SESSION 10

AWS Database

Storage vs Database

Storage is for file storage such as images and PDFs. Storing data on disk(EFS, EBS, S3, EC2 Instance Store) can have its limits

Database is basically a storage but it stores data records which can be queried using a query language. You can structure the data. You build indexes to efficiently query/search through the data. You define relationships between your datasets.

- Highly Scalable: You can scale your database as your application grows without any downtime!
- Fully Managed: Everything, from maintenance to hardware upgrades, is managed by AWS.
- Enterprise Class: You get the same world-class infrastructure used by Amazon's giant ecommerce platform.
- Distributed: Now that your application and database exist on separate machines, your application becomes highly fault-tolerant.
- Workforce Reduction: Since everything is managed by AWS, you don't need a Database Maintenance team in your organization.

Relational Databases

- It contains several tables, and each table has its primary key.
- Due to a collection of an organized set of tables, data can be accessed easily in RDBMS.
- A table is a collection of related data entries and contains rows and columns to store data.
- Each table represents some real-world objects such as person, place, or event about which information is collected.
- Can use SQL language to perform queries.

NoSQL Databases

- NoSQL= non-SQL= non relational databases
- It is an approach to database design that enables the storage and querying of data outside the traditional structures found in relational databases.
- Instead of the typical tabular structure of a relational database, NoSQL databases, house data within one data structure, such as JSON document.
- They have following benefits: Flexibility, Scalability, High-performance, Highly functional
- NoSQL can handle a huge amount of data because of scalability, as the data grows NoSQL scale itself to handle that data in an efficient manner.

When should NoSQL be used:

- When a huge amount of data needs to be stored and retrieved.
- The relationship between the data you store is not that important
- The data changes over time and is not structured.
- Support of Constraints and Joins is not required at the database level
- The data is growing continuously and you need to scale the database regularly to handle the data.

NoSQL databases are generally classified into four main categories:

1. **Document databases:** These databases store data as semi-structured documents, such as JSON or XML, and can be queried using document-oriented query languages.
2. **Key-value stores:** These databases store data as key-value pairs, and are optimized for simple and fast read/write operations.
3. **Column-family stores**: These databases store data as column families, which are sets of columns that are treated as a single entity. They are optimized for fast and efficient querying of large amounts of data.
4. **Graph databases:** These databases store data as nodes and edges, and are designed to handle complex relationships between data.

1. **Relational Database**: In relational databases, the data is usually stored in a tabular format. Relational databases particularly use structured query language (SQL) to run queries to perform operations such as insertion, updating, deletion, and more. AWS provides the following relational database services.

- **Amazon RDS**
- **Amazon Redshift**
- **Amazon Aurora**

2. **Key–Value Database:** The key–value database is a type of NoSQL database where the method of having a value attached to a key is used to store data. Meaning that the data is composed of two elements, keys and values.

