What is Cloud Computing?

- **Cloud computing is the on-demand delivery of IT resources (such as computing power/servers, storage, networking, databases etc.) over the Internet with pay-as-you-go pricing.**
- These resources run on server computers that are located in large data centers in different locations around the world.
- When you use a cloud service provider like AWS, that service provider owns the computers/IT resources that you are using.

Benefits of Cloud Computing

- **Scalability:** Cloud computing allows to easily scale their computing resources up or down as their needs change, without having to purchase and manage additional hardware.
- **Cost Savings:** Using cloud computing can be more cost-effective than maintaining on-premises IT infrastructure, as it eliminates the need for costly hardware, software, and maintenance expenses.
- **Accessibility:** Cloud computing enables remote access to applications and data, allowing users to work from anywhere with an internet connection.
- **Reliability:** Cloud providers offer high levels of uptime and redundancy, ensuring that applications and data are available even in the event of hardware failure.
- **Flexibility:** Cloud computing offers a wide range of deployment options, including public, private, and hybrid clouds, which can be tailored to meet the unique needs of an organization.

Traditional Computing Vs. Cloud Computing

- Infrastructure is thought of as hardware.
- Traditional computing uses hardware, software, and other related services from a centralized location.
- This invests in hardware, software, and other services and personnel to manage and maintain them. Traditional computing can be expensive
- They require space, staff, physical security, planning, and capital expenditure.
- Traditional Computing involves the long hardware procurement cycle that involves acquiring, provisioning, and maintaining on-premises infrastructure.
- With a hardware solution, you must ask if there is enough resource capacity or sufficient storage to meet your needs, and you provision capacity by guessing theoretical maximum peaks.

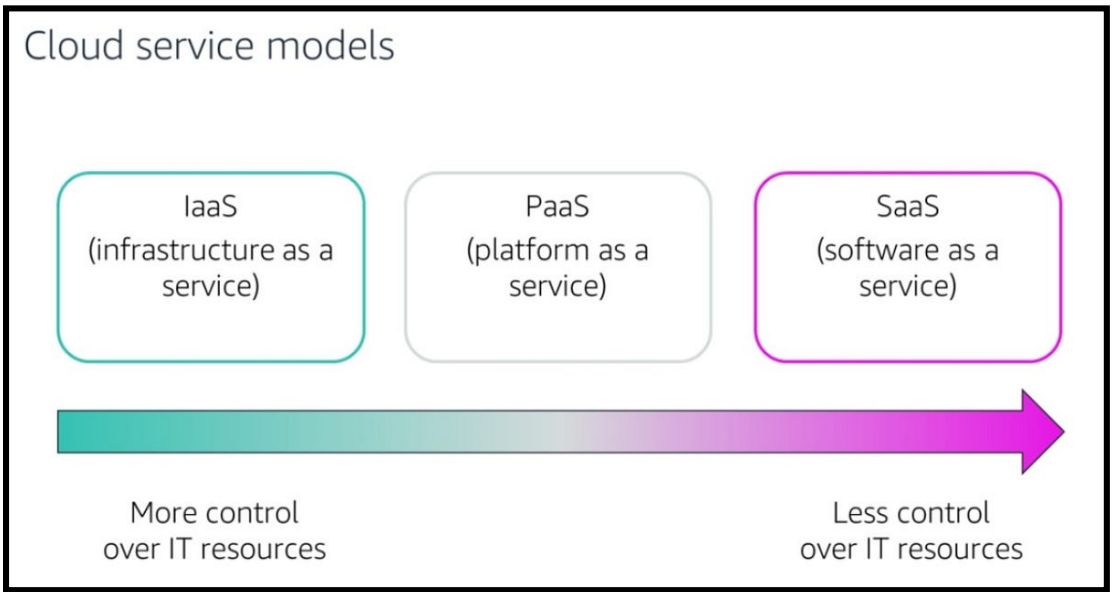
- For example, if you wanted to provision a new website, you would need to buy the hardware, rack and stack it, put it in a data center, and then manage it or have someone else manage it. This approach is expensive and time-consuming.

Cloud Computing

- Cloud computing enables you to think of your infrastructure as software.
- Software solutions are flexible. You can select the cloud services that best match your needs, provision and terminate those resources on-demand, and pay for what you use.
- You can elastically scale resources up and down in an automated fashion. With the cloud computing model, you can treat resources as temporary and disposable.
- Compared to hardware solutions, software solutions can change much more quickly, easily, and cost-effectively.
- Cloud computing helps developers and IT departments avoid undifferentiated work like procurement, maintenance, and capacity planning, thus enabling them to focus on what matters most.

Cloud Service Models

There are three main cloud service models. Each model gives you a different level of control over your IT resources:



IaaS – Inrastructure as a Service

- In the IaaS service model, cloud service providers' delivers on-demand infrastructure resources via the internet, such as computing devices, servers, storage, networking etc.

Customers don't have to manage, maintain, or update their own data center infrastructure.

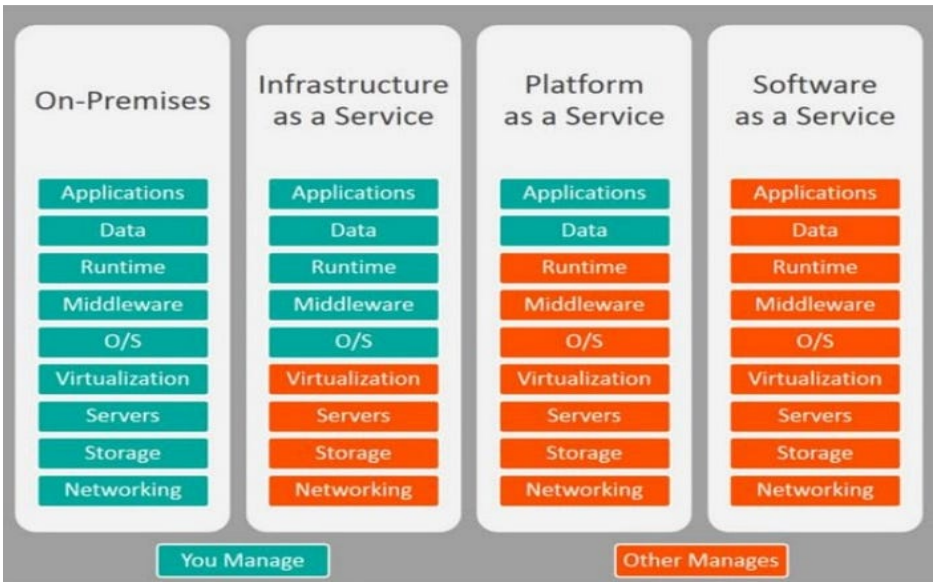
- We manage, maintain and update the servers and the operating system on our own.

