Day1 learning of AWS

Friday, January 24, 2025 10:03 AM

What is the cloud? | Cloud definition

The cloud is made up of servers in data centers all over the world. Moving to the cloud can save companies money and add convenience for users.

What is compute?

In cloud computing, the term "compute" describes concepts and objects related to software computation. It is a generic term used to reference processing power, memory, networking, storage, and other resources required for the computational success of any program.

What is cloud computing?

Cloud computing is the on-demand delivery of IT resources over the Internet with pay-as-you-go pricing.

Traditional Computing vs. Cloud Computing

Understanding Traditional Infrastructure Challenges

Before the rise of cloud computing, businesses relied heavily on traditional infrastructure to host their applications. Setting up a simple website involved significant costs and complexities. Let's break down the process:

- **Server Purchase**:
- You had to purchase a physical server costing around ₹50,000 or more.
- The server needed a secure and cold environment, often requiring renting a dedicated space.
- **Maintenance Costs**:
- Continuous power supply and cooling systems were essential.
- You needed to hire IT professionals to manage the server, handle hardware failures (like hard disk or RAM issues), and perform regular maintenance.
- **Over or Underutilization**:
- If you anticipated 10,000 users but only received 10, the investment in a high-capacity server was wasted.
- Conversely, if the number of users exceeded expectations, the server would crash due to overload, resulting in downtime and a poor user experience.
- **Inflexibility**:
- Scaling up or down based on demand was cumbersome, leading to either over-provisioning or under-provisioning.

These limitations made it difficult for businesses to focus on their core objectives and hindered innovation.

Enter Cloud Computing

Cloud computing revolutionized how infrastructure is managed, offering a solution to all the challenges of traditional setups. Here's how:

On-Demand Resource Provisioning

With cloud computing, you pay only for the resources you use. Need a server for a day? No problem. The cloud provider charges you on an hourly or even per-second basis. This eliminates upfront costs and over-provisioning.

Scalability

Cloud services automatically scale based on your needs. For instance:

- If your website receives 10 users today and 10 million tomorrow, the cloud can seamlessly add or reduce servers to handle the traffic.
- During low-traffic periods, resources scale down, reducing costs.

No Maintenance Hassles

Cloud providers manage the physical infrastructure, so you don't need to worry about hardware failures, power supply, or cooling systems. This allows you to focus entirely on developing and optimizing your application.

Global Accessibility

All your cloud resources are accessible over the internet, ensuring high availability and reliability for users worldwide.

Everyday Cloud Examples

Even if you're not actively setting up servers, you're likely using cloud computing in your daily life:

- **Google Drive**: Stores your files on the cloud.
- **Gmail**: Provides email services without needing to set up a personal mail server.
- **Netflix**: Streams movies and shows via cloud-hosted platforms.

Why Businesses Prefer Cloud Computing

Cost Efficiency

You only pay for what you use. If you stop using a server, you can shut it down, and the billing stops immediately.

Scalability and Flexibility

Handle sudden spikes in traffic without any manual intervention or additional setup. Resources are allocated dynamically to match your application's needs.

Focus on Core Goals

Since infrastructure management is handled by the cloud provider, businesses can focus on building robust applications and delivering excellent user experiences.

Real-Life Example: World Cup Website Traffic

Consider a sports website like ESPN during the World Cup. When no matches are happening, the site might only require minimal resources. However, during a major match, millions of users flock to the website. Cloud computing ensures:

- Servers automatically scale up to handle the increased traffic.
- Once the traffic subsides, resources scale down to minimize costs.

Introduction to Cloud Services

Cloud services have revolutionized how businesses manage and deploy their applications by offering flexibility, scalability, and cost efficiency. Let's explore the three primary types of cloud services: **Infrastructure as a Service (IaaS)**, **Platform as a Service (PaaS)**, and **Software as a Service (SaaS)**.

What Are Cloud Services?

Cloud services provide resources, tools, and services to businesses over the internet, eliminating the need for complex on-premises infrastructure. Understanding these services can help identify the right fit for your needs.

