CLOUD COMPUTING

NIST Definition

Cloud computing is a model for enabling ubiquitous, convenient, and on-demand network access to a shared pool of configurable computing resources— such as networks, servers, storage, applications, and services— that can be rapidly provisioned and released with minimal management effort or service provider interaction. This model shifts IT from a capital- intensive investment to a flexible, service-oriented approach.

Characteristics of the Cloud

- On-demand self-service

  Consumers can independently provision computing resources such as virtual machines, storage, or network capacity whenever needed. Provisioning is automated and does not require direct interaction with the cloud provider, enabling faster deployment and experimentation.

- Broad network access

  Cloud services are available over the network and accessed through standard protocols. This promotes use across heterogeneous client platforms, including mobile devices, laptops, workstations, and thin clients, supporting modern distributed and remote work environments.

- Resource pooling

  The provider pools computing resources to serve multiple consumers using a multi-tenant model. Physical and virtual resources are dynamically assigned and reassigned according to demand. Customers generally do not

know the exact physical location of resources, but may specify location at a regional or data-center level. Pooled resources include CPU, memory, storage, and network bandwidth.

- Rapid elasticity

  Capabilities can be elastically scaled up or down, often automatically, in response to workload demand. To consumers, available resources may appear unlimited, allowing applications to handle sudden spikes in traffic without manual intervention.

- Measured service

  Cloud systems automatically monitor, control, and optimize resource usage through metering mechanisms. Usage is measured at an appropriate level (e.g., storage used, compute hours, network traffic), enabling pay-per-use, chargeback, or showback models with transparency for both provider and consumer.

As-a-Service Models

As-a-Service refers to delivering IT capabilities as managed services over the cloud. Instead of owning and maintaining infrastructure or software, organizations consume services provided and operated by third parties. The three primary service models differ in how responsibility is shared between provider and consumer: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

Platform as a Service (PaaS)

PaaS provides a complete development and deployment environment in the cloud, abstracting infrastructure concerns such as servers, operating systems, and middleware.

Development teams can build, test, deploy, and scale applications without managing the underlying infrastructure, allowing them to focus on code, application logic, and innovation.

PaaS often acts as an internal platform offering APIs, development tools, managed databases, CI/CD pipelines, documentation, and support services, forming a key pillar of digital and organizational transformation.

Infrastructure as a Service (IaaS)

- IaaS delivers virtualized computing resources that replace traditional on-premises hardware.
- The cloud provider manages physical infrastructure such as servers, storage, networking, and data centers.
- Organizations retain control over operating systems, middleware, applications, and configurations.
- Resources are managed through APIs or web-based dashboards, enabling automation and infrastructure-as-code practices.
- IaaS is the most flexible cloud service model and is well suited for workloads requiring customization, migration of legacy systems, or granular control over infrastructure.

Software as a Service (SaaS)

SaaS delivers fully functional applications over the internet, eliminating the need to install, manage, or maintain software locally.

The provider is responsible for updates, patches, security, and availability, while customers focus solely on usage.

SaaS applications typically use a multi-tenant architecture to isolate customer data and are commonly offered through subscription or usage-based pricing models.

Private Cloud

A private cloud provides cloud services exclusively to a single organization, either on-premises or hosted by a third party.

It is often chosen by organizations that cannot use public clouds due to:

- Strict security policies
- Regulatory or compliance requirements
- Data sovereignty or latency constraints

Private Cloud Aspects

- Security

  Data and applications remain within the organization's control, making private clouds suitable for sensitive or regulated workloads.

- Control and customization

  Infrastructure can be tailored to specific enterprise standards, performance requirements, and security policies.

- Cost considerations

While operational costs may be lower long-term, private clouds require significant upfront capital investment and ongoing maintenance.

- Scalability limitations

  Scaling resources requires planning and procurement, making elasticity more limited than in public clouds.

Public Cloud

A public cloud is owned and operated by a third-party provider and delivers shared computing resources to multiple customers over the internet. Resources are provisioned automatically through self-service portals and APIs, enabling rapid deployment and global scalability.

Public Cloud Aspects

- Scalability and elasticity

  Near-unlimited resources allow organizations to scale workloads dynamically based on demand.

- Cost efficiency

  Low upfront costs and pay-as-you-go pricing reduce capital expenditure.

- Shared responsibility

  Providers manage infrastructure security, while customers remain responsible for data, applications, and access control.

Hybrid Cloud

A hybrid cloud combines private and public cloud environments, allowing workloads and data to move between them as business needs change.

This model supports flexibility, cost optimization, regulatory compliance, and gradual cloud migration while maintaining a unified operational model.

6 Rs of Cloud Migration

- Rehosting

  Also known as lift-and-shift, this approach migrates applications with minimal or no code changes. It is fast but does not fully leverage cloud-native capabilities.

- Replatforming

  Introduces limited optimizations to better utilize cloud services while keeping the core architecture unchanged.

- Repurchasing

  Replaces existing applications with cloud-native SaaS alternatives, often improving efficiency but increasing vendor dependency.

- Refactoring / Re-architecting

  Redesigns applications to be cloud-native, often using microservices, containers, and managed services. This approach delivers the greatest long-term benefits but requires significant time and expertise.

- Retiring

  Decommissions applications that no longer provide business value, reducing complexity and operational costs.

- Retaining

Keeps certain workloads on existing infrastructure due to security, compliance, latency, or cost considerations, commonly used during long-term hybrid cloud strategies.