Storage and Database

CSYE 6225: Network Structure & Cloud Computing
Northeastern University

Instructor: Raja Alomari,PhD

M6: Overview

Topics include:

- Cloud Storage: Object Store, S3, EFS, EBS.
- Overview of Databases.
- Databases in cloud environments including RDS, DynamoDB, Redshift.

Database and Cloud Storage

Objectives:

- Understand the various types of cloud storage.
- Clarify key aspects of cloud storage and databases.
- Understand the various types of cloud databases.
- Clarify key aspects of cloud databases.
- Hands-on with various storage types and cloud databases.

Introduction to Database Management Systems (DBMS)

A **database** is an organized collection of **structured** data or information stored electronically.

Key features:

- Databases enable efficient storage, retrieval, and management of large amounts of data.
- They are critical for applications ranging from websites, finance systems, and ERP systems to IoT and cloud platforms.

What is a Database Management System (DBMS)?

Definition:

- A DBMS is software that interacts with the user, applications, and the database itself to capture and analyze data.
- It facilitates data management, storage, manipulation, and retrieval in a database.

Key Functions:

- **Data Definition:** Allows the creation, modification, and removal of definitions that define the organization of the data in the database.
- **Data Updating:** Inserting, modifying, and deleting data within the database.
- **Data Retrieval:** Provides querying capabilities to extract useful information.
- **Administration:** Handles tasks like security, backup, recovery, and access control.

Types of Databases

Relational Databases (RDBMS)

- Stores data in tables (rows and columns).
- Uses SQL (Structured Query Language) to manage and query data.
- Examples: MySQL, PostgreSQL, Oracle, SQL Server, AWS RDS.

NoSQL Databases

- Designed for large-scale data and real-time web applications.
- Uses various data models like key-value, document, graph, and wide-column stores.
- Examples: MongoDB, AWS DynamoDB, Cassandra.

Types of Databases

Object-Oriented Databases

- Integrates object-oriented programming and databases.
- Stores data as objects, similar to how data is stored in object-oriented programming.
- Examples: db4o, ObjectDB.

Data Warehouses

- Optimized for querying and reporting, rather than transaction processing.
- Example: Amazon Redshift, Snowflake.

Relational Database Management Systems (RDBMS)

Definition:

- RDBMS is a type of DBMS based on the relational model.
- Data is stored in relations (tables) with relationships between them.

Key Features:

- ACID Properties: Atomicity, Consistency, Isolation, Durability.
- Normalization: Organizing data to reduce redundancy.
- SQL: Standard language for managing relational databases.

Popular RDBMS Examples:

- MySQL: Open-source, widely used.
- PostgreSQL: Advanced open-source RDBMS with powerful features.
- Oracle Database: Enterprise-level database with high security and performance.

NoSQL Databases

Definition:

- NoSQL databases provide a mechanism for storage and retrieval of data that is modeled differently than in relational databases (e.g., not necessarily using tables).

Data Models in NoSQL:

- Key-Value Stores: Stores data as a collection of key-value pairs (e.g., DynamoDB, Redis).
- Document Databases: Stores data in documents (e.g., MongoDB, Couchbase).
- Graph Databases: Stores data in graph structures with nodes, edges, and properties (e.g., Neo4j).

Advantages:

- Scalability: NoSQL databases scale horizontally across many servers.
- Flexibility: Ideal for semi-structured or unstructured data.
- High Performance: Low-latency access for real-time applications.

SQL vs NoSQL

- SQL (Relational):
  - Structure: Tables (rows and columns).
  - Schema: Fixed schema, data must follow the structure.
  - Query Language: SQL (Structured Query Language).
  - Use Case: Best for structured data and complex querying.
- NoSQL:
  - Structure: Key-value pairs, documents, graphs, etc.
  - Schema: Dynamic schema, no fixed structure.
  - Query Language: Varies by database (e.g., MongoDB query language).
  - Use Case: Best for unstructured, semi-structured data, or large-scale real-time data.

ACID Properties in DBMS

- Atomicity:
  - Ensures that all the steps in a transaction are completed successfully or none at all.
- Consistency:
  - Ensures that a transaction brings the database from one valid state to another valid state.
- Isolation:
  - Ensures that transactions are executed independently without interference.
- Durability:
  - Ensures that the results of a committed transaction are permanent, even in the case of a system failure.

CAP Theorem in Distributed Databases

- Definition:
  - CAP theorem states that in a distributed system, you can only guarantee two out of the three following properties:
  - Consistency: Every read receives the most recent write.
  - Availability: Every request receives a response, without guarantee that it contains the most recent write.
  - Partition Tolerance: The system continues to function even if there is a network partition.
- Trade-offs:
  - No distributed system can simultaneously achieve all three properties, leading to trade-offs based on the application needs.

