UNIT 4 - DATA STORAGE AND SECURITY IN CLOUD

CLOUD FILE SYSTEMS

Cloud file systems are specifically designed to be distributed and operated in the cloud based environment.

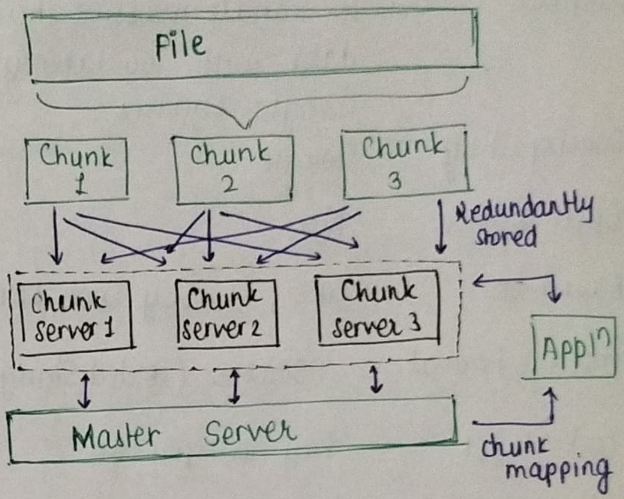
GFS (Google File System) DSBDA unit 3

A distributed file system developed by Google to store and manage huge files across many servers.

Key feature :- Master slave architecture
- fault tolerant (chunk replicated 3times)
- Optimized for large files, sequential read & write.

Used for :- Google search
Gmail
Google Maps backend.

Components :- Master Node, Chunkservers, Clients



HDFS (Hadoop Distributed file system) refer DSBDA unit 3

Developed by :- Apache

Purpose :- Open source storage for Big Data processing using Hadoop.

Features :- Name node (Master): manages metadata.
Data node (Slaves): store blocks (128 MB each).
Replication : Default 3 copies of each block.
Batch processing friendly (works great with MapReduce)

Used for :- Analytics, Data warehousing,
Offline data crunching.

GFS is Google's private file system to handle their own massive data needs.
HDFS is open source version inspired by GFS, used in Hadoop ecosystem for BigData processing.

Feature	GFS	HDFS
Developer	Google	Apache
Use case	search, maps	MapReduce, spark
chunk size	64 MB	128 MB / 256 MB
Architecture	Master - Slave	NameNode - DataNode
Datastorage unit	Chunks	Blocks
Programming language	C++	Java
Implementation	Proprietary (not open source)	Open Source.

MapReduce refer DSBDA unit 3

Stages :
Splitting phase
Mapping phase
Shuffle & sorting phase
Reduction phase

Cloud Database (build on-top of GFS/HDFS)

BigTable and HBase

Aspects	Big Table (Google)	HBase (Apache)
Definition	A distributed, column-oriented NoSQL database built by Google for handling big data.	An open-source, distributed NoSQL database modeled after Big Table, built on Hadoop.
Type	NoSQL	NoSQL
Purpose	To handle massive structured data with low latency across Google services.	To provide scalable, random access to big data in Hadoop ecosystem.
Developed by	Google	Apache Software foundation
Open source	No	Yes
Based on	GFS, Chubby lock service	HDFS, Zookeeper
Storage format	SSTables (sorted String Tables)	HFiles
Cost Model	Pay-as-you-go	Free software; pay only for infra
Use cases	Google Search, Gmail, Google Maps, YouTube Analytics.	Facebook Messenger, IoT sensor data, time series analytics, logs.

Dynamo cloud data stores aka DynamoDB (simple DB and datastore and 2 more)

Amazon DynamoDB is a fully managed, serverless NoSQL database service offered by AWS, inspired by original Dynamo system developed at Amazon

Features :-

- Managed by AWS
- Key-value & Document store
- Highly scalable
- Single digit millisecond performance at any scale (Superfast read & write)
- High Availability (Replicates data across multiple availability zones (AZ))
- Serverless
- Built in security
- Time-to-Live (Autodeleted expired items)
- ACID transactions

Advantages :-

- fully managed.
- Auto scaling (handles sudden traffic spikes without manual scaling).
- High durability & availability.
- Same performance at scale (large no. of users).
- flexible schema.
- Pay only for what you use.
- No downtime deployments (updates apps without downtime)
- Low latency (Fast read/writes)

Used By :- Amazon, Netflix, Airbnb, Snapchat, Zoom, Lyft, etc

Limitations :- No complex transactions, Size limits (400KB), Complex data modelling

Dynamo Cloud Data Stores - is a category of cloud based NOSQL databases that follow the principles of original Dynamo system created by Amazon in 2007.

