What is Cloud Computing

The NIST Definition of Cloud Computing

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 is composed of five essential characteristics, three service models, and four deployment models.

Essential Characteristics of Cloud

1. On-demand self-service

A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed **automatically without requiring human interaction** with each service provider.

2. Broad network access

Capabilities are **available over the network** and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

3. Resource pooling

The provider's computing resources are **pooled to serve multiple consumers** using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data centre). Examples of resources include storage, processing, memory, and network bandwidth.

4. Rapid elasticity

Capabilities can be elastically provisioned and released, in some cases automatically, to **scale rapidly** outward and inward commensurate **with demand**. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

5. Measured service.

Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource **usage can be monitored**, **controlled, and reported**, providing transparency for both the provider and consumer of the utilized service.

Cloud Service Models

1. Software as a Service (SaaS)

The capability provided to the consumer is to use the provider's **applications running on a cloud infrastructure**. The applications are accessible from various client devices through either 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, with the possible exception of limited user specific application configuration settings.

2. Platform as a Service (PaaS)

The capability provided to the consumer 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 **consumer does not manage or control the underlying cloud infrastructure** including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

3. Infrastructure as a Service (IaaS)

The capability provided to the consumer is to **provision processing, storage, networks**, **and other fundamental computing resources** where the consumer is able to 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, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).

Cloud Deployment Models

1. Private cloud

The cloud infrastructure is provisioned for **exclusive use by a single organization** comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

2. Community cloud

   The cloud infrastructure is provisioned for **exclusive use by a specific community of consumers from organizations** that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

3. Public cloud

   The cloud infrastructure is provisioned **for open use by the general public**. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

4. Hybrid cloud

   The cloud infrastructure is a **composition of two or more distinct cloud infrastructures** (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

AWS Cloud Infrastructure

AWS Regions

AWS has the concept of a Region, which is a **physical location** around the world where they **cluster data centres**. Each AWS Region consists of multiple, isolated, and physically separate AZs within a geographic area.

Availability Zones

An Availability Zone (AZ) is **one or more discrete data centres** with redundant power, networking, and connectivity in an AWS Region. All AZs in an AWS Region are interconnected with high-bandwidth, low-latency networking, over fully redundant, dedicated metro fibre providing high-throughput, low-latency networking between AZs.

AWS Local Zones

AWS Local Zones place compute, storage, database, and other select AWS services closer to end-users. Each AWS Local Zone location is an **extension of an AWS Region** where you can run your latency sensitive applications using AWS services such as Amazon Elastic Compute Cloud, Amazon Virtual Private Cloud, Amazon Elastic Block Store, Amazon File Storage, and Amazon Elastic Load Balancing in geographic proximity to end-users.

AWS Wavelength

AWS Wavelength Zones is an AWS infrastructure deployment that is **embed** AWS compute and storage services **within the telecommunications providers' datacentres** at the edge of the 5G networks, and seamlessly access the breadth of AWS services in the region. This enables developers to deliver applications that require single-digit millisecond latencies such as game and live video streaming, machine learning inference at the edge, and augmented and virtual reality (AR/VR). AWS Wavelength brings AWS services to the edge of the 5G network, minimizing the latency to connect to an application from a mobile device.

AWS Outposts

AWS Outposts bring native AWS services, infrastructure, and operating models to virtually any data centre, co-location space, or **on-premises facility**. You can use the same AWS APIs, tools, and infrastructure across on-premises and the AWS cloud to deliver a truly consistent hybrid experience. AWS Outposts is designed for connected environments and can be used to support workloads that need to remain on-premises due to low latency or local data processing needs.

Basics of AWS Services and Account

AWS Core Service Categories

1. Compute Services
2. Storage Services
3. Networking and Content Delivery
4. Security, Identity, and Compliance
5. Database Services
6. Management and Governance

Type of User Account

1. Root Account

a. The root user is the owner of the AWS account.
b. Never share your root user credentials with anyone.
c. Do not use the root user for my day-to-day work.
d. Enable MFA for Root user.
e. Do not create access for the root user.