PaaS - Platform as a Service

- In the PaaS service model, delivers and manages all the hardware and software resources to develop applications through the cloud.
- Developers and IT operations teams can use PaaS to develop, run, and manage applications without having to build and maintain the infrastructure or platform on their own.
- Customers still have to write the code and manage their data and applications, but the environment to build and deploy apps is managed and maintained by the cloud service provider.
- PaaS provides frameworks to work and build on for developers.

SaaS - Software as a Service

- Software as a service, or SaaS, provides the entire application stack, delivering an entire cloud-based application that customers can access and use.
- SaaS products are completely managed by the service provider and come ready to use, including all updates, bug fixes, and overall maintenance.
- Most SaaS applications are accessed directly through a web browser, which means customers don't have to download or install anything on their devices.
- Files and applications are all that we manage here.
- Twitter and Gmail are example of SaaS.



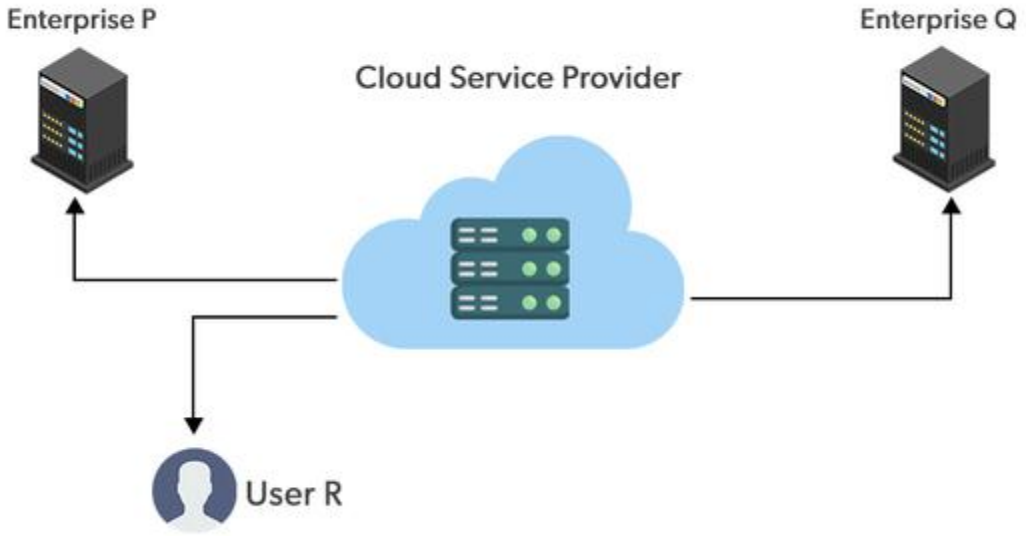
Key Difference between IaaS, Paas, SaaS

Parameter	IaaS	PaaS	SaaS
Definition	It provides virtualized computing resources (server, storage, and networking) over the internet.	It gives developers a platform to create and deploy apps without having to manage the underlying infrastructure.	It gives users online access to software applications that are hosted and controlled by a different supplier.
Stand for	Infrastructure as a Service	Platform as a Service	Software as a Service
Used By	Network Architect	Developers	End Users.
Pros	Affordable, flexible, accessible, reliable	Cost-effective, increased productivity, easy scalability, easy accessibility	Scalable and accessible, affordable, easy to upgrade, easy deployment
Cons	Lack of control, Data security issue	Compatibility issue, vendor changes	Insufficient data security, less control
Example Services	Amazon Web Service (AWS), Microsoft Azure, Google Cloud Platform	Heroku, OpenShift, Beanstalk, Apache Stratos	Google Workspace, Salesforce, Dropbox

Cloud Deployment Models

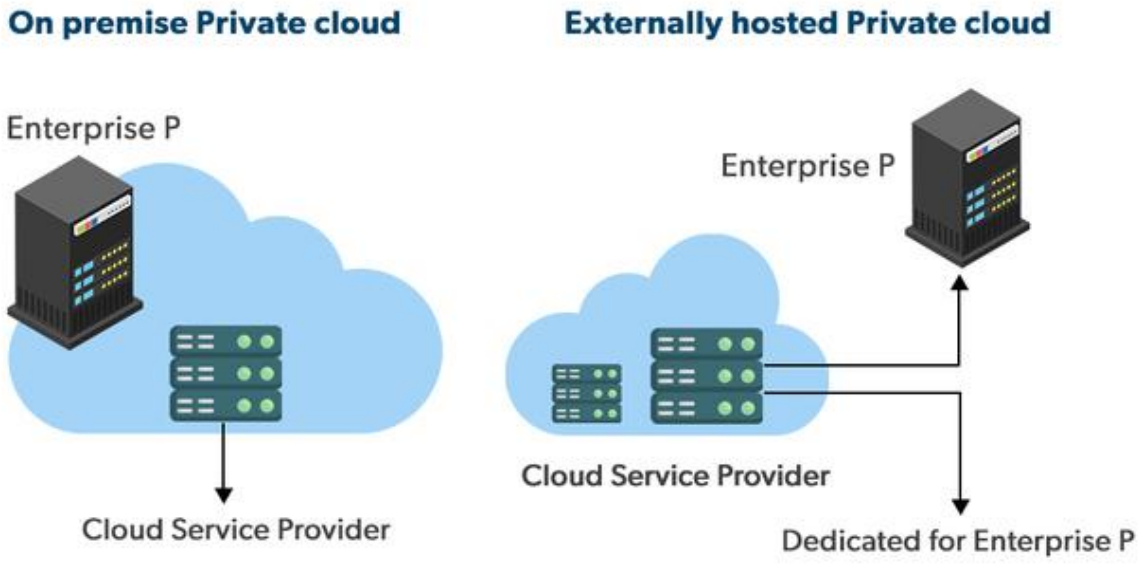
Public Cloud

- The public cloud makes it possible for anybody to access systems and services.
- The public cloud may be less secure as it is open to everyone.
- The public cloud is one in which cloud infrastructure services are provided over the internet to the general people or major industry groups.
- For example, Google App Engine etc.