Types of Cloud Services

1. Infrastructure as a Service (IaaS)

IaaS provides businesses with essential IT infrastructure, including virtual machines, storage, and networking. Here's how it works:

- **Key Features**:
- The cloud provider offers hardware and virtualization, along with an operating system.
- Users decide the type of operating system (e.g., Ubuntu, Fedora) and are responsible for managing applications on top of it.
- **Example Use Case**:
- You've purchased a virtual server and chosen your preferred OS.
- All software, configurations, and maintenance are your responsibility.
- **Examples of IaaS**:
- Amazon EC2
- Microsoft Azure Virtual Machines
- Google Compute Engine

2. Platform as a Service (PaaS)

PaaS provides a platform that allows developers to build, deploy, and manage applications without worrying about the underlying infrastructure.

- **Key Features**:
- The provider manages the OS, middleware, and runtime.
- You focus solely on application development and data management.
- **Example Use Case**:
- A database administrator leverages PaaS to create and configure a database platform, such as MySQL or PostgreSQL, with built-in replication and security features.
- Developers can directly interact with the database without managing hardware or OS configurations.
- **Examples of PaaS**:
- Amazon RDS
- Google App Engine
- Heroku

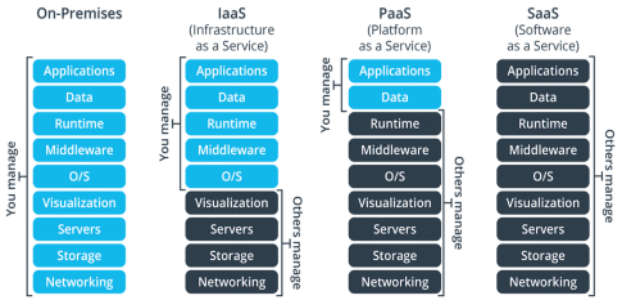
3. Software as a Service (SaaS)

SaaS delivers fully functional software applications over the internet. These applications are managed entirely by the service provider.

- **Key Features**:
- No setup or maintenance is required on the user's end.
- The service provider handles infrastructure, application maintenance, and updates.
- **Example Use Case**:
- Email services like Gmail or Microsoft Outlook allow businesses to create custom email addresses (e.g., **admin@learning-ocean.com**) without setting up mail servers.
- **Examples of SaaS**:
- Gmail
- Dropbox
- Salesforce

Visual Representation of Virtualization Layers

To help visualize the differences between these services, the following diagram illustrates their respective virtualization layers:



- **IaaS**: Hardware and virtualization layers are provided, with users managing the OS and applications.
- **PaaS**: Includes everything in IaaS, along with middleware and runtime. Users focus only on data and applications.
- **SaaS**: All layers, from hardware to applications, are fully managed by the provider.

How to Identify the Right Service for Your Needs

When choosing a cloud service, consider the following:

1. What is this service managing on my behalf?
2. What responsibilities do I have?
3. Does it align with my project requirements?

By answering these questions, you can determine whether a service falls under IaaS, PaaS, or SaaS.

Types of Cloud Models

Cloud computing has transformed how businesses manage resources by offering flexible and scalable solutions. There are three primary types of cloud models:

- **Public Cloud**
- **Private Cloud**
- **Hybrid Cloud**

Each of these models serves different purposes and caters to specific business needs. Let’s dive deeper into each one.

1. Public Cloud

The **Public Cloud** is a cloud environment accessible over the internet, allowing anyone to create and manage resources. It is hosted and maintained by third-party providers and is available to the general public.

Key Features:

- Resources are shared among multiple users.
- Accessible via the internet.
- Cost-effective and highly scalable.

Example Use Case:

A startup can use the public cloud to host its website or application without worrying about infrastructure management. Resources can be scaled up or down based on demand.

Examples of Public Cloud Providers:

- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)
- Microsoft Azure

2. Private Cloud

The **Private Cloud** is a cloud environment dedicated to a single organization. It provides enhanced security and control, making it ideal for businesses with strict compliance or regulatory requirements.

Key Features:

- Exclusive access to resources.
- Hosted on-premises or in a dedicated data center.
- Customizable to meet specific organizational needs.