2. IAM Account

a. Create an IAM user with Admin Privileges for doing admin work in AWS.
b. Enable MFA for high privilege IAM users.

AWS Access Methods

1. AWS Management Console
2. AWS CLI and AWS Cloud Shell
3. AWS API and SDK

IAM Basics

AWS IAM stand for AWS **Identity and Access Management**. AWS IAM allows you to control access and permissions for other AWS web services and resources by defining the following.

1. Users

A user is an entity that you create in AWS to **represent the person or application** that uses it to interact with AWS.

2. Groups

   An IAM user group is a **collection of IAM users**. User groups let you specify permissions for multiple users, which can make it easier to manage the permissions for those users.

3. Policies

   A policy is an object in AWS that, when associated with an identity or resource, **defines their permissions**. Most policies are stored in AWS as **JSON documents**.

4. Roles

   An IAM role is an IAM identity that you can create in your account that has specific permissions. An IAM role is similar to an IAM user, in that it is an AWS identity with permission policies that determine what the identity can and cannot do in AWS. However, instead of being uniquely associated with one person, **a role is intended to be assumable by anyone who needs it**.

EC2 Basics

Amazon **Elastic Compute Cloud** (Amazon EC2) provides **scalable computing capacity** in the Amazon Web Services (AWS) Cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster.

EC2 Security Basics

Creating and Accessing an EC2 Instance requires at least the following security objects.

1. Key pair

   A key pair, **consisting of a private key and a public key**, is a set of security credentials that you use to prove your identity when connecting to an instance. Amazon **EC2 stores the public key**, and **you store the private key**. You use the private key, instead of a password, to securely access your instances. Anyone who possesses your private keys can connect to your instances, so it is important that you store your private keys in a secure place.

2. Security Group

   A security group acts as a **virtual firewall for your EC2 instances** to control incoming and outgoing traffic. **Inbound rules** control the incoming traffic to your instance, and **outbound rules** control the outgoing traffic from your instance.

EC2 Remote Connection

You have following methods to connect from your local computer to your Linux EC2 Instance.

1. SSH client

You can use any standard SSH client to **remotely connect to your EC2 Instance**.

2. EC2 Instance Connect

Amazon EC2 Instance Connect provides a simple and secure way to connect to your Linux instances and **internally uses SSH**. You can use EC2 Instance Connect to using a **browser-based client also**.

3. AWS Session Manager

Session Manager is a fully managed AWS Systems Manager capability that provides secure and auditable instance management **without the need to open inbound ports, maintain bastion hosts, or manage SSH keys**.

AWS Free tier and Budget

AWS Free tier

The AWS Free Tier provides customers the ability to explore and try out AWS services free of charge up to specified limits for each service. The Free Tier is comprised of three different types of offerings.

1. 12-month Free Tier
2. An Always Free offer
3. Short term trials

Services with a **12-month Free** Tier allow customers to use the product for free up to specified limits **for one year from the date the account was created**.

Services with an Always Free offer allow you to use the product for **free up to specified limits** as long as you have a valid AWS account.

Services with a short-term trial are free to use for a **specified period of time or up to a one-time limit** depending on the service selected.

AWS Budget

AWS Budgets allows you to set custom budgets to track your cost and usage from the simplest to the most complex use cases. With AWS Budgets, you can choose to be alerted by email or SNS notification when actual or forecasted cost and usage exceed your budget threshold.

References & Links

1. https://csrc.nist.gov/publications/detail/sp/800-145/final
2. https://aws.amazon.com/about-aws/
3. https://aws.amazon.com/about-aws/global-infrastructure/regions_az/?p=ngi&loc=2
4. https://docs.aws.amazon.com/IAM/latest/UserGuide/best-practices.html
5. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/managing-users.html
6. https://aws.amazon.com/free