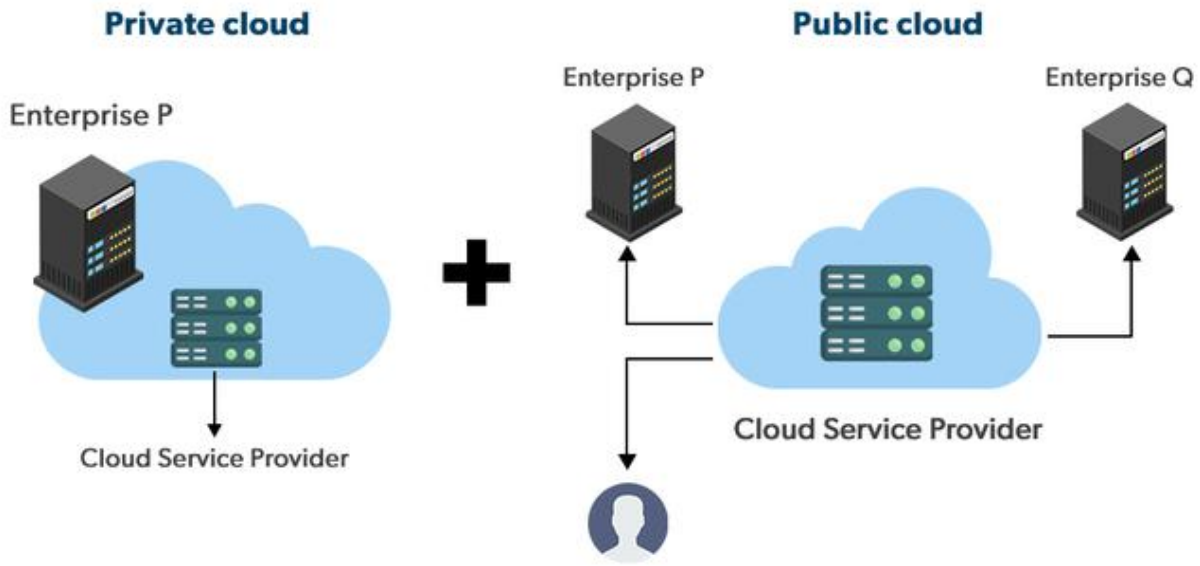
Private Cloud

- The private cloud deployment model is the exact opposite of the public cloud deployment model.
- It's a one-on-one environment for a single user (customer).
- There is no need to share your hardware with anyone else.



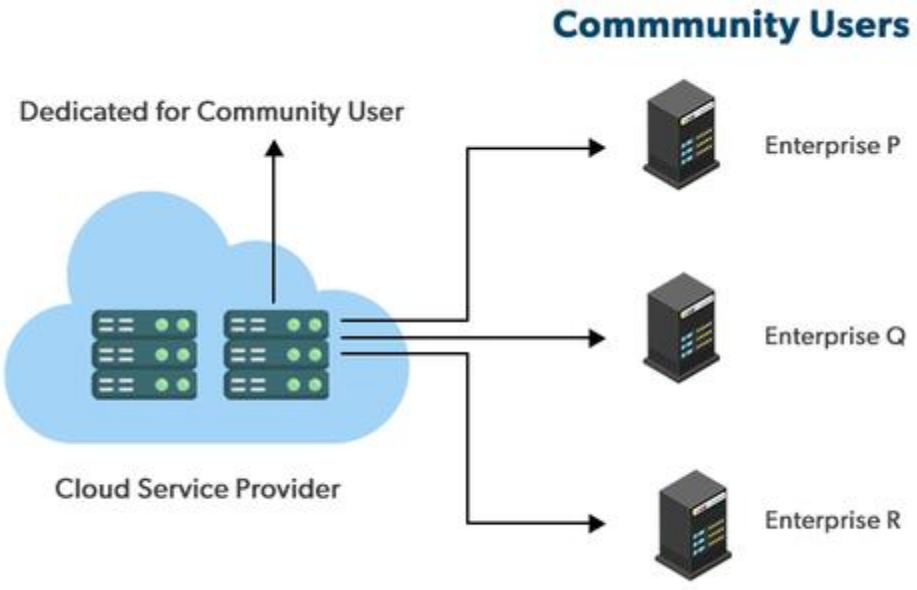
Hybrid Cloud

- A hybrid cloud is a combination of both public and private cloud environments.
- It allows organizations to take advantage of the benefits of both types of clouds.



Community Cloud

- It allows systems and services to be accessible by a group of organizations.
- It is a distributed system that is created by integrating the services of different clouds to address the specific needs of a community, industry, or business.
- The infrastructure of the community could be shared between the organization which has shared concerns or tasks.
- It is generally managed by a third party or by the combination of one or more organizations in the community.



Difference between Public Cloud vs Private Cloud vs Hybrid Cloud

Factors	Public Cloud	Private Cloud	Hybrid Cloud
Resources	Resources are shared among multiple customers	Resources are shared with a single organization	It is a combination of public and private clouds. based on the requirement.
Tenancy	Data of multiple organizations is stored in the public cloud	Data of a single organization is stored in a clouds the public cloud	Data is stored in the public cloud, and provide security in the public cloud.
Pay Model	Pay what you used	Have a variety of pricing models	It can include a mix of public cloud pay-as-you-go pricing, and private cloud fixed pricing. It has other pricing models such as consumption-based, subscription-based, etc.
Operated by	Third-party service provider	Specific organization	Can be a combination of both
Scalability and Flexibility	It has more scalability and flexibility,	It has predictability and consistency	It has scalability and flexibility by allowing organizations to use a combination of public and private cloud services.
Expensive	less expensive	More expensive	Can be more expensive, but it can also be less expensive , depending on the specific needs and requirements of the organization.
Availability	The general public (over the internet)	Restricted to a specific organization	Can be a combination of both.

stateful architecture?

Stateful architecture or application describes a structure that allows users to store, record, and return to already established information and processes over the internet. It entails transactions that are performed using past transactions as a reference point. In stateful applications, the current transaction can be affected by the previous ones.

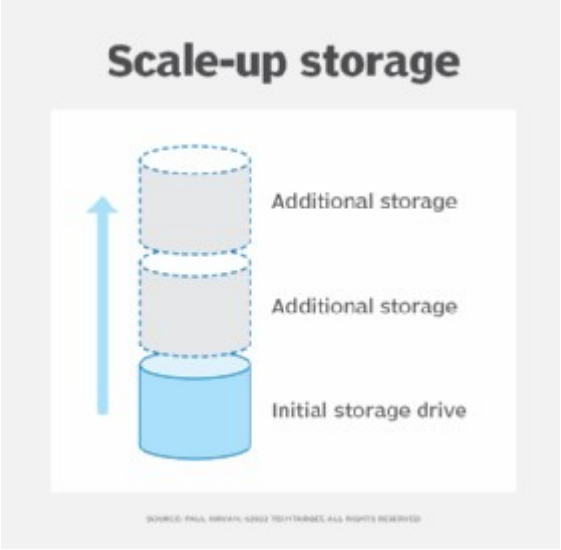
stateless architecture?

