307401 Big Data

Introduction to Cloud Computing and Amazon AWS

2024-2

What is Cloud Computing?

Definition:

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

Key Features:

- Access to computing services like servers, storage, databases, networking, and software.
- Available anytime, from anywhere, via the internet.
- Scalable and flexible according to business needs.

Traditional IT vs. Cloud Computing

Feature	Traditional IT	Cloud Computing
Infrastructure	Physical hardware	Virtual infrastructure
Cost	Capital expenditure (CapEx)	Operational expenditure (OpEx)
Setup time	Weeks to months	Minutes
Scalability	Limited by physical capacity	Instantly scalable
Maintenance	Handled by in-house teams	Handled by cloud provider

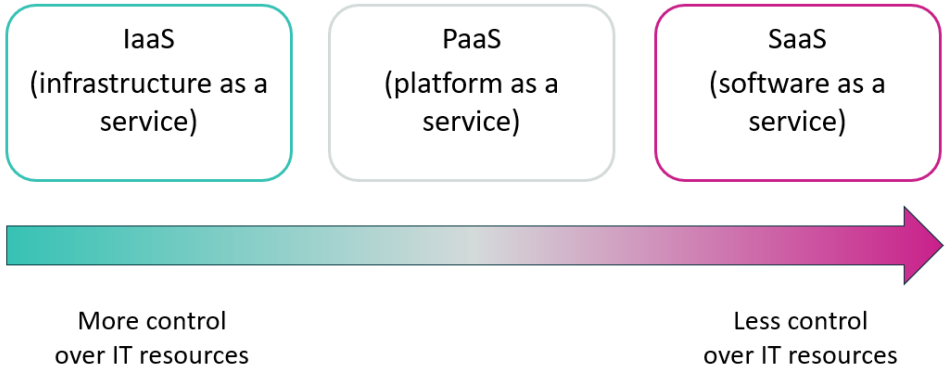
Infrastructure as Software

- In cloud computing, infrastructure (servers, storage, networks) is treated as software.
- Users can provision and manage resources using code or simple interfaces.
- This eliminates the need to purchase and manage physical hardware.

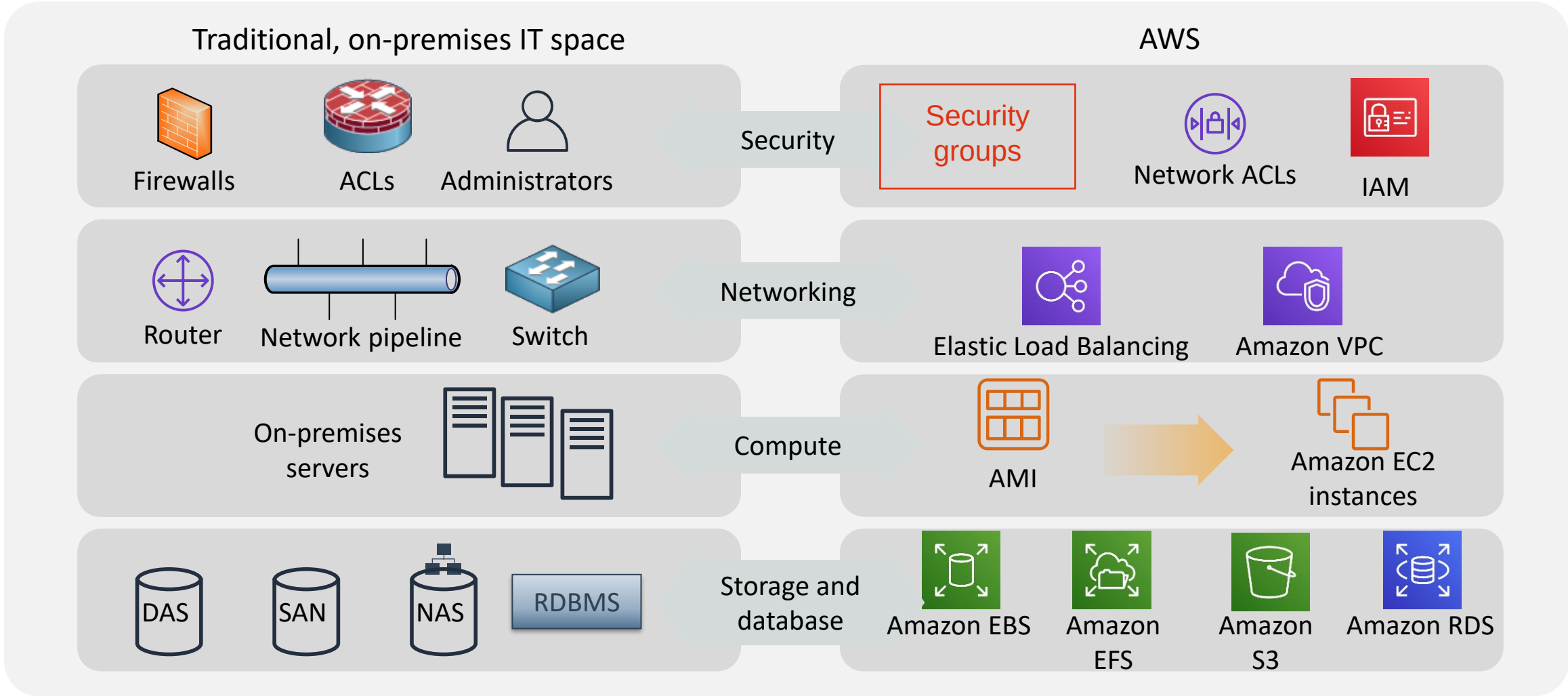
Cloud Computing Models

Three Primary Service Models:

1. **Infrastructure as a Service (IaaS)**
    1. Provides virtualized computing resources over the internet.
    2. Examples: Amazon EC2, Google Compute Engine
2. **Platform as a Service (PaaS)**
    1. Provides hardware and software tools over the internet (usually for app development).
    2. Examples: AWS Elastic Beanstalk, Google App Engine
3. **Software as a Service (SaaS)**
    1. Delivers software applications over the internet.
    2. Examples: Google Workspace, Microsoft 365



Similarities between AWS and traditional IT



Cloud Deployment Models

1.Public Cloud

- Services delivered over the public internet, available to anyone.
- Example: AWS, Azure

2.Private Cloud

- Services maintained on a private network for a single organization.

