the foundation for the cloud computing platform.

The following is the definition of IaaS from the NIST:

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer can deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of selected networking components (e.g., host firewalls).

Platform-as-a-Service (PaaS)

Platform-as-a-Service

Platform-as-a-Service (PaaS) is another primary type of cloud computing platform or cloud computing type. PaaS provides platform technology infrastructure-related services, for example, databases, web servers, and messaging, to build, test, and deploy software. In other words, PaaS offers complete development and deployment environment in the cloud.

The following is the definition of PaaS from the NIST:

The capability provided to the customer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The customer does not manage or control the underlying cloud infrastructure, including network, servers, operating systems, or storage, but has to control the deployed applications and possibly configuration settings for the application-hosting environment.

Software-as-a-Service (SaaS)

Have you used Gmail? Did you happen to watch movies on Netflix? Do you have a Facebook account? Does your workplace use Zoom for meetings? Have you used Microsoft Office 365? If your answer is "Yes" to any of these questions, essentially, you are using software-as-a-service (SaaS).

Software-as-a-Service advantages

Like IaaS and PaaS, SaaS is another main cloud computing platform or cloud computing type. If IaaS is about infrastructure, PaaS is about the platform – then SaaS is about software. In SaaS, software solutions are delivered as a service over the Internet. Therefore, SaaS software is mostly executed directly within a web browser. This feature of SaaS eliminates the need to install or download software to run it.

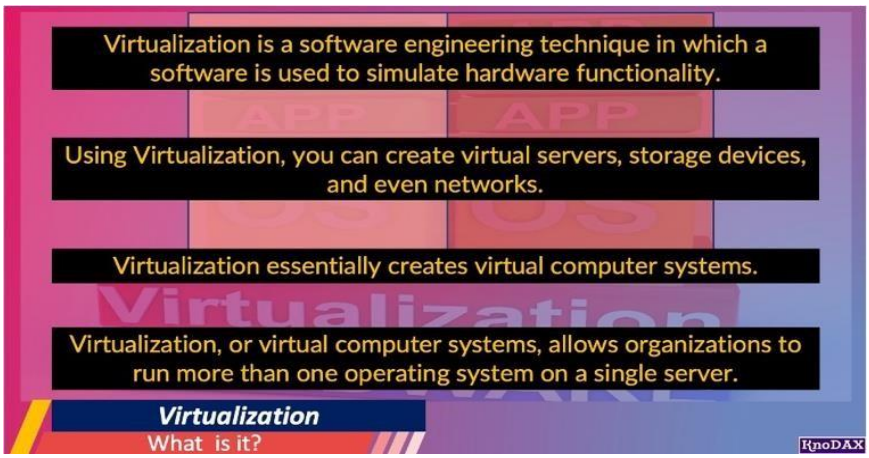
For example, Gmail, Netflix, Facebook, Zoom, and Microsoft Office 365 are some common examples of SaaS. There are countless examples of SaaS, but I'll limit it to a few to keep it simple. To illustrate further, to use Gmail, you don't need to install Gmail on your local computer. You open a web browser, type the Gmail Web URL in the address bar, and start using Gmail. On the same token, to watch a movie on Netflix, since Netflix is a SaaS solution or SaaS software, you don't need to install Netflix software on your local computer. Just open a web browser, type the Netflix web URL in the address bar, and you're ready to start watching movies on Netflix. As you can notice from these illustrations, in SaaS, you don't need to install or download software to execute it.

The following is the definition of SaaS from NIST.

The capability provided to the customer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface, such as a web browser (e.g., web-based email) or a program interface. The consumer does not manage or control the underlying cloud infrastructure, including network, servers, operating systems, storage, or even individual application capabilities, except for limited user-specific application settings. With the understanding of cloud computing that underpins AWS architecture, let's dive into AWS.

Virtualization

Virtualization essentially creates virtual computer systems. Virtualization, or in practical terms, virtual computer systems, allows organizations to run more than one operating system on a single server. As a result, virtualization helps in reducing physical servers' needs.