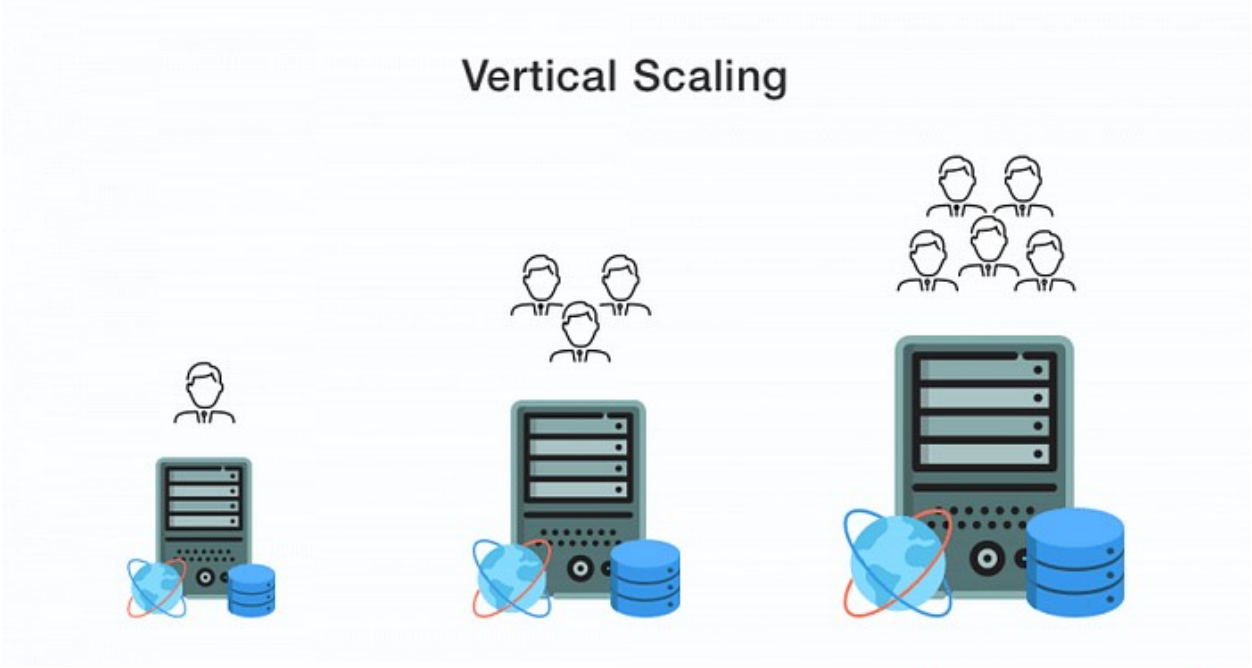
A stateless architecture or application is a type of Internet protocol where the state of the previous transactions is neither stored nor referenced in subsequent transactions. Each request sent between the sender and receiver can be interpreted and does not need earlier requests for its execution.

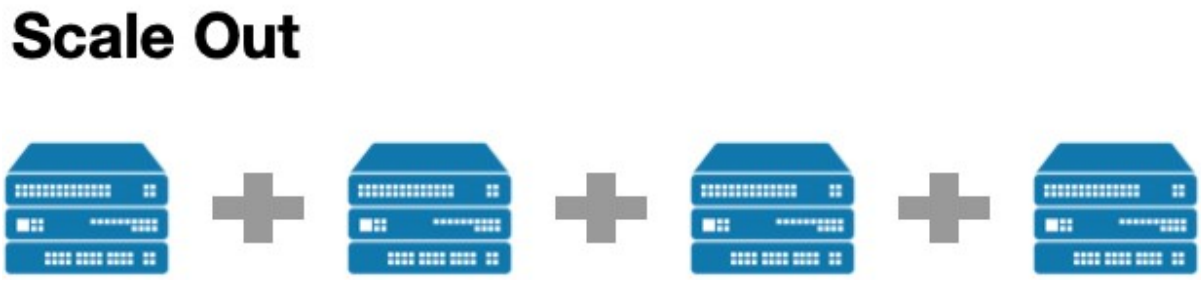
Cloud scalability in cloud computing refers to increasing or decreasing IT resources as needed to meet changing demand.

Scaling Up Vs. Scaling Out

Scaling Up/Down approach is also known as *vertical scaling*.This approach focuses on enhancing the power and capacity of individual resources to handle greater workloads.