3.Hybrid Cloud

- Combines public and private clouds, allowing data and apps to move between them.

Benefits of Cloud Computing

- **Cost Efficiency:** Pay only for what you use.
- **Scalability:** Instantly adjust resources to meet demand.
- **Agility:** Deploy applications and services rapidly.
- **Global Reach:** Deliver services worldwide with minimal latency.
- **No Maintenance:** Cloud provider handles updates and maintenance.

Examples of Cloud Services

Category	Example Service	Description
Compute	Amazon EC2	Virtual servers in the cloud
Storage	Amazon S3	Scalable object storage
Database	Amazon RDS	Managed relational database service
Networking	Amazon VPC	Isolated virtual networks

Advantages of Cloud Computing

Overview

Cloud computing transforms the way organizations operate by providing a more efficient, scalable, and cost-effective approach to IT.

In this section, we will cover:

- Cost transformation
- Scalability and flexibility
- Speed and agility
- Operational efficiency
- Global accessibility

From CapEx to OpEx

Traditional IT	Cloud Computing
High upfront costs (CapEx)	Pay-as-you-go pricing (OpEx)
Costly overprovisioning	Scale only when needed
Long procurement cycles	Instant resource provisioning

Key Benefit: You pay only for what you use, reducing waste and improving budgeting.

Economies of Scale

Cloud providers serve millions of customers.

This **massive scale** allows them to:

- Purchase infrastructure at lower costs
- Optimize usage of physical resources
- Pass cost savings to customers
- **Analogy:** Like buying in bulk at wholesale prices.

Elasticity and Scalability

Cloud platforms automatically scale resources based on demand.

Two forms of scaling:

- **Vertical Scaling:** Increase power (CPU/RAM) of existing resources
- **Horizontal Scaling:** Add more resources to handle load (e.g., more servers)
- **Result:** No more guessing capacity in advance or overprovisioning.

Speed and Agility

- Traditional IT takes **weeks or months** to deploy infrastructure.
- Cloud computing enables deployment in **minutes**.
- **Use Case:** Developers can test, deploy, and iterate quickly—reducing time-to-market.

Operational Efficiency

No need to manage physical data centers.

Cloud providers handle:

- Power, cooling, physical security
- Server maintenance and upgrades
- Hardware lifecycle management
- **Your team can focus on innovation instead of infrastructure.**

Global Reach

Cloud services are available across **geographic regions and availability zones**.

What this means for you:

- Serve customers around the world with minimal latency
- Comply with data residency and legal requirements
- Deploy across multiple regions for resilience

BUSINESS INTELLIGENCE AND DATA ANALYTICS
– WE PREDICT THE FUTURE –

Institute of Analytics
Recognized Programme

University of Petra

Introduction to Amazon Web Services (AWS)

What is AWS?

Amazon Web Services (AWS) is a **comprehensive cloud computing platform** provided by Amazon.

Key Features:

- Delivers **on-demand** access to compute, storage, databases, networking, analytics, and more.
- Offers a **pay-as-you-go** pricing model.
- Operates as a **secure, global infrastructure** serving millions of customers.

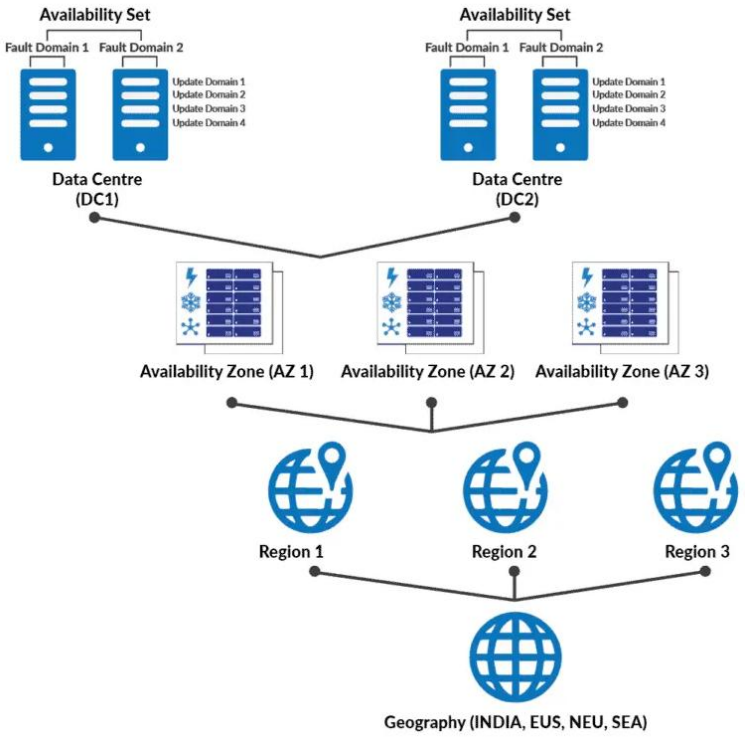
Characteristics of AWS

- **Secure:** Data encryption, identity management, and compliance.
- **Flexible:** Choose the right tools, platforms, and configurations.
- **Scalable:** Instantly adapt to demand changes.
- **Reliable:** Built-in redundancy and failover mechanisms.
- **Global:** Data centers in multiple regions and availability zones.