Example Use Case:

A financial institution that handles sensitive customer data may opt for a private cloud to ensure data remains within its control and meets compliance standards.

Implementation:

Private clouds can be set up using tools like OpenStack. For instance, an organization might create virtual machines (VMs) within its private network, accessible only to employees through a secure connection.

3. Hybrid Cloud

The **Hybrid Cloud** combines the features of both public and private clouds. It allows organizations to leverage the benefits of both models by seamlessly integrating their private cloud with a public cloud environment.

Key Features:

- Flexibility to run sensitive workloads on a private cloud while utilizing the public cloud for less critical tasks.
- Enhanced scalability and cost-efficiency.

Example Use Case:

A company may store customer data on a private cloud for security reasons while hosting its web application on a public cloud to handle

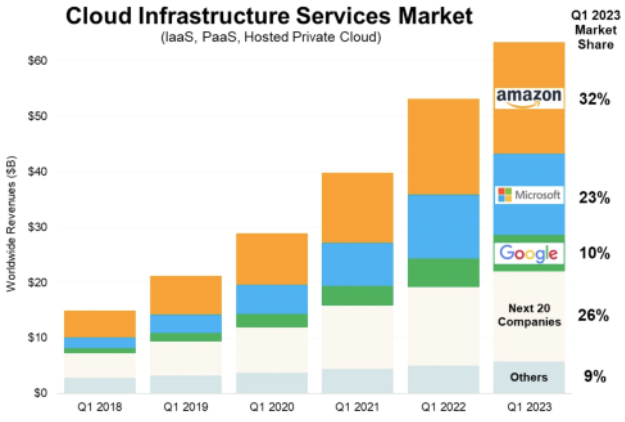
unpredictable traffic spikes.
Implementation:
Hybrid clouds can be implemented using services like AWS Outposts or Azure Arc, which enable smooth integration between public and private environments.

Comparing Cloud Models

Feature	Public Cloud	Private Cloud	Hybrid Cloud
Accessibility	Open to everyone	Restricted to one organization	Combines both
Cost	Pay-as-you-go	Higher initial setup cost	Flexible
Scalability	Highly scalable	Limited by hardware	Scalable
Security	Standard security	Enhanced security	Mixed security

How AWS Charge?

There are three fundamental drivers of cost with AWS: compute, storage, and outbound data transfer. These characteristics vary somewhat, depending on the AWS product and pricing model you choose.



AWS Regions

A **Region** is a distinct geographical area where AWS operates its data centers. Each Region is independent to ensure better control and performance for users in that area.
Key Points:

- AWS currently has **27 Regions** across the globe, with new ones being added over time.
- Each Region provides localized services to minimize latency and meet compliance requirements.

Example:

- **Mumbai Region**: Designed to cater to Indian businesses for faster service delivery.
- **Hyderabad Region** (Upcoming): Expanding AWS's footprint in India.

Why AWS Uses Global Regions

AWS distributes its infrastructure across various Regions for several reasons:

1. **Performance Optimization**:
   - Hosting data closer to users reduces latency.
   - Example: Indian customers accessing services in the Mumbai Region experience better speeds compared to using services hosted in the US.
2. **Compliance**:
   - Some industries or governments require data to stay within specific countries or Regions.
   - Example: Government projects often mandate local data storage.
3. **Disaster Recovery**:
   - Data redundancy across Regions minimizes risks from natural disasters or technical failures.
   - Example: Data stored in Mumbai can be backed up in another Region, such as Singapore.
4. **Flexibility**:
   - Businesses can choose Regions based on their specific needs, such as cost-effectiveness or proximity

Understanding AWS Availability Zones

In this post, we will explore **AWS Availability Zones (AZs)**, how they function within Regions, and why they are essential for building resilient and highly available applications.

What Are Availability Zones?

An **Availability Zone** is an isolated location within an AWS Region. Each AZ consists of one or more data centers that operate independently but are interconnected to other AZs within the same Region.

