Specialization - Cloud Computing - I

Prof. Mohit K

Unit – 4

Cloud Storage

Introduction to cloud data storage

Enterprise Level Data

refers to the large-scale, diverse, and mission-critical data generated, processed, and managed by organizations to support their business operations, strategic decision-making, and long-term goals. This data is often characterized by the following features:

Volume: It involves handling large amounts of structured, semi-structured, and unstructured data generated from various business processes, customer interactions, and operational activities.

Variety: Enterprise data comes from multiple sources, including databases, CRM systems, IoT devices, social media, and third-party applications, encompassing formats like text, images, videos, and logs.

Velocity: It is often generated and processed in real-time or near-real-time, requiring robust systems to handle continuous streams of data efficiently.

Storage Systems In Cloud

Effective storage systems enhance cost-efficiency, performance, and ease of management, comprising Direct Attached Storage (DAS), Storage Area Network (SAN), and Network Attached Storage (NAS).

- **Direct Attached Storage (DAS):** DAS is the foundational storage system providing block-level storage and is critical in building SAN and NAS. Its high performance comes from direct system connections and it uses devices like SCSI, SATA, SAS, and Flash.
- **Storage Area Network (SAN):** SAN enables multiple hosts to connect to a single storage device, providing block-level storage. It is suitable for clustering environments but does not allow simultaneous access. Technologies include Fibre Channel (FC), iSCSI, and ATA over Ethernet (AoE).
- **Network Attached Storage (NAS):** NAS offers file-level storage, leveraging SAN and DAS as base systems. Known as a "File Server," NAS allows multiple hosts to share a single volume simultaneously, unlike SAN and DAS.
- **Layered Structure:** DAS serves as the base for SAN and NAS, with SAN positioned between DAS and NAS in the storage hierarchy.

Cloud Storage Types

Two primary types:

- **Structured Storage**: Organized data.
- **Unstructured Storage:** Unorganized or loosely structured data.

Structured Storage

- Data is stored in predefined formats like tables with rows and columns.
- High performance for transactions and queries.

Common cloud solutions:

- Amazon RDS
- Google Cloud Spanner
- Azure SQL Database

Use cases:

- Banking and finance
- CRM systems
- Inventory management
- Visuals: Database icons, relational table example

Unstructured Storage

Handles data with no predefined schema (e.g., images, videos, emails, and logs).

Scalable to store massive amounts of data.

Common cloud solutions:

- Amazon S3
- Google Cloud Storage
- Azure Blob Storage

Use cases:

- Media storage and streaming
- IoT data
- Backup and archival
- Visuals: File folder icons, media icons

Semi-Structured Data

Combines elements of structured and unstructured storage.

Stores data in flexible formats like JSON, XML, or NoSQL databases.

Common cloud solutions:

- Amazon DynamoDB
- MongoDB Atlas
- Azure Cosmos DB

Use cases:

- Real-time analytics
- E-commerce catalogs
- Social media feeds
- Visuals: JSON file structure illustration

Unstructured Storage

Feature	Structured Storage	Unstructured Storage	Semi-Structured Storage
Format	Predefined schema	Flexible, no schema	Flexible, schema optional
Examples	SQL Databases	S3, Blob Storage	NoSQL, JSON, XML
Performance	Optimized for queries	Optimized for scalability	Optimized for flexibility
Use Cases	Financial data	Media, backups, IoT	Real-time analytics

Unstructured Storage

Structured Storage: Best for transactional and relational data.

Unstructured Storage: Ideal for media, backups, and IoT data.

Semi-Structured Storage: Flexible for mixed data types and real-time use cases.

Storage Systems

- Effective storage system design ensures cost-efficiency, high performance, and easy management.
- Types of Storage Subsystems:
  - Direct Attached Storage (DAS) – Basic storage system.
  - Storage Area Network (SAN) – Connects multiple hosts to a single storage device.
  - Network Attached Storage (NAS) – Provides file-level storage, built on DAS and SAN.