AWS Global Footprint

AWS operates in:

- **Regions:** Geographic locations (e.g., US East, EU Central)
- **Availability Zones:** Isolated data centers within regions
- **Edge Locations:** For content delivery and low-latency access



Purpose for this Infrastructure:

1. High availability
2. Fault tolerance
3. Low latency

https://aws.amazon.com/about-aws/global-infrastructure/#AWS_Global_Infrastructure_Map

https://aws.amazon.com/about-aws/global-infrastructure/regions_az/

AWS Regions

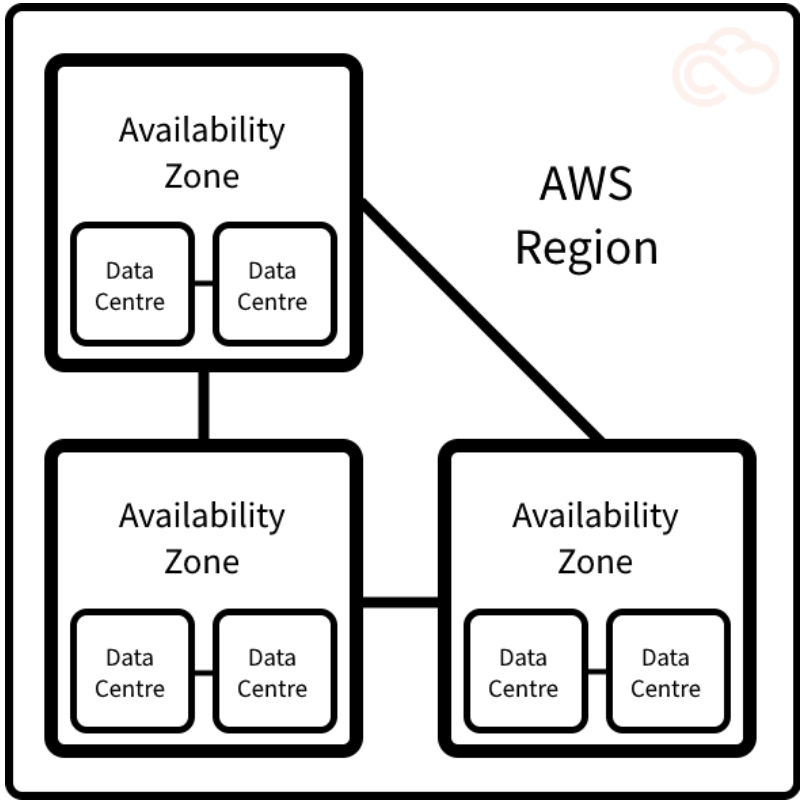
- Each **Region** is a separate geographic area.
- Contains **2 or more Availability Zones**.
- Customers choose Regions based on:
  - **Latency**
  - **Legal compliance**
  - **Service availability**
  - **Cost differences**

Example:

- Europe (Ireland) → eu-west-1
- US East (N. Virginia) → us-east-1

Availability Zones (AZs)

- Each Region includes multiple **Availability Zones**.
- AZs are **physically isolated** but **interconnected** using high-speed private links.
- Designed for **fault isolation** and **resiliency**.
- **Best Practice:**
  Deploy across multiple AZs to build **highly available** and **disaster-resilient** systems.

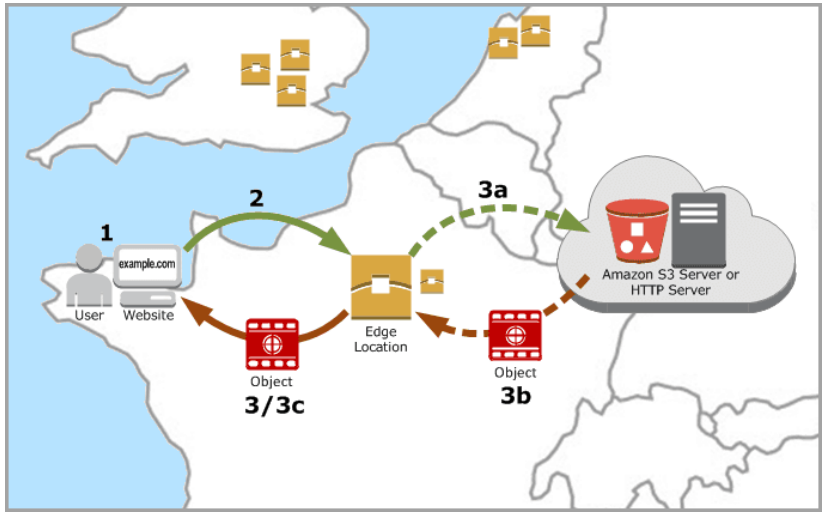


AWS Data Centers

- Secure, fault-tolerant physical facilities.
- Each AZ consists of **one or more data centers**.
- Features include:
  - Redundant power and networking
  - Access controls and physical security
  - Environmental protections (cooling, fire suppression)

Points of Presence (PoPs)

- Includes **Edge Locations** and **Regional Edge Caches**.
- Used primarily for **content delivery** via Amazon CloudFront (CDN).
- Enables **fast delivery** of static and dynamic content with **low latency**.



AWS Infrastructure Capabilities

- **Elasticity:** Automatically scales based on demand.
- **High Availability:** Ensures uptime across services.
- **Fault Tolerance:** Isolates failures at the AZ level.
- **Redundancy:** Multiple copies of data/services across AZs.

Example

```
Region: us-west-2 (Oregon)
  ├── AZ-a (data center 1, 2)
  ├── AZ-b (data center 3)
  └── AZ-c (data center 4, 5)
Edge Location: Seattle
Edge Location: San Francisco
```

This structure enables global applications with local performance.

AWS Service Categories

Why Organize Services into Categories?

- AWS offers over 200 services.
- To make it easier to find, understand, and select the right services, AWS organizes them into **functional categories**.
- These categories reflect how businesses use cloud services to meet specific goals.

Compute services –
- Amazon EC2
- AWS Lambda
- AWS Elastic Beanstalk
- Amazon EC2 Auto Scaling
- Amazon ECS
- Amazon EKS
- Amazon ECR
- AWS Fargate

Storage services –
- Amazon S3
- Amazon S3 Glacier
- Amazon EFS
- Amazon EBS

Management and Governance services –
- AWS Trusted Advisor
- AWS CloudWatch
- AWS CloudTrail
- AWS Well-Architected Tool
- AWS Auto Scaling
- AWS Command Line Interface
- AWS Config
- AWS Management Console
- AWS Organizations

Database services –
- Amazon RDS
- Amazon DynamoDB
- Amazon Redshift
- Amazon Aurora

Security, Identity, and Compliance services –
- AWS IAM
- Amazon Cognito
- AWS Shield
- AWS Artifact
- AWS KMS

Networking and Content Delivery services –
- Amazon VPC
- Amazon Route 53
- Amazon CloudFront
- Elastic Load Balancing