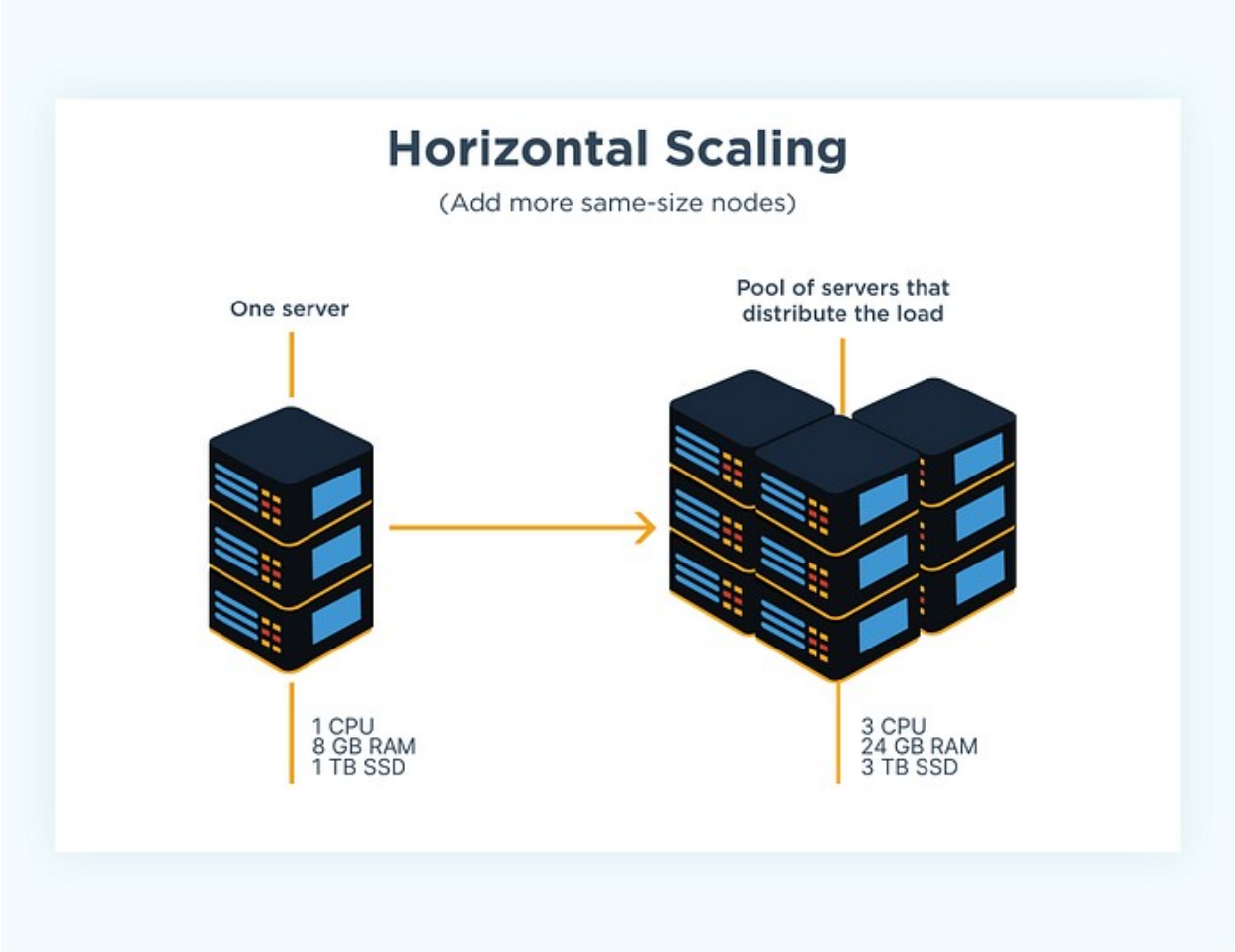


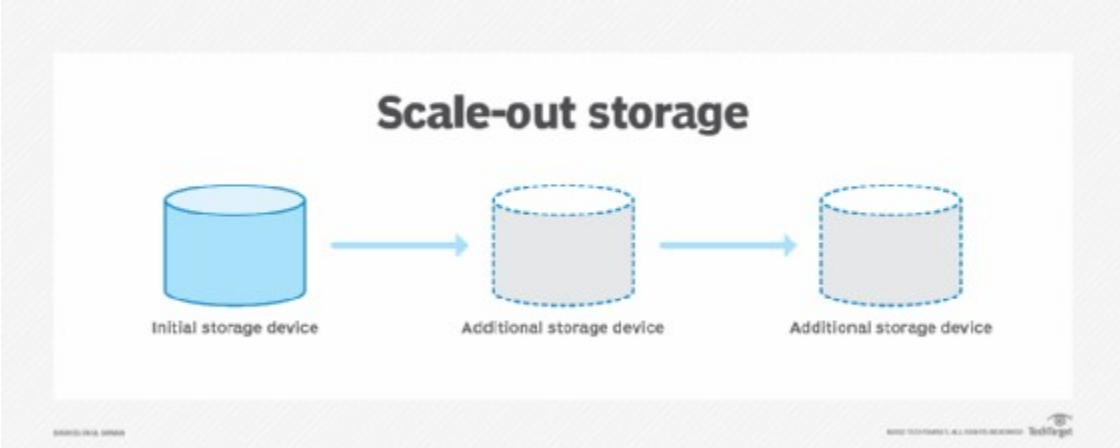




Scaling Out

Horizontal scaling, also referred to as scaling out, involves adding more machines or servers to distribute the workload across multiple resources. This approach focuses on increasing capacity by adding more identical systems.