What is virtualization

As you can see, virtualization is a game-changer with respect to saving costs in buying and maintaining physical servers. Typically, we run one operating system on one server. However, since more than one operating system can be run on single physical hardware in virtualization, organizations can reduce their need to buy and maintain physical servers. This is because virtualization helps them consolidate their servers' needs in fewer servers.

Why is Virtualization Needed?

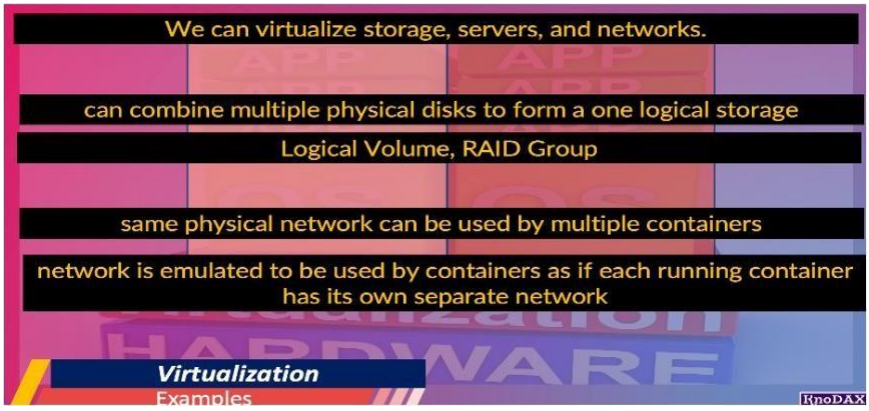
Why is virtualization needed?

Let's continue our discussion about virtualization further, imagine a scenario suppose we have a server that is being utilized minimally. Wouldn't it be better to utilize it in some way where we can use this server's resources to create another server inside it? That's the basic idea behind virtualization.

Let's take another example. As we know, maintaining a consistent SLA is very important in critical applications. How can we achieve consistent SLA when running multiple applications on the physical server? Maintaining consistent SLA would be a guessing game as each application would have to compete with other applications' processes for the resources. One

way to handle this is to run each application in a separate isolated environment on the same physical server. That way, the application would not have to compete with other processes for the resources. Running applications in different environments would help provide consistent service level agreement (SLA). We can use the Virtualization technique to create a different independent running environment for each running application on the same physical server.

What can be Virtualized?



What can be virtualized

Let's talk about what we can virtualize. We can virtualize servers, storage, and networks. This means these hardware constructs can be created in software form using virtualization.

Using Virtualization, we can run multiple servers on the same physical server. These virtual servers are called virtual machines or VMs. We will talk about virtual machines later in the chapter.

virtualization examples

For example, we can run Windows and Linux operating systems as virtual machines in two entirely different environments on a single physical machine. Each VM would have its RAM, storage, and network.

Not only using Virtualization can we run multiple separate operating systems on the same physical server, but also, using Virtualization, we can run multiple applications in a completely separate isolated environment on the same physical machine. This type of virtualization is called containerization, for example, Docker container. We will learn about Docker later in this book.

Besides server virtualization, storage can also be virtualized using the Virtualization technique. For example, multiple physical disks can be combined to form one logical storage (virtual storage), which can be assigned to a server. Examples are Logical Volume and the RAID (Redundant Array of Independent/Inexpensive Disks) group.

In addition to server and storage virtualization, the network can be virtualized using the Virtualization technique. Using network virtualization, a physical network can be used by multiple containers (separate runtime environments) running on the same physical server. The physical network is emulated so that it would be used by various containers as if each running container has its separate network.

Another type of virtualization is desktop virtualization. Desktop virtualization enables multiple desktop machines on a single physical server. This is also called desktop-as-a-server.

With the understanding of cloud computing that underpins AWS architecture, let's dive into AWS.

Virtual Machine

Now we got an understanding of Virtualization. As we talked about, one of the advantages of virtualization is that we can run multiple instances of operating systems -- also called virtual servers or VM – on single physical hardware. The virtualization technique used to create virtual servers, such as Windows or Linux servers, is called a virtual machine.