AWS Cost Management services –
- AWS Cost & Usage Report
- AWS Budgets
- AWS Cost Explorer

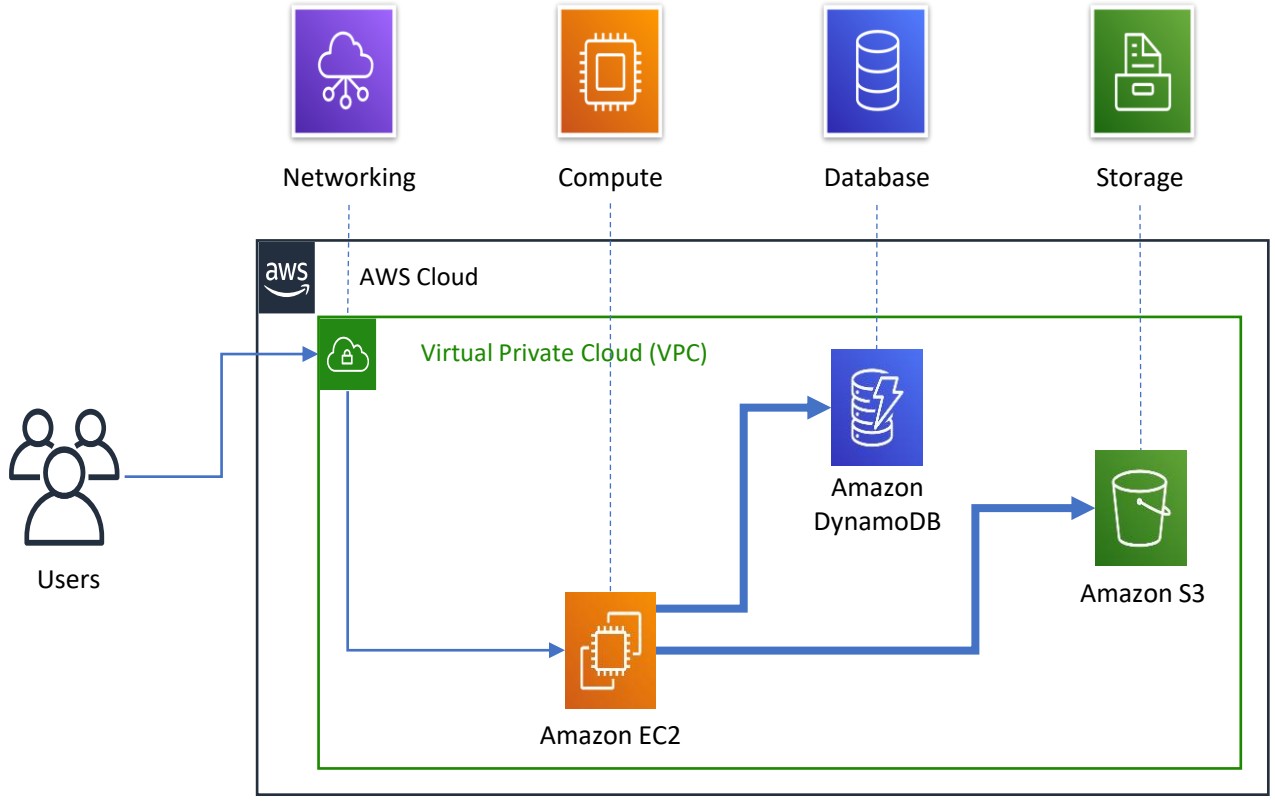
Core AWS Service Categories

Category	Focus Area
Compute	Virtual servers and serverless apps e.g. EC2, Lambda, Elastic Beanstalk
Storage	Data storage (object, block, file) e.g. S3, EBS, Glacier
Database	Managed SQL and NoSQL databases e.g. RDS, DynamoDB, Redshift
Networking & Content Delivery	Secure connectivity and low-latency delivery e.g. VPC, Route 53, CloudFront
Security, Identity & Compliance	Access control, encryption, and governance e.g. IAM, KMS, Shield
Management & Monitoring	Visibility, logging, and optimization e.g. CloudWatch, CloudTrail
Developer Tools	CI/CD, version control, code deployment
Analytics	Big data processing and analysis e.g. Athena, Glue, QuickSight
Machine Learning	AI/ML model training and deployment e.g. SageMaker, Rekognition

Example Use Case

A simple web application hosted on AWS:

- **Networking:** Amazon VPC
- **Compute:** Amazon EC2
- **Storage:** Amazon S3
- **Database:** Amazon DynamoDB
- **Security:** IAM roles and policies
- This modular setup shows how services integrate seamlessly.



Compute Services

Purpose: Deliver computing capacity in various forms

Service	Description
Amazon EC2	Virtual servers in the cloud
AWS Lambda	Serverless functions triggered by events
AWS Elastic Beanstalk	Managed app deployment environment
Amazon ECS/EKS	Containers and Kubernetes management
AWS Fargate	Serverless containers

Storage Services

Purpose: Store, archive, and retrieve data efficiently

Service	Description
Amazon S3	Scalable object storage
Amazon EBS	Block storage for EC2 instances
Amazon EFS	Elastic file storage for shared access
Amazon S3 Glacier	Archival storage with low retrieval cost

Database Services

Purpose: Run databases without managing infrastructure

Service	Description
Amazon RDS	Managed relational databases (e.g., MySQL, PostgreSQL)
Amazon Aurora	High-performance relational DB compatible with MySQL/PostgreSQL
Amazon DynamoDB	Managed NoSQL key-value/document store
Amazon Redshift	Data warehouse for large-scale analytics

Networking & Content Delivery

Purpose: Connect resources securely and deliver content globally

Service	Description
Amazon VPC	Isolated virtual networks
Route 53	Scalable DNS and domain name routing
CloudFront	Global content delivery network (CDN)
Elastic Load Balancing	Distributes traffic to maintain performance

Security, Identity, and Compliance

Purpose: Protect data, manage user access, and meet compliance standards

Service	Description
AWS IAM	Identity and Access Management
AWS KMS	Key Management Service for encryption
AWS Shield	DDoS protection
AWS Artifact	Compliance and audit documentation