AWS Database Services

- **Amazon RDS:** Managed relational database service supporting multiple database engines.
- **Amazon DynamoDB:** Fully managed NoSQL database designed for high availability and scalability.
- **Amazon Redshift:** Data warehousing service for large-scale data analysis.
- **Amazon Aurora:** MySQL and PostgreSQL-compatible relational database with enhanced performance and availability.

Amazon RDS (Relational Database Service)

- Key Features:
  - Supports MySQL, PostgreSQL, Oracle, SQL Server, and MariaDB.
  - Automated backups, patches, and scaling.
  - Multi-AZ deployment for high availability and durability.
- Use Cases:
  - Enterprise applications, websites, and online transaction systems.
- Security:
  - Encryption with AWS KMS.
  - VPC integration for secure, isolated databases.

Amazon DynamoDB (NoSQL Database Service)

Key Features:

- Fully managed NoSQL database.
- Supports key-value and document data models.
- Automatically scales throughput capacity.

Use Cases:

- Mobile apps, gaming applications, IoT, and real-time data processing.

Security:

- Encryption at rest and in transit.
- IAM policies for fine-grained access control.

Amazon Redshift (Data Warehouse)

- Key Features:
  - Fully managed, petabyte-scale data warehouse service.
  - Optimized for complex SQL queries and analytics.
  - Supports integration with business intelligence tools.
- Use Cases:
  - Big data analytics, reporting, and business intelligence workloads.

Amazon Aurora (High-Performance RDBMS)

- Key Features:
  - MySQL and PostgreSQL-compatible relational database engine.
  - 5x performance of MySQL, 3x performance of PostgreSQL.
  - Auto-scaling, fault-tolerant, with automated backups and failover.
- Use Cases:
  - Mission-critical, high-performance, relational workloads.

Cloud Storage

Cloud Storage

Key Concepts in Cloud Storage

- **Durability and Availability:** Data replication across regions
- **Latency and Performance:** Differences between cold and hot data storage
- **Security:** Encryption, access control, and data compliance
- **Cost Models:** Pay-as-you-go, upfront savings, and data lifecycle management.

Types of Cloud Storage

- Object Storage (e.g., Amazon S3)
- Block Storage (e.g., Amazon EBS)
- File Storage (e.g., Amazon EFS)
- Archive Storage (e.g., Amazon S3 Glacier)

Amazon S3

Amazon S3 (Simple Storage Service) offers a variety of storage classes to accommodate different types of data access patterns and cost requirements. These classes help optimize costs depending on the frequency of data access and retrieval times.

Amazon S3 storage classes are designed to offer different levels of performance, availability, durability, and cost. These storage classes cater to various use cases, from frequently accessed data to long-term archival storage.

Amazon S3 Standard (S3 Standard)

Description:

- Designed for frequently accessed data.
- High durability (99.999999999% - 11 nines) and high availability (99.99%).
- Suitable for general-purpose storage with low latency and high throughput.

Use Cases:

- Frequently accessed data such as websites, content distribution, mobile applications, and gaming.
- Short-term storage for big data analysis.

Pricing (approximate/may change):

- Storage cost: $0.023 per GB/month.
- Request costs: GET requests $0.0004 per 1,000 requests, PUT requests $0.005 per 1,000 requests.

Amazon S3 Intelligent-Tiering

- Description:
  - Automatically moves data between two access tiers (frequent and infrequent) based on changing access patterns.
  - Suitable for data with unknown or unpredictable access patterns.
  - No retrieval fees; only monitoring and automation charges.
- Use Cases:
  - Data with changing access patterns, such as user-generated content, data lakes, or analytics data.
- Pricing (approximate/may change):
  - Frequent access tier: $0.023 per GB/month.
  - Infrequent access tier: $0.0125 per GB/month.
  - Monitoring and automation: $0.0025 per 1,000 objects/month.
  - No retrieval fees.

Amazon S3 Standard-Infrequent Access (S3 Standard-IA)

- Description:
  - Designed for infrequently accessed data but that still needs rapid access when required.
  - Offers lower storage costs but higher retrieval costs.
  - Same durability as S3 Standard.
- Use Cases:
  - Data that is accessed less frequently but must be available immediately when needed, such as backups or disaster recovery files.
- Pricing (approximate/may change):
  - Storage cost: $0.0125 per GB/month.
  - Retrieval cost: $0.01 per GB.
  - Request costs: GET requests $0.001 per 1,000 requests, PUT requests $0.01 per 1,000 requests.

Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA)

- Description:
  - Stores data in a single availability zone (AZ), offering lower cost than Standard-IA.
  - Suitable for non-critical, infrequently accessed data that does not require multi-AZ resilience.
- Use Cases:
  - Secondary backups, data that can be recreated if lost, or geographically distributed applications.
- Pricing (approximate/may change):
  - Storage cost: $0.0100 per GB/month.
  - Retrieval cost: $0.01 per GB.
  - Request costs: Same as S3 Standard-IA.

Amazon S3 Glacier

- Description:
  - Low-cost archival storage for data that is rarely accessed.
  - Retrieval times vary between minutes to hours depending on retrieval options.
  - Ideal for long-term archives where quick retrieval is not critical.
- Use Cases:
  - Regulatory archives, compliance data, long-term backups.
- Pricing (approximate/may change):
  - Storage cost: $0.004 per GB/month.
  - Retrieval cost:
    - Standard: $0.0036 per GB for 3-5 hours.
    - Expedited: $0.03 per GB for 1-5 minutes.
    - Bulk: $0.0025 per GB for 5-12 hours.
  - Request costs:
    - GET requests: $0.05 per 1,000 requests.

Amazon S3 Glacier Deep Archive

- Description:
  - Lowest-cost storage class designed for long-term retention of data that is rarely accessed.
  - Retrieval times range from 12 hours to 48 hours.
  - Suitable for data archiving where compliance or other regulations require long-term data storage.
- Use Cases:
  - Long-term backup, digital preservation, historical data archives.
- Pricing (approximate/may change):
  - Storage cost: $0.00099 per GB/month.
  - Retrieval cost:
    - Standard: $0.02 per GB (12-hour retrieval time).
    - Bulk: $0.0025 per GB (48-hour retrieval time).
  - Request costs:
    - GET requests: $0.05 per 1,000 requests.

Amazon S3 Outposts

Description:

- Brings S3 to on-premises environments using AWS Outposts.
- For workloads requiring on-premises data storage with low latency.

Use Cases:

- Applications requiring on-premises storage due to regulatory or latency requirements.

Pricing (approximate/may change):

- Custom pricing, based on configuration and use case.

S3 Overview

Storage Class	Availability	Retrieval Time	Storage Cost ($/GB) Approximate/May change	Use Case
S3 Standard	99.99%	Immediate	$0.023	Frequently accessed data, content distribution
S3 Intelligent-Tiering	99.9% - 99.99%	Immediate	$0.023 (frequent)	Data with unpredictable access patterns
S3 Standard-IA	99.90%	Immediate	$0.0125	Infrequently accessed data but needs fast retrieval
S3 One Zone-IA	99.50%	Immediate	$0.01	Infrequent access, single AZ for non-critical data
S3 Glacier	Varies	Minutes to hours	$0.004	Long-term archival storage
S3 Glacier Deep Archive	Varies	Hours to days	$0.00099	Ultra-low-cost, long-term archival
S3 Outposts	Varies (on-prem)	Immediate (on-prem)	Custom Pricing	On-premises storage with AWS Outposts

Amazon Elastic Block Store (Amazon EBS)

High-performance, block storage service designed to support Amazon EC2 instances.

EBS is highly reliable, durable, and scalable, allowing users to create, manage, and configure storage volumes optimized for different workloads.

Key Features of Amazon EBS

- **Durable and Reliable Storage:**
  - Provides 99.999% availability for each EBS volume.
  - Automatically replicates data within an Availability Zone (AZ) to protect against failures.
- **Scalable:**
  - Users can scale EBS volumes from 1 GiB to 16 TiB.
  - Elasticity allows adjusting volume size, performance, or type without downtime.
- **Snapshots for Backup and Disaster Recovery:**
  - Snapshots are point-in-time backups of volumes stored in Amazon S3.
  - Snapshots can be used to recover data or create new volumes.
- **Encryption:**
  - EBS offers AES-256 encryption at rest and in transit.
  - Integration with AWS Key Management Service (KMS) for easy key management.
- **Multiple Volume Types:**
  - Designed for diverse performance needs including general-purpose, provisioned IOPS, throughput-optimized, and cold storage.