Direct Attached Storage (DAS) & Storage Area Network (SAN)

DAS: Direct Attached Storage

Provides block-level storage, directly connected to a system.

High performance; foundation for SAN and NAS.

Storage devices: SCSI, SATA, SAS, FC, Flash, RAM.

SAN: Storage Area Network

Used when multiple hosts need a single storage device.

Provides block-level storage but allows access to one host at a time.

Technologies: Fibre Channel (FC), iSCSI, AoE (ATA over Ethernet).

Network Attached Storage (NAS)

Network Attached Storage

- NAS is built on DAS and SAN, providing file-level storage.
- Also known as File Server.
- Key Advantages:
- Multiple hosts can share a single volume simultaneously.
- Unlike SAN and DAS, which allow only one client per volume.

Data Storage Management

Introduction

Storage is expensive; tiered storage helps balance cost and performance.

- Fibre Channel provides high performance but is costly.
- SAS/DAS is cost-effective but lower in performance.
- IT organizations use a mix of storage technologies.

Data Storage Management Tools

- Storage administrators use tools to manage and monitor storage devices.
- Key tasks: configuration, migration, provisioning, archiving, and monitoring.

Data Storage Management Tools

- Configuration tools handle the set-up of storage resources. These tools help to organize and manage RAID(Redundant Array of Independent Disk) devices by assigning groups, defi ning levels or assigning spare drives.
- Provisioning tools define and control access to storage resources for preventing a network user from being able to use any other user's storage.
- Measurement tools analyse performance based on behavioural information about a storage device. An administrator can use that information for future capacity and upgrade planning.

Storage Management Process

Storage management relies on policies to govern storage device usage.

Encompasses three key areas:

- Change Management
- Performance & Capacity Planning
- Tiering (Tiered Storage)

Change Management

Process to request, schedule, implement, and evaluate storage adjustments.

Defines:

- How a request is made & approved
- Steps for configuring & provisioning storage
- Data migration processes to ensure integrity & availability

Performance & Capacity Planning && Data Storage Challenges

Performance & Capacity Planning

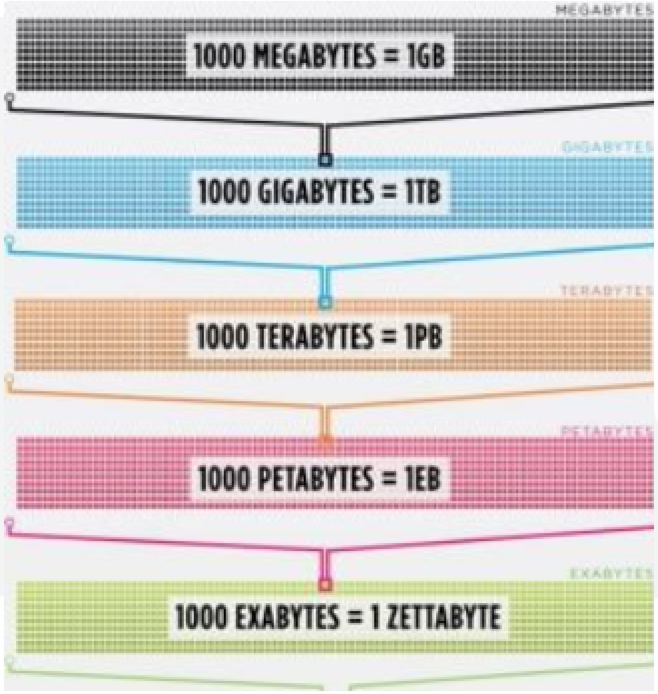
- **Measures system performance in terms of storage & utilization.**
- **Analysis helps in making informed decisions for future storage purchases.**

Data Storage Challenges

- Storage management must address key challenges:
  - Massive Data Demand
  - Performance Barrier
  - Power Consumption & Cost

Massive Data Demand

Digital data expected to grow exponentially (45 ZB by 2020).