Monitoring & Management

Purpose: Improve visibility, control, and efficiency

Service	Description
Amazon CloudWatch	Monitors metrics, logs, alarms
AWS CloudTrail	Tracks user activity and API usage
AWS Config	Audits configuration changes
AWS Trusted Advisor	Optimization and best practices review

Developer Tools & Others

Developer Tools:

- AWS CodeCommit, CodeBuild, CodeDeploy, CodePipeline

Other Categories (Examples):

- **Machine Learning:** Amazon SageMaker, Rekognition
- **Analytics:** Athena, Glue, QuickSight
- **IoT, Blockchain, Robotics, AR/VR**
  Each with targeted services

BUSINESS INTELLIGENCE AND DATA ANALYTICS
– WE PREDICT THE FUTURE –

Institute of Analytics
Recognized Programme

University of Petra

AWS Cloud Security and Identity Access Management (IAM)

Shared Responsibility Model

AWS uses a shared responsibility model to divide security duties between AWS and the customer.

Responsibility Area	AWS (Provider)	Customer (User)
Physical security	✓	
Infrastructure (hardware, VMs)	✓	
Network configuration		✓
OS, apps, and data		✓
Access control and permissions		✓

Summary:
AWS secures the **cloud itself**, customers secure what they do **in the cloud**.

AWS Responsibilities

AWS is responsible for:

- Securing data centers, servers, networking, and storage infrastructure.
- Providing encrypted services and physical redundancy.
- Maintaining hardware, hypervisors, and virtualization layers.

Customer Responsibilities

You (the customer) are responsible for:

- Securing your applications and data.
- Managing access through IAM (users, roles, groups).
- Configuring firewalls (security groups) and encryption.
- Regularly updating your OS, patches, and credentials.

Security at Different Service Levels

Cloud Model	Responsibility Level
IaaS	Customer manages OS, apps, firewalls
PaaS	Customer manages data and app logic
SaaS	Customer manages only the data and access

Example:

- EC2 (IaaS) requires OS patching.
- Lambda (PaaS) abstracts OS, customer only writes code.

What is IAM (Identity and Access Management)?

IAM is the AWS service that allows you to:

- **Authenticate:** Who is trying to access the system.
- **Authorize:** What they are allowed to do.

Key Functions:

- Define **users**, **groups**, **roles**, and **policies**
- Control access to AWS **resources** (e.g., S3, EC2)

IAM Key Concepts

Concept	Description
User	Represents a person or service needing access
Group	A collection of users with shared permissions
Role	Temporary access granted to trusted entities
Policy	A JSON document defining allowed or denied actions on resources

IAM Policies

- Written in **JSON** format
- Two types:
  - **Identity-based:** Attach to users, groups, or roles
  - **Resource-based:** Attach directly to resources (e.g., S3 bucket)
- **Example Allow Policy:**

```
{
  "Effect": "Allow",
  "Action": ["s3:*"],
  "Resource": "arn:aws:s3:::example-bucket/*"
}
```

Principle of Least Privilege

- Always follow this best practice:
- "Grant only the permissions required to perform a task, and no more.“
- This minimizes security risks by limiting exposure.

IAM Roles – Use Cases

IAM Roles are used to:

- Grant applications (e.g., EC2 instances) permission to access AWS services.
- Delegate access to users in other AWS accounts.
- Enable temporary access with conditions (e.g., session duration).

AWS Compute Services

What is Compute in Cloud Computing?

"Compute" refers to the processing power required to run applications, systems, and workloads.

AWS provides flexible options:

1. Virtual machines (EC2)
2. Serverless functions (Lambda)
3. Containers (ECS, EKS)
4. Platform-as-a-Service (Elastic Beanstalk)

Amazon EC2 – Virtual Servers

Amazon Elastic Compute Cloud (EC2) allows you to rent virtual servers (called *instances*) in the cloud.

Key Features:

- Choose OS, CPU, memory, storage, and network capacity.
- Launch and terminate instances on demand.
- Full control of the server (root/admin access).

Use Cases:

- Web servers
- Application servers
- Development and testing environments

EC2 Highlights

- **AMI (Amazon Machine Image):** Pre-configured templates used to launch instances.
- **Instance Types:** Choose based on CPU, memory, storage (e.g., t2.micro, m5.large).
- **Security Groups:** Virtual firewalls for EC2.
- **Elastic IPs:** Static public IP addresses.
- **Auto Scaling:** Automatically increase/decrease the number of instances.

Main Categories of EC2 Instance Families

Prefix	Category	Purpose / Optimized For	Example Instance Type
t	**General Purpose (Burstable)**	Cost-effective with CPU bursts — great for low traffic workloads	t3.micro
m	**General Purpose**	Balanced CPU, memory, networking — good for most applications	m5.large
c	**Compute Optimized**	High-performance processors — great for compute-heavy workloads	c6g.xlarge
r	**Memory Optimized**	More RAM — for memory-intensive applications like in-memory DBs	r5.2xlarge
x	**Memory Optimized (High Mem)**	Ultra-high memory (SAP HANA, in-memory DBs, analytics)	x1e.8xlarge
z	**High CPU / Memory Ratio**	High GHz performance — electronic design, gaming	z1d.large
d	**Storage Optimized**	Local NVMe or HDD storage — great for big data and analytics	d3.4xlarge
i	**Storage Optimized (IOPS)**	High IOPS SSD — databases, low latency storage needs	i3en.large
h	**Storage Optimized (HDD)**	High-throughput HDD storage — for large sequential datasets	h1.8xlarge
p	**Accelerated Computing (GPU)**	Powerful GPUs — ML training, HPC, graphics rendering	p4d.24xlarge
g	**GPU-based Inference**	GPU optimized for ML inference, gaming, graphics	g5.xlarge
f	**FPGA-based Computing**	Field Programmable Gate Arrays — custom hardware acceleration	f1.4xlarge
inf	**Inferentia-based**	AWS Inferentia chips — high-performance ML inference	inf1.xlarge
trn	**Trainium-based**	AWS Trainium chips — high-performance ML training	trn1.32xlarge
u	**Bare Metal / Specialized**	Ultra high memory bare metal instances	u-6tb1.metal

Instance Type Naming Format: <family><generation>.<size>

Example:

m5.large → General purpose (m), 5th gen, large size

c6g.xlarge → Compute optimized (c), 6th gen, ARM-based (g), xlarge size

AWS Lambda – Serverless Computing

Lambda lets you run code *without provisioning or managing servers*.