- **Amazon DynamoDB**

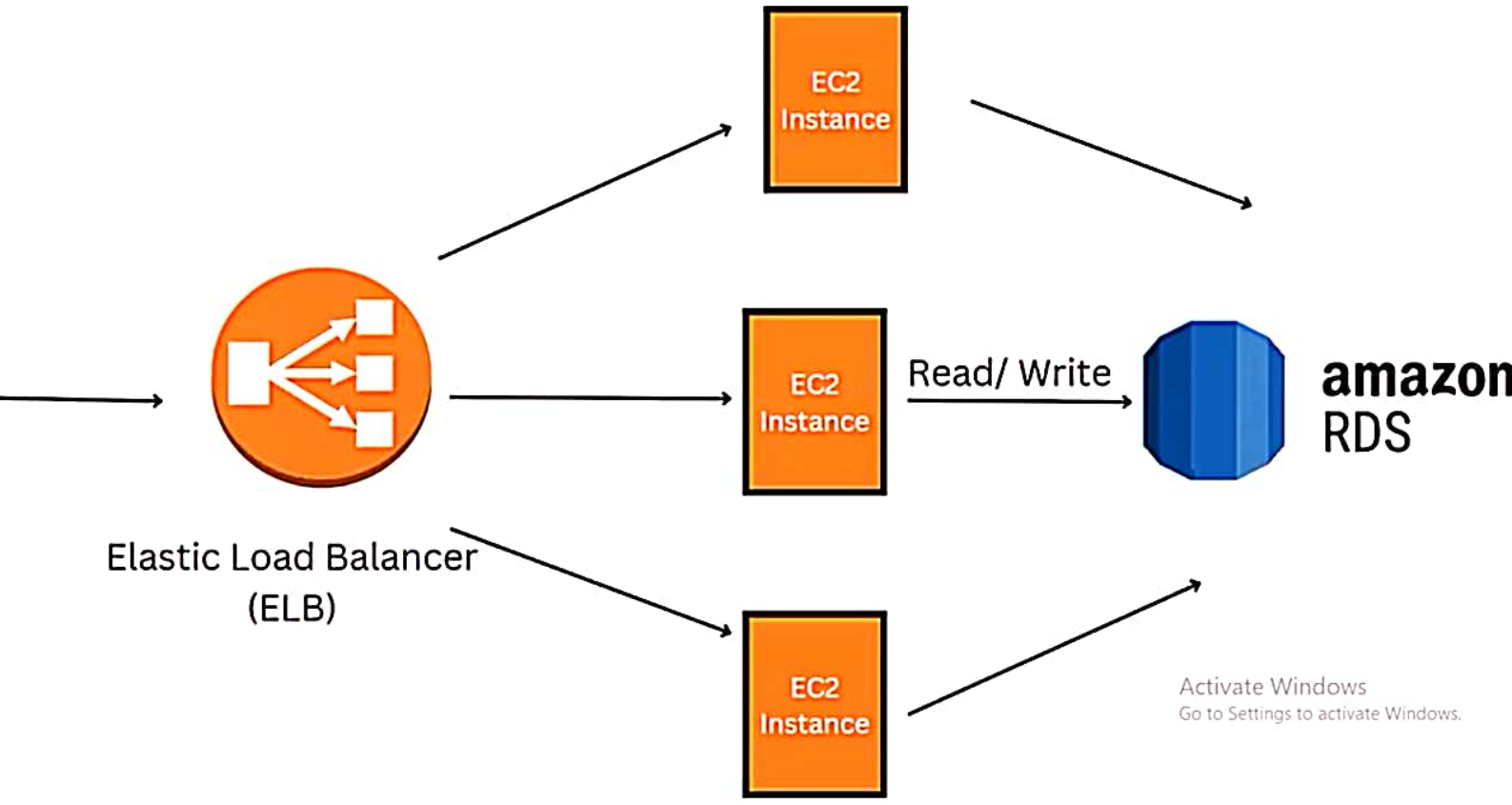
3. **In-memory Database:** This type of database is primarily based on the main memory for computer data storage. Basically, an in-memory database keeps the whole data in the RAM. Meaning that each time you access the data, you only access the main memory and not any disk. And the reason that the main memory is faster than any disk is why in-memory databases are so popular.