Cloud data stores

A data store is a repository where data is stored as objects.
It includes:

- Databases
- Flat files
- Other storage systems

Types of Data Stores && Distributed Data Store

- Relational Databases (MySQL, PostgreSQL, SQL Server, Oracle)
- Object-Oriented Databases
- Operational Data Stores
- Schema-less Data Stores (Apache Cassandra, DynamoDB)
- Paper Files
- Data Files (Spreadsheets, Flat Files)

DISTRIBUTED DATA STORE

Similar to distributed databases, data is stored across multiple nodes. These are non-relational databases optimized for fast searches.

Examples:

- Google BigTable
- Amazon Dynamo
- Windows Azure Storage

Benefits of Distributed Data Stores

- **High Availability**: Data is stored across multiple nodes.
- **Scalability:** Grows with data needs.
- **Fault Tolerance:** Error correction and file recovery techniques.
- **Performance**: Optimized for quick searches over large datasets.

Distributed Data Stores

Software	Features
Apache Accumulo	Built on Hadoop & ZooKeeper, Java-based
Apache Cassandra	Combines Dynamo & BigTable features
HBase	Supports BigTable & Java programming
Hypertable	Designed for clustered storage & processing
KDI (Kosmix)	BigTable clone in C++

Relational vs. Distributed Data Stores

Feature	Relational Database	Distributed Data Store
Data Storage	Tables & Rows	Objects & Key-Value
Scalability	Limited	Highly Scalable
Fault Tolerance	Single Point Failure	Redundant & Distributed
Performance	Moderate	High Speed Queries

Provisioning cloud storage

- Cloud Storage allows sharing third-party resources via the Internet on a need basis.
- Increases efficiency by using remote storage devices from service providers.
- Scalability: Storage capacity expands as needed using multi-tenancy.
- Private Storage Cloud: Located behind an organization's firewall for in-house use.
- Cloud Data Management Interface (CDMI):
- Standardizes storage metering & billing.
- Allows IT organizations to connect with multiple providers without custom adapters.

Data-Intensive Computing Overview

Data-intensive computing processes large volumes of data (big data) using parallel computing techniques. It differs from compute-intensive computing, which focuses on execution time for complex computations.

Processing Approach:

- Uses parallel computing with multiple processors and disks in clusters connected via high-speed networks.
- Enables scalability and performance improvement by distributing data processing across multiple computing resources.

Processing Approach & Key Characteristics

Processing Approach:

Uses parallel computing to distribute workload across multiple processors and disks.

Data is processed independently within computing clusters.

Improves performance & scalability.

Key Characteristics of Data-Intensive Systems:

Mechanism for data collection & computation.

Programming model used.

Reliability & availability of data.

Scalability in both hardware & software.

Key Characteristics

1. Mechanism for data collection and computation.
2. Programming model used.
3. Reliability and availability of the system.
4. Scalability of both hardware and software.

System Architecture:

MapReduce (Hadoop)

- Developed by Google, later implemented as Hadoop (open-source).
- Uses a key-value pair approach for data processing.
- Automates data partitioning, scheduling, and execution, making it user-friendly for non-experts.

HPCC (High-Performance Computing Cluster)

- Developed by LexisNexis Risk Solutions.
- Uses commodity hardware running on Linux OS.
- Includes custom middleware and a high-level language ECL for efficient data-intensive computing.

Cluster Computing: A network of interconnected computers working together as a unified system to process large-scale computations

Parallel Computing: The simultaneous execution of multiple computations to speed up processing, often using specialized hardware like GPUs (NVIDIA CUDA, AWS EC2 GPU instances) and TPUs (Google Cloud TPUs).

Distributed Computing: A system where multiple nodes process data across different machines, enabling scalable computing

Cloud Storage: From LANs to WANs Introduction & Evolution

- Cloud storage has revolutionized data management.
- Transition from Local Area Networks (LANs) to Wide Area Networks (WANs).
- Benefits include scalability, flexibility, and cost efficiency.