Key Concepts:

- Upload your function code
- Define event triggers (e.g., file uploaded to S3, API call)
- Code runs automatically and scales based on demand

Benefits:

- Pay only when code runs
- Built-in fault tolerance
- Supports multiple programming languages (Python, Node.js, Java, etc.)

Common Lambda Use Cases

- Real-time file processing (e.g., generate image thumbnails)
- Scheduled tasks (e.g., backup at midnight)
- Backend services for APIs
- Real-time stream processing (e.g., Kinesis, DynamoDB Streams)

Container Services Overview

Containers are lightweight, portable packages of software that include everything needed to run an application.

Benefits over Virtual Machines:

- Faster to start/stop
- Consistent environments (code runs the same on any machine)
- Lower overhead

AWS container services:

1. **Amazon ECS:** AWS-native container orchestration
2. **Amazon EKS:** Managed Kubernetes
3. **AWS Fargate:** Serverless containers

AWS Fargate – Serverless Containers

AWS Fargate is a fully managed **serverless compute engine** for containers offered by **Amazon Web Services (AWS)**.

It lets you run containers without having to manage the underlying servers or clusters.

With **Fargate**, you:

- Do not manage EC2 instances
- Only define and run containers
- Pay per vCPU and memory used
- **Ideal for:** Developers who want to focus only on the application, not the infrastructure.

AWS Elastic Beanstalk – Platform-as-a-Service (PaaS)

Elastic Beanstalk allows you to deploy applications without managing the underlying infrastructure.

Supported languages/platforms:

- Java, Python, .NET, Node.js, PHP, Ruby, Go, Docker

What it does:

- Handles provisioning (EC2), load balancing, scaling, monitoring
- You simply upload your code
- **Use Case:** Web developers who want speed and ease.

Choosing the Right Compute Service

Scenario	Recommended Service
Full control of server environment	EC2
Event-driven, serverless execution	Lambda
Microservices with Docker/Kubernetes	ECS / EKS
Managed deployment of web apps	Elastic Beanstalk

AWS Storage Services

Why Storage Matters in the Cloud

Cloud storage allows you to store, access, and protect data without managing physical hardware.

AWS provides multiple storage types to support:

- Backup and archiving
- File systems
- Application data
- Media and content delivery

Main Types of AWS Storage

Storage Type	Description	Use Case
Object Storage	Stores data as objects with metadata	Static content, backups, big data
Block Storage	Data stored in blocks, like hard drives	Operating systems, databases
File Storage	Shared file system for applications	File shares, content management
Archival Storage	Long-term, low-cost cold storage	Compliance, infrequently accessed data

Amazon S3 – Object Storage

Amazon Simple Storage Service (S3) is AWS's core object storage service.

Features:

- Scalable to store petabytes
- High durability (99.999999999% or "11 nines")
- Lifecycle management and versioning
- Ideal for storing any type of unstructured data

Use Cases:

- Media files, backups, logs, big data lakes

Amazon EBS – Block Storage

Amazon Elastic Block Store (EBS) provides persistent block storage for EC2.

Features:

- Used as a virtual hard drive for EC2 instances
- High performance and low latency
- Data is retained when instances stop

Use Cases:

- Operating systems
- Relational and NoSQL databases
- Transactional apps

Amazon EFS – File Storage

Amazon Elastic File System (EFS) provides a scalable and elastic network file system.

Features:

- Fully managed
- Shared across multiple EC2 instances
- Grows/shrinks automatically

Use Cases:

- Web servers needing shared access to files
- Enterprise content management systems

Amazon S3 Glacier – Archival Storage

S3 Glacier is designed for long-term, infrequently accessed data.

Features:

- Extremely low-cost
- Retrieval options from minutes to hours
- Secure and durable

Use Cases:

- Regulatory data retention
- Historical archives
- Compliance logs

Storage Classes in Amazon S3

Class	Use Case	Retrieval Time
S3 Standard	Frequent access	Immediate
S3 Intelligent-Tiering	Unknown or changing patterns	Immediate
S3 Standard-IA	Infrequent access	Immediate
S3 Glacier	Archival	Minutes to hours
S3 Glacier Deep Archive	Lowest-cost, rarely accessed data	Up to 12 hours

Durability and Availability

- **Durability**: Probability that your data will not be lost (e.g., S3: 99.999999999% durability)
- **Availability**: System uptime and accessibility (e.g., S3 Standard: 99.99% availability)

AWS Database Services

Why Use Managed Databases?

Traditional database management involves:

- Provisioning hardware
- Installing software
- Applying patches
- Performing backups and scaling manually

AWS Database Services offer:

- Fully managed infrastructure
- Scalability and high availability
- Backup, patching, and monitoring automation
- Choice between **SQL** and **NoSQL** options

Relational vs. Non-Relational Databases

Type	Characteristics	Examples
Relational (SQL)	Structured data, tables, schemas, ACID compliance	Amazon RDS, Amazon Aurora
Non-Relational (NoSQL)	Key-value, document, flexible schema	Amazon DynamoDB

Amazon RDS – Relational Database Service

Amazon RDS is a managed service that runs traditional relational databases (like MySQL, PostgreSQL, SQL Server, etc.) for you — but on AWS infrastructure. You don't need to install or maintain the database software yourself — AWS does that part.

Key Features:

- Supports MySQL, PostgreSQL, SQL Server, Oracle, and MariaDB
- Automated backups and patching
- Multi-AZ deployment for high availability
- Read replicas for performance
- **Use Cases:**
  Enterprise applications, ERP systems, websites with structured data

Amazon Aurora

Aurora is a **MySQL and PostgreSQL-compatible** relational database built for the cloud.

Key Advantages:

- Up to 5x faster than standard MySQL
- Up to 3x faster than PostgreSQL
- Storage auto-scales up to 128 TB
- High durability (replicated across 6 copies)
- **Ideal for:** Mission-critical, high-performance applications

Amazon DynamoDB – NoSQL Database

Amazon DynamoDB is a fully managed **key-value and document** database.