Principles are :- High available, Eventually consistent, Distributed and horizontal scalable, NOSQL

Examples :- Amazon Dynamo DB, Amazon Simple DB, Google Cloud Datastore, LinkedIn Voldemort, facebook Cassandra

Simple DB - is a NOSQL key value cloud database service launched by AWS in 2007. It was designed for simple queries, schema-less storage, & high availability - inspired by Amazon's Dynamo System.

feature	Simple DB	Dynamo DB
Definition	Early AWS NOSQL DB service (2007), attribute-value store.	Modern, fully managed NOSQL DB (2012), key-value + document store.
Storage Limit	~10GB per domain	Virtually unlimited (upto 400KB per item, scale across partition).
Scalability	Limited	High
Query language	Basic SQL like syntax	NOSQL styled query
Performance	Good	High
DAX	Not available	Available

Comparisions :

Aspects	Dynamo DB	RDBMS
Type	NOSQL	SQL
Datastructure	key-value, items in table	Rows & columns
Schema	Schemaless	fixed schema
Scalability	Horizontal	Vertical
Joins	No	Yes
Use case	Real time apps, IoT, gaming, etc	Banking systems, Reporting, etc
Examples Systems	Amazon DynamoDB, Cassandra, Mongo DB	MySQL, PostgreSQL, Oracle, SQL Server

Cloud Storage - storing data in remote servers accessed over the internet (cloud) managed by cloud providers

- Features : Accessible, Backup, recovery, durability, supports files, blobs and objects.

Cloud Storage Providers

Amazon AWS (S3), Google Cloud (Google Cloud Storage), Microsoft Azure (Azure Blob Storage), IBM Cloud, Oracle Cloud, etc.

SECURING THE CLOUD

Issues in securing the cloud.

→ Data Breaches sensitive data maybe exposed)
→ Unauthorised access
→ Data loss (hardware failure)
→ DOS
→ Insecure API's
→ Insecure Endpoints (User Devices)

Cloud security involves protecting data, applications, and services hosted in the cloud. It includes technical solutions, policies and best practices to prevent unauthorized access, data loss or breaches.

General Security Advantages of Cloud Based Solutions.

1) Advanced security Infrastructure - cloud provides strong physical & network security
2) Automatic security updates & patching - systems are regularly updated.
3) Data Encryption - to prevent unauthorized access.
4) Access Control and Identity Management - RBAC, MFA and SSO enhance security.
5) Disaster Recovery and Data Backup - help to protect against data loss..
6) Centralized Security Management - Security policies and monitoring are managed from a central dashboard.
7) Compliance with Industry Standard - ISO27001, HIPAA, GDPR, etc.
8) Security Monitoring and Threat Detection - real time monitoring & AI powered threat detection improve early warning & response.
9) Staffing and Expertise - Cloud vendors employ specialized security teams, allowing customers to benefit from their expertise without direct hiring.
10) Cross Pollination of Security learnings - lessons from one customer's threats improve protection for all cloud users.

Introducing Business Continuity and Disaster Recovery

-Business continuity ensures that an organization can maintain essential functio during and after disruption (outrages, cyberattacks or disasters)

Disaster Recovery focuses on restoring IT systems and data access following a catastropic event.

- BC - Cloud services help by providing features such as data replication, automated failover, and remote access, so employees and customers can continue to interact with your systems even if something goes wrong.

DR - In cloud, this usually involves backing up data to multiple locations, using automated tools to switch operations to backup systems, and quickly recovering lost or corrupted data. Cloud based DR is faster and often cheaper than traditional methods because you don't need to maintain your own backup infrastructure.

How to Approach Business Continuity !?

Assemble Team → Identify Tasks → Business Impact Analysis → Develop strategies
↓
Train Staff ← Test & Update ← Document & Implement ← Backup & Redundancy

Disaster Recovery - Understanding the Threats

Types of Disasters

1) Natural Disasters - Earthquakes, floods, fires, etc.
These can destroy on premises data centers, connectivity to cloud resources
2) Hardware or System failures - server crashes, power outages, disk failure
Physical Hardwares can fail.
3) Cyber Attacks - Ransomware, DDoS, Data breaches.
These target cloud infra or user data, causing server interruption or data loss.
4) Human Errors
5) Software Bugs
6) Network failures

Disaster Recovery on Cloud Platform.

1) Backup & Restore
2) Pilot Light - A minimal version of your system runs in the cloud. During a disaster, you quickly scale it into a full production system.
3) Warm Standby - A scaled-down version of your system runs continuously. Can quickly scaleup to full capacity if needed.
4) Multi-AZ Deployment - Applications & databases are replicated across availability zones.
5) Cloud DR as a Service (DRaaS) - Fully managed by cloud provider.

Threats in DR

→ Lack of DR plan
→ Inadequate testing
→ Complexity of modern systems
→ Slow Response and Recovery Times
→ Data Consistency & Integrity issues.
→ Cybersecurity threats.
→ Inappropriate Datacenter locations.
→ Network & Infra Limitations.

Architect of failure - designing systems with expectation that things will go wrong - hardware might break, software may crash or networks could go down. Instead of hoping for perfect uptime, you build in redundancy, automatic failover, and monitoring so the systems can keep running or recover quickly when something fails. This makes system more reliable & resilient.

Fault Tolerance - ability of system to continue operating properly even when one or more of its components fail. This means that if hardware, software or network components encounter errors or stop working, the system as whole remains functional & users experience little or no disruption.

Characteristics :—

No single point of failure - one failure does not affect whole system.

Redundancy - Duplicate components included.

Error Detection & Recovery

Continuous Availability

Load Balancing - distributes traffic.