Evolution
Traditional Storage: Physical servers and LAN-based storage.

- Modern Cloud Storage: Internet-based storage on distributed servers.

- Key Technologies: Virtualization, Software-Defined Storage (SDS), and Distributed File Systems.

Characteristics of Cloud Storage

Scalability – Expands with demand.

Elasticity – Pay-as-you-use model.

On-Demand Access – Accessible anytime, anywhere.

Multi-Tenancy – Shared resources among users.

Reliability & Redundancy – Data replication for fault tolerance.

Distributed Data Storage - Concept & Architecture

- Data is stored across multiple geographically distributed locations.
- Ensures data availability and disaster recovery.
- Uses technologies like Object Storage, Distributed File Systems (DFS), and Block Storage.

Architecture

Storage Nodes: Multiple storage units in different locations.

- Data Replication: Ensures redundancy and fault tolerance.

- Load Balancing: Manages data requests efficiently.

Applications Utilizing Cloud Storage

- Backup and Disaster Recovery – Redundant cloud storage.
- Big Data Analytics – Cloud storage supports data-intensive computations.
- IoT Data Storage – Stores sensor and real-time data.
- Enterprise Collaboration – Google Drive, Dropbox.
- Content Delivery Networks (CDN) – Faster data delivery via caching.

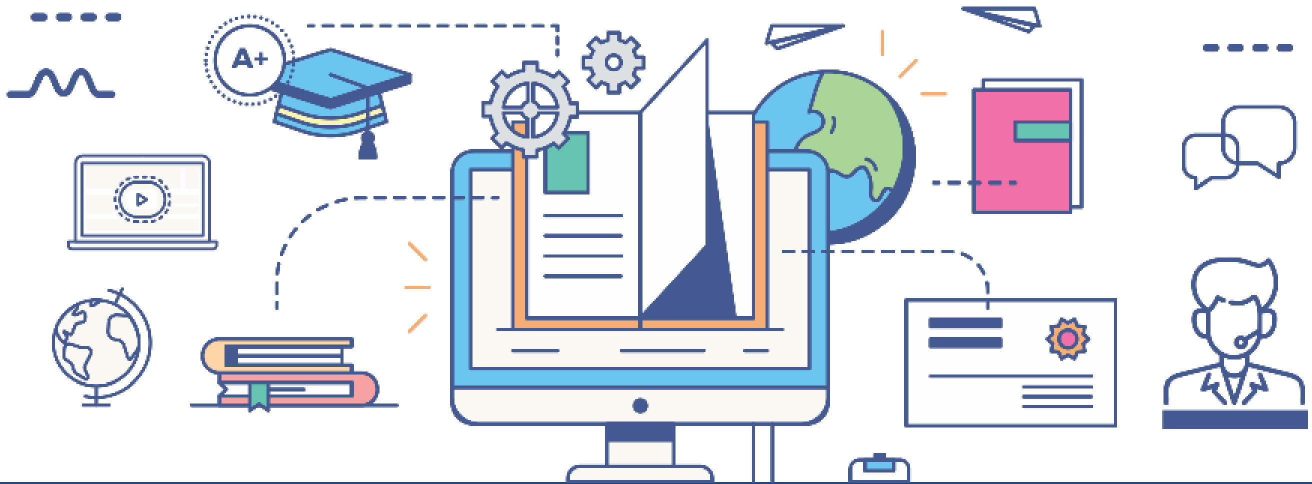
Benefits & Challenges and Future Trends

- Cost-effectiveness.
- Improved accessibility and collaboration.
- Security and compliance measures.
- Automatic updates and maintenance.

Challenges: Security risks, latency, regulatory compliance.

Future Trends: AI-driven storage optimization, decentralized cloud storage (Blockchain), Edge computing integration.

DIGITAL LEARNING CONTENT



Parul® University

www.paruluniversity.ac.in