Features:

- Single-digit millisecond latency at any scale
- Automatically scales throughput and storage
- Built-in security, backup, and caching (via DAX)
- **Use Cases:**
  Gaming leaderboards, IoT, real-time bidding engines, session management

Amazon Redshift – Data Warehousing

Amazon Redshift is a fully managed **data warehouse** optimized for analytics.

Features:

- Petabyte-scale analytics
- SQL-based querying
- Integrates with S3 and BI tools (e.g., Tableau, QuickSight)
- Supports complex joins, aggregations, and OLAP operations
- **Use Cases:**
  Business intelligence, reporting, data lake analytics

AWS Database Selection Guide

Requirement	Recommended Service
Managed relational DB	Amazon RDS
High-performance SQL	Amazon Aurora
NoSQL key-value store	Amazon DynamoDB
Data warehousing & analytics	Amazon Redshift

AWS Networking and Content Delivery

Why Networking in the Cloud Matters

In AWS, networking controls:

- How services communicate with each other
- How traffic flows between your applications and users
- Security boundaries and availability zones

AWS provides tools to:

- Build private networks
- Route global traffic
- Deliver content efficiently
- Secure communication

Amazon VPC – Virtual Private Cloud

Amazon VPC allows you to provision a **logically isolated network** within AWS.

Key Features:

- Define subnets, IP ranges, route tables
- Configure security groups and network ACLs
- Enable private and public subnets
- Connect VPC to on-premises via VPN or Direct Connect
- **Use Case:**
  Run secure applications in a private environment

Elastic Load Balancing (ELB)

ELB automatically distributes incoming traffic across multiple targets.

Types of Load Balancers:

- Application Load Balancer (HTTP/HTTPS, Layer 7)
- Network Load Balancer (TCP/UDP, Layer 4)
- Gateway Load Balancer (3rd party virtual appliances)

Benefits:

- Fault tolerance
- Auto scaling integration
- Health checks for high availability

Amazon Route 53 – DNS Service

Amazon Route 53 is a **highly available and scalable DNS** (Domain Name System) service.

Functions:

- Register and manage domain names
- Route users to AWS or external services
- Health checks and latency-based routing
- Geo DNS support
- **Use Case:**
  Direct users to the fastest and healthiest endpoints globally

Amazon CloudFront – Content Delivery Network (CDN)

CloudFront speeds up delivery of content (e.g., HTML, CSS, images, videos, APIs).

Key Features:

- Uses edge locations worldwide
- Reduces latency and increases speed
- Integrates with S3, EC2, and Lambda@Edge
- Supports HTTPS and access control
- **Use Case:**
  Distribute web content globally with fast, secure access

AWS Direct Connect

AWS Direct Connect establishes a **dedicated private network connection** from your data center to AWS.

Benefits:

- Lower latency
- Consistent network performance
- Avoids public internet
- **Use Case:**
  Enterprises with sensitive workloads or large data transfers

AWS VPN – Virtual Private Network

AWS VPN allows you to securely connect on-premises networks to AWS over the public internet.

Types:

- Site-to-Site VPN
- Client VPN
- **Use Case:**
  Secure communication between offices, remote workers, and AWS cloud resources

BUSINESS INTELLIGENCE AND DATA ANALYTICS
– WE PREDICT THE FUTURE –

Institute of Analytics
Recognized Programme

University of Petra

AWS Monitoring, Cost Management, and Support Services

Why Monitoring and Cost Management Matter

Operating in the cloud gives flexibility—but also requires visibility.

Key goals:

- Monitor system health and performance
- Control and optimize cloud spending
- Track changes and maintain compliance
- Get support for operational issues

Amazon CloudWatch – Monitoring and Observability

Amazon CloudWatch provides data and actionable insights for:

- Metrics
- Logs
- Alarms
- Dashboards

Use Cases:

- Monitor EC2 CPU usage
- Set alarms on Lambda errors
- Visualize service performance
- Trigger auto-scaling

CloudWatch Features

Feature	Functionality
Metrics	Track performance data (CPU, memory, etc.)
Logs	Collect logs from AWS and applications
Alarms	Alert you when thresholds are breached
Dashboards	Visualize key indicators in real time

Integration: Works with EC2, Lambda, RDS, ECS, and custom applications

AWS CloudTrail – Governance and Auditing

AWS CloudTrail records all API calls made in your AWS account.

Tracks:

- Who did what, and when
- Service-level activity logs
- Security, troubleshooting, and audit trails

Use Cases:

- Investigate unusual activity
- Audit resource usage
- Prove compliance

AWS Config – Resource Tracking

AWS Config helps track changes to your AWS resources over time.

Benefits:

- View current and historical configurations
- Detect drift from desired states
- Get alerts for non-compliance

Example:

- Notify when an S3 bucket becomes publicly accessible

AWS Trusted Advisor – Best Practices Insights

Trusted Advisor provides real-time recommendations for:

- Cost optimization
- Performance improvement
- Security enhancements
- Service limits
- **Available to:**
  All accounts (with limited checks); full checks with AWS Support plans

AWS Cost Explorer

Cost Explorer provides detailed **visual insights** into your AWS usage and spending.

Features:

- Analyze historical costs (up to 12 months)
- Forecast future spending
- Identify usage patterns and trends
- **Use Case:**
  Set budgets and detect unexpected cost spikes

AWS Budgets

AWS Budgets allows you to set custom thresholds for:

- Cost
- Usage
- Reservation coverage and utilization

Benefits:

- Receive alerts before you exceed a limit
- Monitor multiple services or accounts

AWS Support Plans

Plan	Support Level
Basic	Account access, billing, limited Trusted Advisor
Developer	Business hours tech support
Business	24/7 tech support, access to full Trusted Advisor
Enterprise	Dedicated Technical Account Manager (TAM), concierge support

AWS for Big Data and Machine Learning

Why Big Data and ML in the Cloud?

Modern applications generate large volumes of data requiring:

- Scalable processing
- Real-time analysis
- Cost-effective storage
- Tools for predictive modeling

AWS offers managed services that cover the full data lifecycle—from ingestion to storage, analysis, and prediction.