The virtual machine is also called a virtual computer system, or VM, the more popular term for virtual machines.

virtual machine features

We can think of a virtual machine or VM as a separate isolated container having its operating system and applications. VMs are discrete, disconnected, isolated, self-contained, and completely independent.

Because they are self-contained and completely independent, we can launch multiple VMs on a single physical server. For example, we can have a Linux virtual machine and a Windows virtual machine, which can be run on a single physical server in a separate isolated environment.

Hypervisor

As we discussed, using virtualization, we can run multiple instances of operating systems on the same physical hardware. In other words, using virtualization, we can set up virtual machines. Now the question is: how do virtual machines -- as they run on the same physical hardware -- get the computing resources such as processors, memory, or storage?

There is a concept called Hypervisor or Virtual Machine Monitor, using virtual machines to get computing resources such as processors, memory, or storage. The hypervisor is software that creates and manages virtual machines and mediates communication between hosts and virtual machines.

With the understanding of cloud computing that underpins AWS architecture, let's start with what AWS is in the next chapter.

Related YouTube Videos

- What is Cloud Computing: https://youtu.be/Ir7oo_S_3jo
- Cloud Computing Types: https://youtu.be/DMePTTvmsZ0
- Infrastructure-as-a-Service (IaaS) : https://youtu.be/UXGNafGjBQQ
- Platform-as-a-Service (PaaS): https://youtu.be/mszigptiVQI
- Software-as-a-Service (SaaS): https://youtu.be/yL5AhrTO6Is
- Desktop-as-a-Service: https://youtu.be/MUqHwm5PRoc
- Data-as-a-Service: https://youtu.be/32YvzBQYP9g
- Function-as-a-Service (FaaS): https://youtu.be/Qs5EmkB5s_I
- Cloud Computing Deployment Models: https://youtu.be/FAOSS8A9-Rg
- Multitenancy: https://youtu.be/-_yJjAmVlXU

Chapter Review Questions

For the questions given below, please mark them if they are true or false.

1. The word "cloud" in cloud computing is a metaphor for
2. "the Internet." True / False

3. The term "cloud computing" refers to Internet-based computing.
4. True / False

5. Cloud computing is essentially about providing cloud services over the Internet. True / False

6. The cloud architecture enables cloud providers to organize and consolidate massive hardware, such as computing resources, storage, network, and software, to make it available over the Internet. True / False

7. AWS, Google, and Microsoft are cloud providers because they provide public cloud services. True / False

8. The back-end system providing cloud services is called a "cloud computing platform." True / False

9. Availability and Reliability are the same concepts just two different names. True/ False

10. An elastic system adds or removes resources based on how it has been configured. True / False

11. Multitenancy is a feature of a cloud computing platform that enables more than one user to access the software deployed on the cloud concurrently. True / False

12. SaaS software is mostly executed directly within a web browser. True / False

13. Function-as-a-Service (FaaS) is not a good solution for method or function calls, which have a dynamic workload that fluctuates considerably. True / False 1. A public cloud generally has a limited amount of computing resources and storage available, and it is not easily scalable. True / False

14. Organizations with solid security and regulatory requirements, such as banks and healthcare providers, prefer private cloud – particularly total on-premises cloud solutions. True / False

15. In a hybrid cloud, integration is involved between private and public clouds. This can cause potential performance issues because of network latency and security risks as data are shared between public and private clouds. True / False

16. The multi-cloud model use case generally fits into a larger organization, where one department's cloud needs and budgets may not be aligned with the other departments. True / False

17. Multitenancy is a software architecture where multiple end-users or distinct user groups can use an instance of a single software. True / False

Please select the correct answer from the given choices for the questions below.