Key Characteristics:

- **Isolation**: Each AZ is designed to operate independently to minimize the risk of failure spreading across zones.
- **Redundancy**: Multiple AZs ensure that if one zone experiences an issue, others can continue to function.
- **Interconnection**: AZs within a Region are connected via high-speed, low-latency networks to facilitate efficient data replication and communication.

Why Does AWS Use Availability Zones?

1. **Disaster Recovery**:
   - AZs are designed to handle localized failures such as power outages or natural disasters.
   - Example: If a power failure affects one AZ in Mumbai, resources in other AZs in the same Region will remain operational.
2. **High Availability**:
   - By distributing applications across multiple AZs, businesses can ensure minimal downtime.
   - Example: A web application hosted across three AZs will continue to serve users even if one AZ is temporarily unavailable.
3. **Fault Tolerance**:
   - Each AZ is connected to independent power sources and networks, reducing the likelihood of a single point of failure.
4. **Performance Optimization**:
   - High-speed connections between AZs enable efficient data synchronization and load balancing.

How Are Availability Zones Structured?

Within an AWS Region, AZs are geographically separated but close enough to maintain low-latency connectivity. Here's how they are structured:

- **Distance Between AZs**: Typically 60-100 kilometers (37-62 miles) apart to prevent simultaneous impact from disasters.
- **Independent Power Supply**: Each AZ has its own power source and backup generators.
- **Independent Networking**: AZs are connected to separate network grids, ensuring continued connectivity even if one network fails.

Example: Mumbai Region

The **Mumbai Region** has multiple Availability Zones, strategically placed to ensure:

- High availability for Indian users.
- Compliance with data residency requirements.
- Reliable disaster recovery options.

Best Practices:

- Deploy applications across multiple AZs within the Region to ensure redundancy and fault tolerance.
- Use services like Elastic Load Balancing to distribute traffic across AZs.

Benefits of Using Multiple Availability Zones

1. **Reduced Latency**:
   - Low-latency connections between AZs ensure real-time data replication and high performance.
2. **Improved Reliability**:
   - Applications distributed across AZs can handle failures without impacting end users.
3. **Enhanced Security**:
   - Data transfers between AZs are encrypted, ensuring secure communication.
4. **Compliance**:
   - AZs help meet regulatory requirements by offering localized data storage options.

AWS Local Zones

AWS Local Zones are extensions of AWS Regions designed to bring select AWS services closer to specific locations. These zones enable businesses to reduce latency and deliver seamless user experiences, especially for workloads requiring real-time data processing.

What Are AWS Local Zones?

Local Zones are designed to provide **low-latency access** to specific applications by deploying AWS infrastructure closer to users. They are particularly useful for industries like gaming, media, and financial services that demand rapid response times.

Key Features:

- **Single-Digit Millisecond Latency**: Local Zones reduce latency significantly compared to standard AWS Regions.
- **Targeted Use**: Ideal for workloads with location-specific demands.
- **Limited Services**: Only select AWS services are available in Local Zones.

Why Use AWS Local Zones?

1. **Low Latency**:
   - For applications requiring real-time responsiveness, like gaming or live video streaming.
   - Example: Hosting a gaming server in a Local Zone ensures single-digit millisecond latency for users in nearby areas.
2. **Proximity to Users**:
   - Brings applications closer to end-users, improving the user experience.
3. **Compliance Requirements**:
   - Allows data to remain within a specific city or region to meet local regulations.

Services Available in Local Zones

Not all AWS services are available in Local Zones. You can click here and check the available services in aws local zone.

Example:

In a Local Zone, you can create EC2 instances and utilize EBS volumes for storage. More services may become available in the future as AWS continues to expand its Local Zones.

Current and Upcoming Local Zones

AWS is actively increasing the number of Local Zones worldwide. For instance:

- Local Zones are available in several major cities globally.
- More zones are expected to launch, catering to additional cities and regions.

Use Cases for AWS Local Zones

1. **Gaming**:
   - Hosting game servers close to players ensures low latency, enhancing the gaming experience.
2. **Financial Services**:
   - Enables real-time transaction processing for users in specific locations.
3. **Media and Entertainment**:
   - Low-latency video rendering and live streaming.