- **Amazon ElastiCache**

AWS RDS

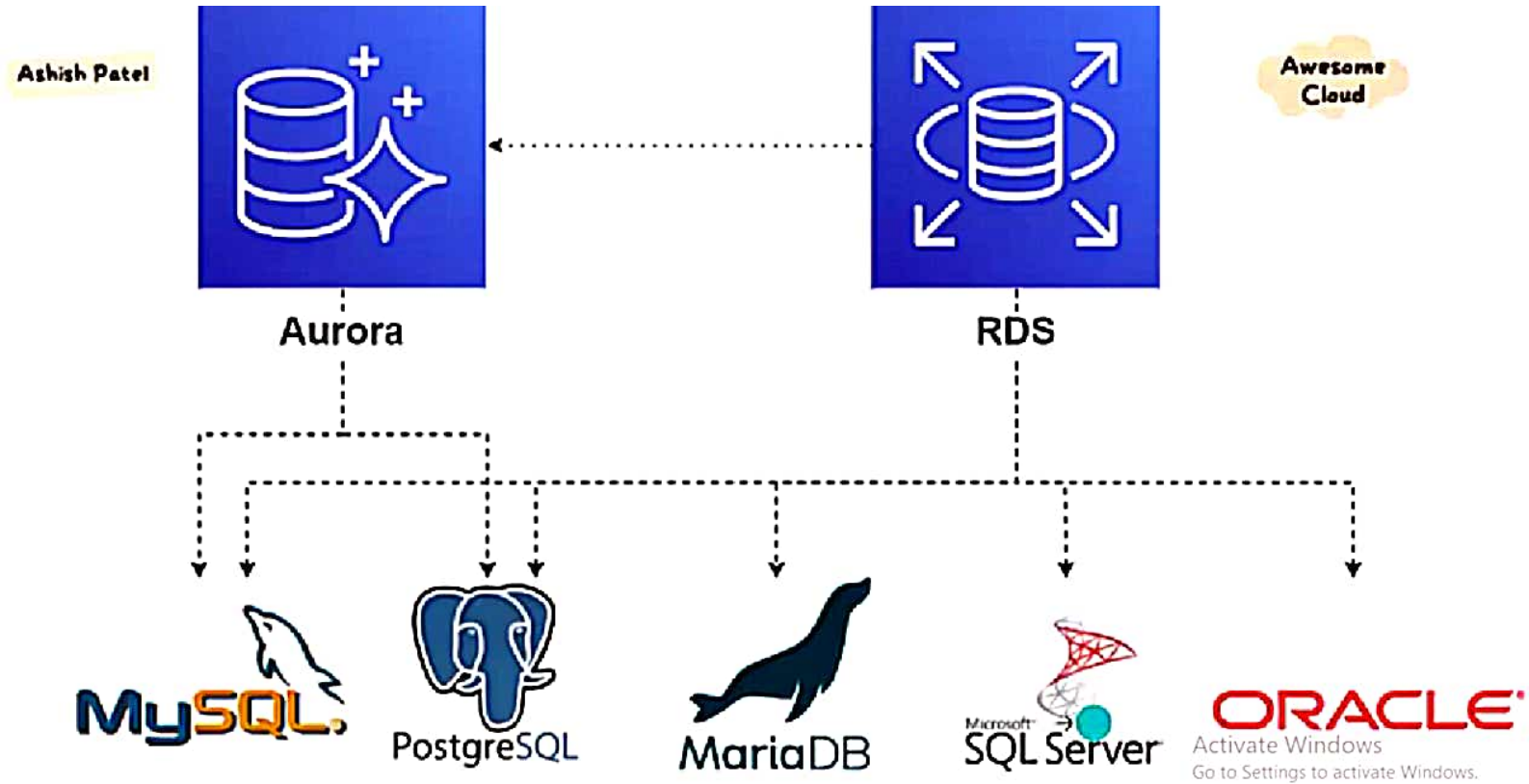
- RDS= Relational Database Service
- It is a managed DB service for DB use SQL as a query language
- It provides you with six familiar database engines to choose **Postgres, MySQL, MariaDB, Oracle, Microsoft SQL Server, Aurora**

Using RDS instead of deploying own EC2:
Since RDS is managed service, therefore continuous backups, monitoring dashboards, Multi AZ setup for disaster recovery, scaling capability



Amazon Aurora

- Aurora is a properietary technology from AWS(not open source)
- PostgreSQL and MySQL are both supported by Aurora DB. You have to write proper queries for creating tables and storing data in it as you do in MySQL.
- It is five times(5x) faster than the actual MySQL and three times(3x) faster than the actual PostgreSQL.
- It can auto-scale up to 128 TB per database instance.
- Aurora costs more than RDS (20% more) but is more efficient
- Not in free tier
- The main advantage is speed, security, and availability which is done by replicating data over three availability zones. It provides the feature of self-healing i.e. it performs the automatic error scanning of your data and the blocks and also provides fault tolerance i.e it provides the ability to continue operating without interruption when their ifs fault in one or more component. It also provides autoscaling, according to your database size it scales out and scales down you don't have to worry about the servers or their charges just have to pay what you are using