AWS Services for Big Data Workloads

Function	Service
Data Storage	Amazon S3, Amazon Glacier
Batch Processing	Amazon EMR (Hadoop/Spark)
Real-Time Streams	Amazon Kinesis, Amazon MSK (Kafka)
Data Integration	AWS Glue, AWS Data Pipeline
Analytics	Amazon Athena, Amazon Redshift, QuickSight

Amazon S3 as a Data Lake

Amazon S3 is often used as a **data lake** for storing structured and unstructured data.

Why?

- Unlimited scale
- High durability and availability
- Integration with analytics and ML tools

Common Use: Central repository for data before processing

Amazon EMR – Big Data Frameworks

Amazon EMR (Elastic MapReduce) runs big data frameworks like:

- Apache Hadoop
- Apache Spark
- HBase
- Hive

Use Cases:

- Distributed data processing
- ETL pipelines
- Log analysis
- Machine learning model training at scale

Amazon Kinesis – Real-Time Data Streams

Amazon Kinesis is designed for streaming data (e.g., website clicks, IoT events).

Components:

- Kinesis Data Streams
- Kinesis Data Firehose (load to S3/Redshift)
- Kinesis Data Analytics (SQL queries on streams)
- **Use Case:** Real-time dashboards, alerts, fraud detection

AWS Glue – Data Integration and ETL

AWS Glue is a fully managed **Extract, Transform, Load (ETL)** service.

Features:

- Automatically discovers schema
- Schedules and runs ETL jobs
- Integrates with S3, RDS, Redshift, etc.
- **Use Case:** Clean and prepare data before analytics or ML

Amazon Redshift – Scalable Data Warehouse

Amazon Redshift enables fast queries on petabyte-scale datasets.

Key Features:

- Columnar storage for performance
- Massively parallel processing (MPP)
- Integrates with BI tools and S3
- **Use Case:** Enterprise reporting, dashboards, analytical queries

Amazon Athena – Serverless Query Engine

Amazon Athena allows you to query S3 data using SQL.

Benefits:

- No infrastructure to manage
- Pay per query
- Ideal for ad-hoc queries on large data

Amazon QuickSight – Business Intelligence

Amazon QuickSight enables interactive dashboards and reports.

Features:

- Connects to S3, Redshift, RDS, and more
- Uses ML for forecasting and anomaly detection
- Embeddable into web apps

Amazon SageMaker – Machine Learning Platform

SageMaker is a fully managed service for the ML lifecycle:

Phase	Capability
Build	Jupyter notebooks, built-in algorithms
Train	Auto-scaling, distributed training
Deploy	One-click deployment, A/B testing
Monitor	Model drift and real-time performance

Use Case: Image recognition, NLP, fraud detection, recommendation engines

Other AI/ML Services

AWS also provides pre-built AI tools:

Service	Use Case
Amazon Rekognition	Image and video analysis
Amazon Polly	Text-to-speech synthesis
Amazon Transcribe	Speech-to-text conversion
Amazon Comprehend	Natural language processing (NLP)
Amazon Lex	Chatbots and conversational agents

Some of the Main Big Data Providers on the Cloud

Feature/Service Category	Amazon AWS	Google Cloud Platform	Microsoft Azure	Cloudera Data Platform (CDP)
Object Storage	Amazon S3	Google Cloud Storage	Azure Blob Storage	Cloudera Data Lake Storage (S3, ADLS, GCS)
Data Warehouse	Amazon Redshift	BigQuery	Azure Synapse Analytics (SQL Data Warehouse)	Cloudera Data Warehouse (Impala, Hive LLAP)
Batch Data Processing	Amazon EMR	Google Cloud Dataproc	Azure HDInsight	Cloudera Data Engineering (Apache Spark)
Stream Data Processing	Amazon Kinesis	Google Cloud Dataflow	Azure Stream Analytics	Cloudera Streams Messaging (Apache Kafka)
ETL/Data Integration	AWS Glue	Google Cloud Data Fusion	Azure Data Factory	Cloudera DataFlow (Apache NiFi)
Real-time Messaging	Amazon Kinesis Data Streams	Google Cloud Pub/Sub	Azure Event Hubs	Cloudera Streams Messaging (Apache Kafka)
Serverless Query Service	Amazon Athena	Google BigQuery	Azure Synapse Analytics (Serverless)	Cloudera Data Warehouse (Impala, Hive LLAP)
Data Lake Storage	Amazon S3 (with AWS Lake Formation)	Google Cloud Storage (with Dataproc)	Azure Data Lake Storage	Cloudera Data Lake Storage (S3, ADLS, GCS)
Big Data Analytics	AWS Glue/Athena	Google Cloud Dataflow/BigQuery	Azure Synapse Analytics/HDInsight	Cloudera Data Engineering, Data Science Workbench
Managed Apache Spark	Amazon EMR	Google Cloud Dataproc	Azure Databricks	Cloudera Data Engineering
NoSQL Database	Amazon DynamoDB	Google Cloud Bigtable	Azure Cosmos DB	Cloudera Operational Database (HBase, Kudu)
Machine Learning	Amazon SageMaker	Google AI Platform	Azure Machine Learning	Cloudera Machine Learning (CDSW)
Interactive Analysis	Amazon QuickSight	Looker	Power BI	Cloudera Data Warehouse, Hue
Data Exploration	Amazon Elasticsearch Service (OpenSearch)	Google Cloud Dataflow/Data Studio	Azure Data Explorer	Cloudera Data Science Workbench
Workflow Orchestration	AWS Step Functions	Google Cloud Composer	Azure Data Factory	Cloudera DataFlow (Apache NiFi)
Serverless Computing	AWS Lambda	Google Cloud Functions	Azure Functions	Cloudera Data Engineering (serverless Spark)
Managed Database	Amazon RDS	Google Cloud SQL	Azure SQL Database	Cloudera Operational Database (HBase, Kudu)
Pre-built AI Services	AWS AI Services (Rekognition, Polly, etc.)	Google Cloud AI (Vision, Speech, etc.)	Azure Cognitive Services	Cloudera Machine Learning (CDSW)