Types of EBS Volumes

Volume Type	Description	Use Case	Max Throughput	Max IOPS/Volume	Pricing (approximate/may change)
General Purpose SSD (gp3)	Balanced price and performance for everyday workloads	Most common workloads (small/medium DBs, web servers)	1,000 MiB/s	16,000 IOPS	$0.08 per GB/month
General Purpose SSD (gp2)	Legacy SSD-based storage, ideal for general use	Boot volumes, small-to-medium databases	250 MiB/s	16,000 IOPS	$0.10 per GB/month
Provisioned IOPS SSD (io2)	High-performance SSD for mission-critical applications	High-performance databases (e.g., MySQL, Oracle)	1,000 MiB/s	64,000 IOPS	$0.125 per GB/month
Throughput Optimized HDD (st1)	Low-cost HDD for frequently accessed, throughput-intensive workloads	Big data, log processing, data warehouses	500 MiB/s	500 IOPS	$0.045 per GB/month
Cold HDD (sc1)	Lowest cost HDD for infrequently accessed data	Large volumes of cold data	250 MiB/s	250 IOPS	$0.015 per GB/month

Snapshots

Description:

- EBS Snapshots provide point-in-time backups of EBS volumes. Snapshots are stored in Amazon S3 and are incremental, meaning only changed data is saved in subsequent snapshots.

Key Features:

- Snapshots can be used to restore volumes or create new volumes.
- Can be shared across AWS accounts or used in different regions.
- Automated snapshots can be scheduled using AWS Backup or other automated policies.

Pricing:

- Snapshots are charged based on the storage used: approximately $0.05 per GB/month for snapshot data.

Encryption and Security

Data Encryption:

- EBS provides AES-256 encryption for volumes at rest, in transit, and for snapshots.
- Integrated with AWS Key Management Service (KMS) to manage encryption keys.
- Supports encrypted snapshots and encryption of existing unencrypted volumes via snapshots.

IAM Policies:

- Fine-grained permissions and role-based access using AWS Identity and Access Management (IAM) to control who can access specific volumes or perform specific actions (such as creating snapshots).

Amazon Elastic File System (Amazon EFS)

Amazon Elastic File System (EFS) is a fully managed, scalable, and cloud-native Network File System (NFS) designed to provide simple and elastic file storage for use with AWS cloud services and on-premises resources. EFS is ideal for applications that require shared access to a common file system and where the performance, durability, and scalability are critical.

Key Features of Amazon EFS

- Scalability and Elasticity:
  - EFS automatically scales up or down to accommodate workloads without provisioning or managing storage capacity.
  - Can scale from small file systems to petabyte-scale, handling millions of IOPS and delivering consistent, low-latency performance.
- Simple Integration with AWS Services:
  - Can be used directly with Amazon EC2, AWS Lambda, and other AWS services.
  - Supports NFS v4.1 and v4.0, making it compatible with a wide range of Linux-based applications.
- Shared Access:
  - Multiple EC2 instances or services can mount the same EFS file system concurrently, making it perfect for shared workloads.

Key Features of Amazon EFS

- Performance Modes:
  - Two performance modes: General Purpose (default) and Max I/O for high-scale workloads.
  - Also includes two throughput modes: Bursting and Provisioned, allowing performance customization.
- Durability and Availability:
  - Data stored in EFS is replicated across multiple AWS Availability Zones (AZs) in a region, ensuring high availability and 11 9s of durability (99.999999999%).
- Security:
  - Supports encryption at rest and in transit using AWS Key Management Service (KMS).
  - Integrates with IAM, VPCs, Security Groups, and Network Access Control Lists (NACLs) to restrict access.
- Cost-Effectiveness:
  - Pay-as-you-go pricing model, and options to save cost with EFS Infrequent Access (IA) for cold data. Optimized for both frequent and infrequent data access patterns.

EFS Storage Classes

- EFS Standard:
  - Optimized for frequently accessed files.
  - High availability and low-latency access to your active data.
  - Ideal for applications needing consistent, low-latency performance for their workloads.
- EFS Infrequent Access (EFS IA):
  - Designed for infrequently accessed files, providing cost savings for less frequently used data.
  - Automatically moves files between the Standard and IA classes based on lifecycle management policies.
  - Offers significant savings (up to 92% cost reduction) compared to Standard for data that is accessed rarely.

Hands-on

Objectives:

- Recognize various AWS storage technologies in the console.
- Recognize KMS and encryption in transit and at rest.
- Deploy RDS (and other) AWS databases.

Module 6 Conclusion

- Cloud computing offers diverse storage and database solutions, such as Amazon S3 and DynamoDB, to meet various application needs.
- High-performance storage and databases such as Amazon EBS and RDS deliver low-latency access for critical applications.
- Scalable solutions such Amazon S3 optimize costs for data management and access, making it ideal for backups and archiving.
- Cloud services such as Amazon Redshift enable efficient data management and analytics for large datasets.
- Cloud computing ensures secure, scalable, and compliant data management across industries, exemplified by AWS's robust security measures.