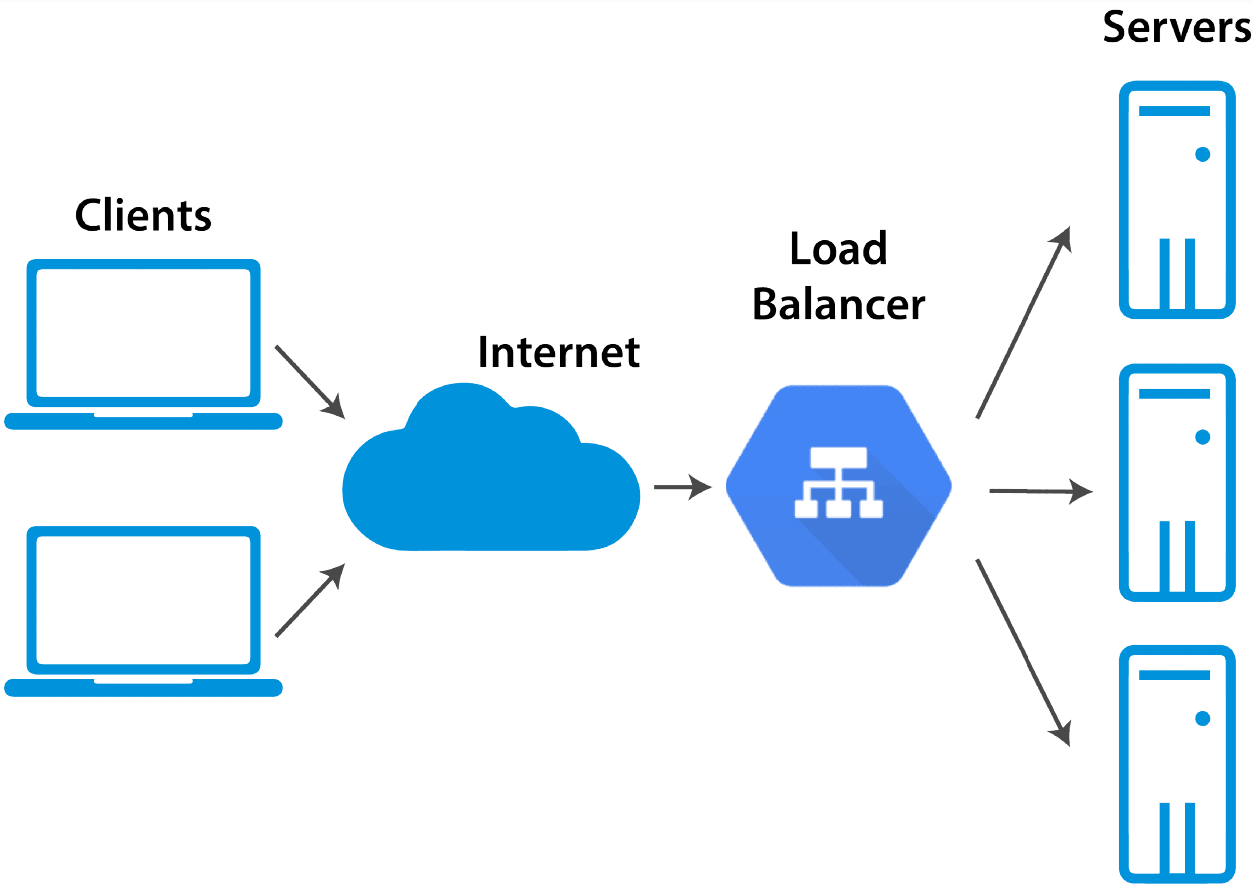
Load Balancing

Cloud load balancing is the practice of evenly distributing traffic, workloads, and client requests across multiple servers running in a cloud environment. This practice ensuring that each cloud resource has a load it can reasonably manage, preventing machines or servers in a cloud environment from being either overloaded or underutilized.

As a result, it improves an application's availability, scalability, security, and performance.

Application availability

Server failure or maintenance can increase application downtime, making your application unavailable to visitors. Load balancers increase the fault tolerance of your systems by automatically detecting server problems and redirecting client traffic to available servers.