1. What is cloud computing?

   a Cloud computing means providing virtual servers over the Internet.
   b Cloud computing means providing virtual storage over the Internet.
   c Cloud computing means providing software as a service over the Internet
   d All the above

2. Which of the following options is the feature of cloud computing?

   a Metered billing model
   b On-demand service
   c Scalability
   d All the above

3. Which of the following organizations is not a cloud service provider?

   a. Google
   b. Apple
   c. Microsoft
   d. Amazon

4. Which of the following statements is not correct?

    a. Cloud computing can be a good choice for applications having highly scalable requirements.
    b. Cloud computing can be a good choice for applications in need of reducing costs on their IT infrastructure.
    c. Applications with low solid latency, security, audit, and regulatory SLA can be the right fit for cloud computing.
    d. Cloud computing can be a good choice for applications having high availability requirements.

20. Which of the following architectures is not related to cloud computing?

    a. Micro-Services
    b. Service-Oriented Architecture (SOA)
    c. Monolith
    d. None of them

21. Which of the following statements is true about cloud computing?

    a. cloud computing provides virtual servers
    b. cloud computing helps to get rid of on-prem data centers for computing needs
    c. cloud computing helps cut costs on maintenance staff for 24x7 operations
    d. all the above

22. Which options will you select to make a system horizontally scalable?

    a. Increase RAM size
    b. Add a new machine to the cluster
    c. Increase CPU clock speed
    d. None of them

23. Which options are the advantages of a flexible pricing model in cloud computing?

    a. The flexible pricing model helps customers to get unlimited bandwidth.
    b. The flexible pricing model enables customers to pay for what they use.
    c. The flexible pricing model enables customers to get storage free but will be changed for virtual servers.
    d. The flexible pricing model enables customers to add resources dynamically when the resources are needed without any extra charge.

24. Cloud computing environments can be of both types- multi-tenant or single-tenant- though the multi-tenant type of cloud computing environment is more common. Which of the following options is an advantage of a multi-tenant cloud computing environment over a single-tenant cloud computing environment?

    a. enhanced data security
    b. faster performance
    c. cost savings
    d. all of them

25. Which of the following SaaS software is a multi-tenant type?

    a. Microsoft Office 365
    b. Facebook
    c. Gmail
    d. all of them

26. A startup company has developed a web application that is doing extremely well with local customers. Looking at the popularity of the software, the company is thinking of rollout this application worldwide. The company engineering president is considering deploying this application on a cloud computing platform. Which of the following attributes of cloud computing helps decide if cloud computing is the right choice for deploying the application on the cloud platform?

    a. Availability
    b. Scalability
    c. Flexible pricing such as pay-as-you-go or metered pricing model
    d. All the above

27. A start-up software organization would like to test and deploy its software solutions on the cloud platform to save cost in avoid buying and maintaining expensive servers. The company is looking for a cloud provider which offers virtual server provisioning and on-demand storage services. Which of the following options of cloud computing delivery models is the startup company looking for?

    a. Software-as-a-Service
    b. Platform-as-a-Service
    c. Application-as-a-Service
    d. Infrastructure-as-a-Service

27. In which distribution model of cloud computing a software application is hosted on the cloud, and users can access the software using the Internet?

a. Software-as-a-Service
b. Platform-as-a-Service
c. Infrastructure-as-a-Service
d. all of them

28. Which of the following statements is correct with respect to a use case of cloud computing?
    a. A company has several hundreds of documents that need to be indexed in a few minutes.
    b. A company needs a CRM solution as its customer base is increasing, and it would like to provide the best customer service to its customers. However, it doesn't have time to build its home-grown solution for CRM as they don't have the resource and time for it.
    c. A company engineering team needs servers to try out some POC type of work for 2-4 weeks. The servers are needed -- lay idle -- when POC is complete.
    d. all of them

29. Which of the following statements is true about an application service provider?
    a. An application service provider uses the software-as-a-service delivery model to provide software as a service over the Internet.
    b. The provider essentially provides virtual servers over the Internet
    c. The provider provides platform-as-a-service.
    d. All of them

30. You are a cloud engineer in a start-up organization. You have asked to provide access to virtual machines in a cloud environment. Which of the following cloud computing delivery models would you use?

    a. Platform-as-a-Service
    b. Infrastructure-as-a-Service
    c. Function-as-a-Service
    d. Software-as-a-Service