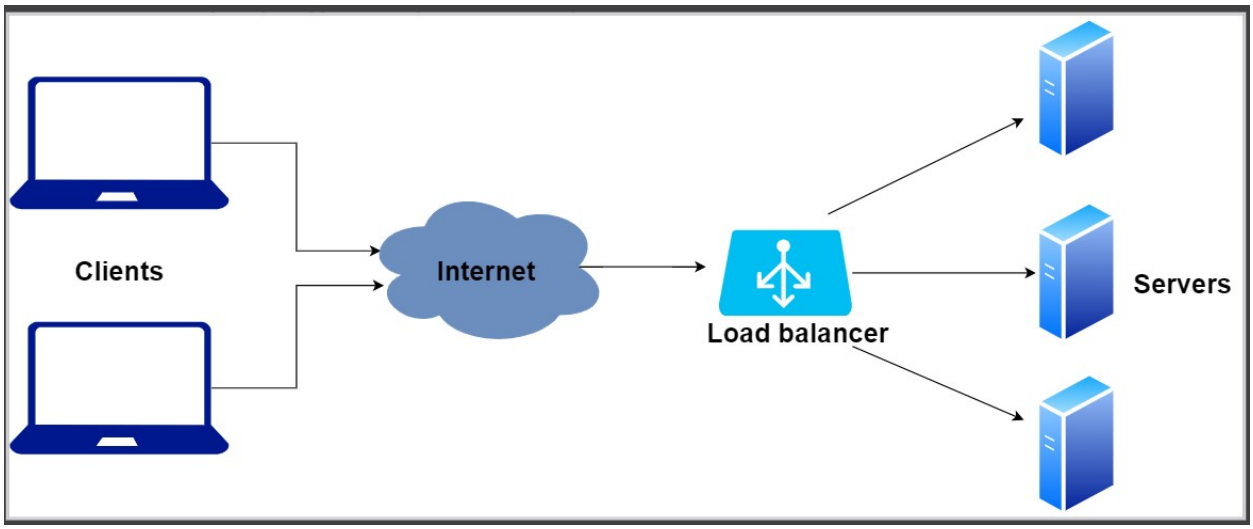


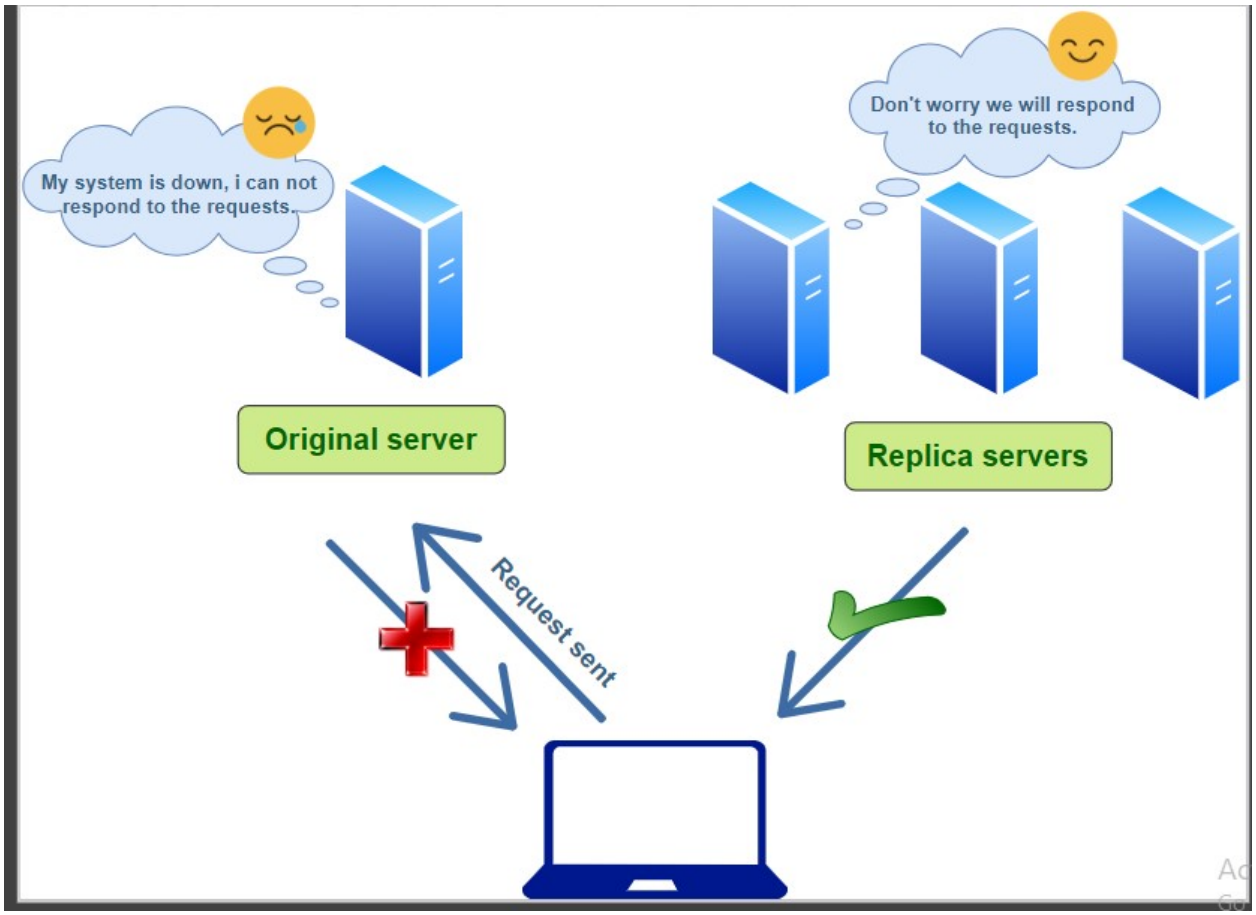
Fault Tolerance

Cloud fault tolerance simply means your infrastructure is capable of supporting uninterrupted functionality of your applications despite failures of components.

In a cloud computing setting that may be due to autoscaling across geographic zones or in the same data centers.

Fault tolerance refers to the ability of the system to keep functioning in even if a software or hardware failure occurs or going through a down state. It is a critical aspect to improve the realibility of a system and keep it useful for the user under all circumstances.


Clients
Internet
Load balancer
Servers

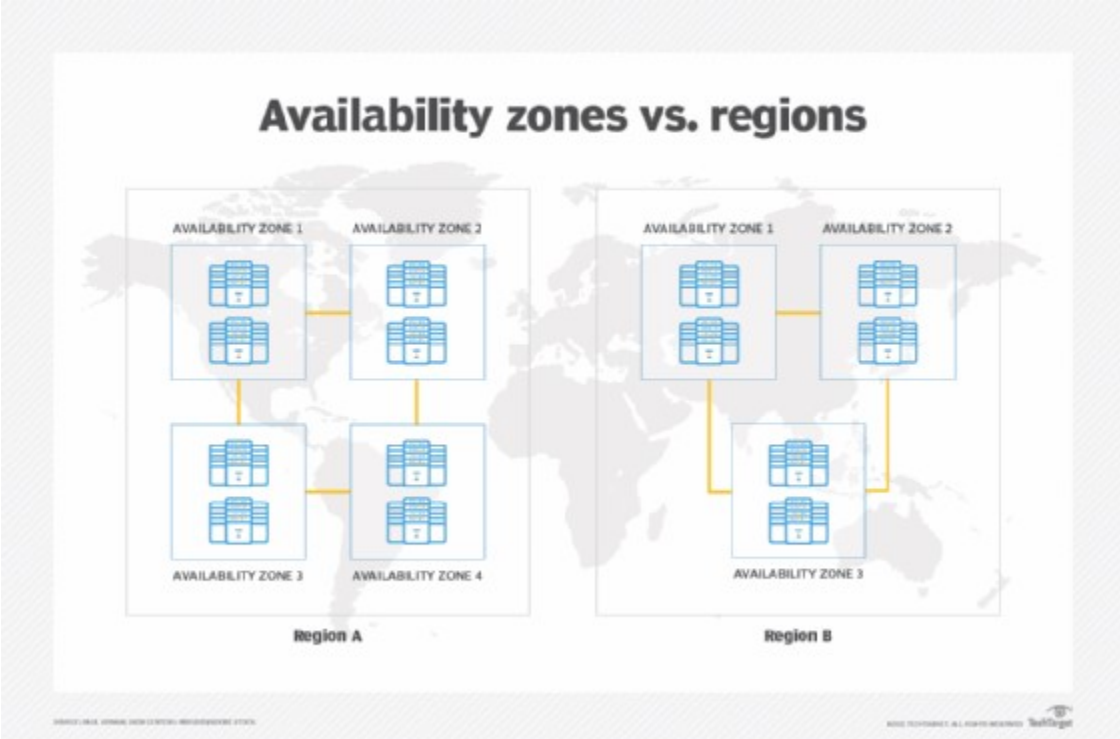

My system is down, i can not respond to the requests.
Don't worry we will respond to the requests.
Original server
Replica servers
Request sent

AWS

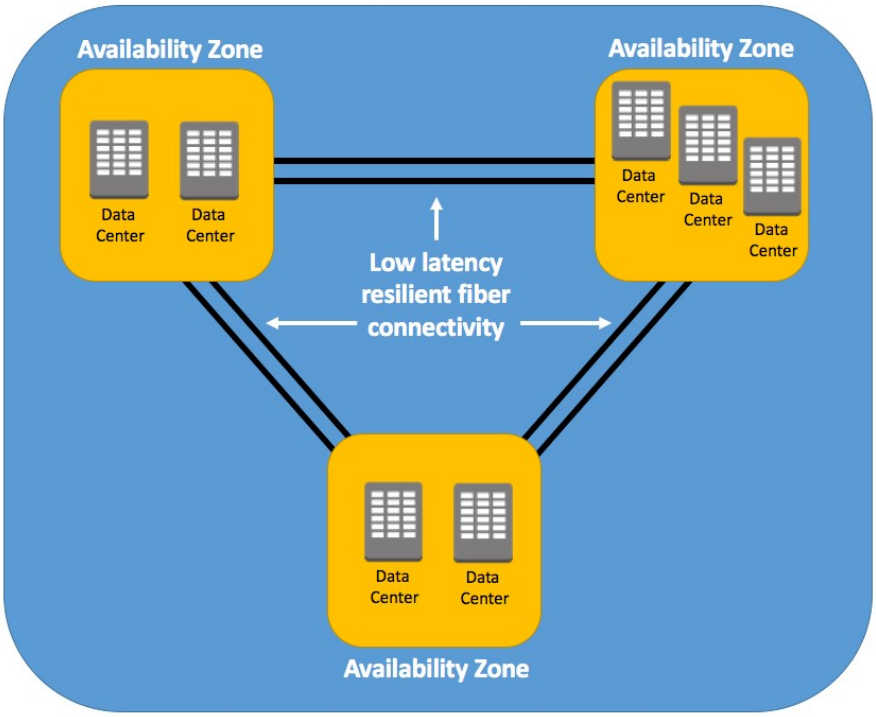
- Amazon Web Services (AWS) is a secure cloud platform that offers a broad set of global cloud-based Services.
- Because these products are delivered over the internet, you have on-demand access to the compute, storage, network, database, and other IT resources that you might need for your projects—and the tools to manage them.
- You can immediately provision and launch AWS resources. The resources are ready for you to use in minutes.
- AWS offers flexibility. Your AWS environment can be reconfigured and updated on demand, scaled up or down automatically to meet usage patterns and optimize spending, or shut down temporarily or permanently.
- The billing for AWS services becomes an operational expense instead of a capital expense.
- AWS services are designed to work together to support virtually any type of application or workload.

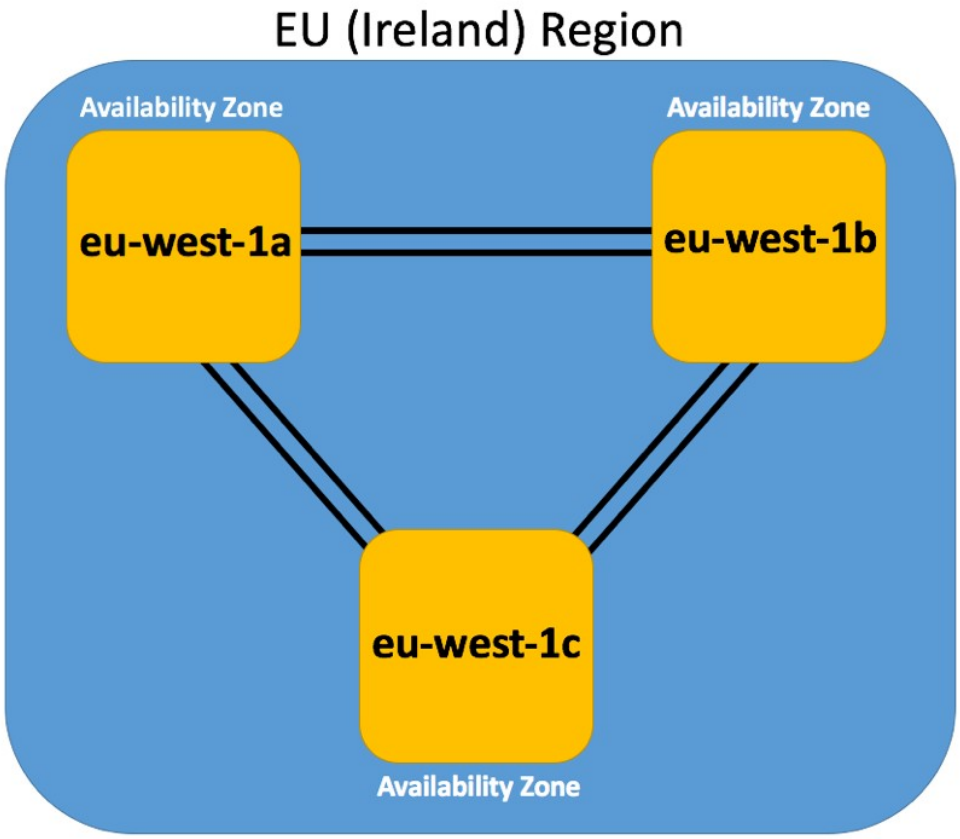
AWS Regions and AZ

- Amazon cloud computing resources are hosted in multiple locations world-wide. These locations are composed of AWS Regions, Availability Zones.
- Each *AWS Region* is a separate geographic area.
- Each AWS Region has multiple, isolated locations known as *Availability Zones*.
- AZs are essentially the physical data centers of AWS.
- Each AZ may contain one or more data centers.
- This is where the actual compute, storage, network, and database resources are located that we use over internet.

- 
Availability zones vs. regions
AVAILABILITY ZONE 1
AVAILABILITY ZONE 2
AVAILABILITY ZONE 3
AVAILABILITY ZONE 4
Region A
AVAILABILITY ZONE 1
AVAILABILITY ZONE 2
AVAILABILITY ZONE 3
Region B

REGION

- 
Availability Zone
Availability Zone
Data Center
Data Center
Data Center
Data Center
Data Center
Low latency resilient fiber connectivity
Data Center
Data Center
Availability Zone



Creating an AWS Account

1. Open the Amazon Web Services home page .
2. Choose **Create an AWS account**.
3. Enter your account information, and then choose **Verify email address**. This will send a verification code to your specified email address.
4. Enter your verification code, and then choose **Verify**.
5. Enter a strong password for your root user, confirm it, and then choose **Continue**.
6. Choose **Business** or **Personal**. Personal accounts and business accounts have the same features and functions.
7. Enter your company or personal information.
8. Read and accept the AWS Customer Agreement. Be sure that you read and understand the terms of the AWS Customer Agreement.
9. Choose **Continue**. At this point, you'll receive an email message to confirm that your AWS account is ready to use. You can sign in to your new account by using the email address and password you provided during sign up.

10. Enter the information about your payment method, and then choose **Verify and Continue**. If you want to use a different billing address for your AWS billing information, choose **Use a new address**. You can't proceed with the sign-up process until you add a valid payment method.
11. Enter your country or region code from the list, and then enter a phone number where you can be reached in the next few minutes.
12. Enter the code displayed in the CAPTCHA, and then submit.
13. When the automated system contacts you, enter the PIN you receive and then submit.
14. Select one of the available AWS Support plans.
15. Choose **complete sign up**. A confirmation page appears that indicates that your account is being activated.
16. Check your email and spam folder for an email message that confirms your account was activated. Activation usually takes a few minutes but can sometimes take up to 24 hours.After you receive the activation message, you have full access to all AWS services.

AWS Shared Responsibility Model

The AWS shared responsibility model is a concept of dividing responsibilities between AWS and a Customer.

The **Customer** is you.

AWS's responsibilities are the security of the cloud.

Customer responsibilities are security in the cloud.

Responsibility of AWS

AWS's responsibility is the security of the cloud.

AWS manages all infrastructure layers.

AWS responsibility “Security of the Cloud” - AWS is responsible for protecting the infrastructure that runs all of the services offered in the AWS Cloud. This infrastructure is composed of the hardware, software, networking, and facilities that run AWS Cloud services.

Some of the infrastructure layers are:

- Data centers
- Hardware and software
- Virtualization
- Networking

Responsibility of a Customer

Customers' responsibility is the security of everything they make in AWS Cloud.

Customers (you) have complete control over your content.

Customer manages AWS services, software, and access to the data.

Customer responsibility “Security in the Cloud” – Customer responsibility will be determined by the AWS Cloud services that a customer selects.