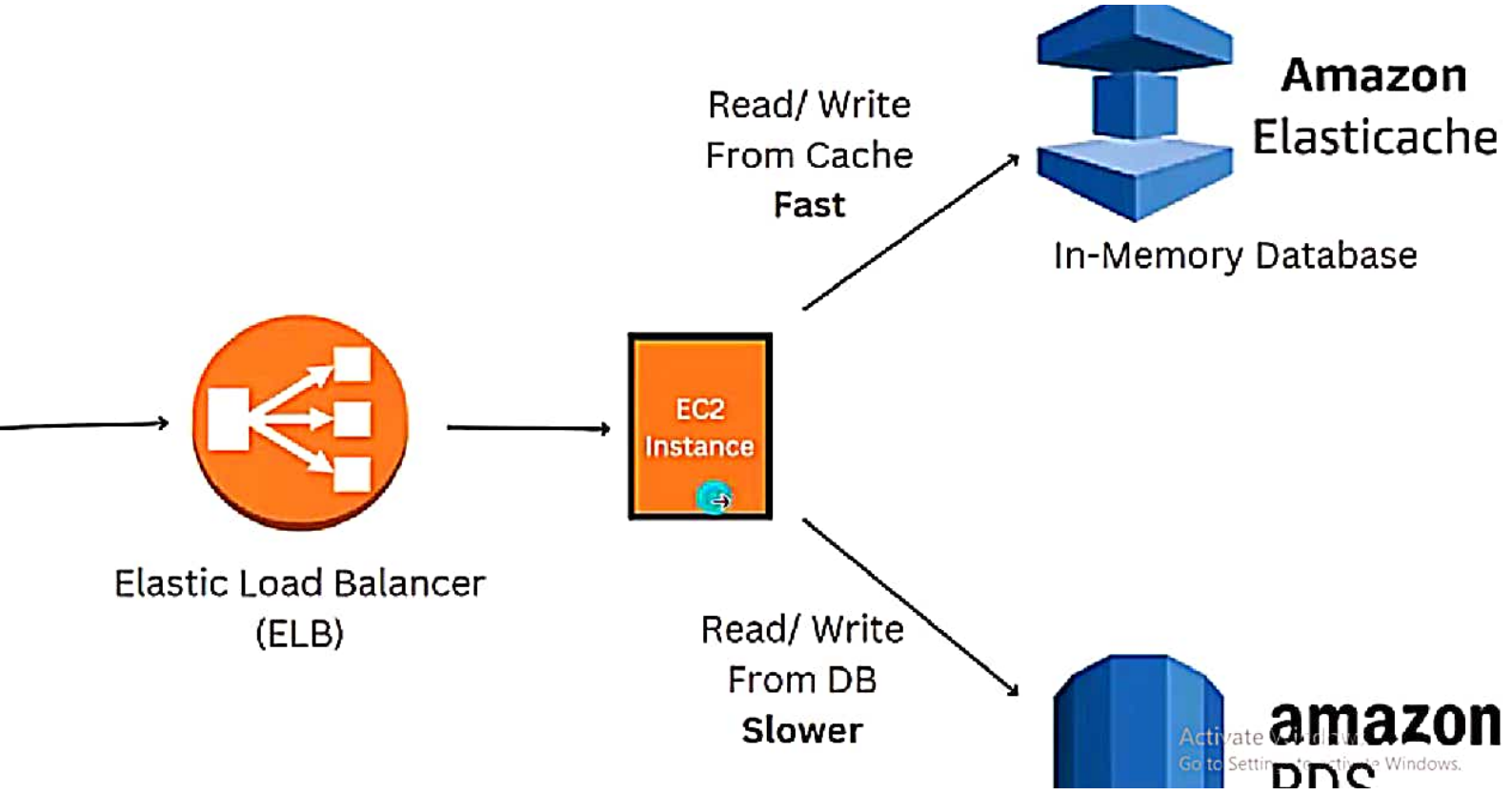


ElasticCache

- Amazon ElastiCache provides two caching engines, namely, Memcached and Redis.
- Cache are in-memory databases with high memory and low latency

Following are some of the main reasons why ElastiCache is immensely useful:

- **Response time:** ElastiCache reduces the response time as it retrieves data from a fast in-memory system. It reduces the dependence on disk-based databases which are usually slower.
- **Scalability**: Amazon ElastiCache is designed to be able to modify itself, automatically. It can scale out or scale up depending on the fluctuating application requirements.
- **Complete management:** Amazon ElastiCache is fully managed, so the common administrative tasks such as hardware provisioning, failure recovery, backups, and more are all automated.


Read/ Write
From Cache
Fast
Amazon
Elasticache
In-Memory Database
EC2
Instance
Elastic Load Balancer
(ELB)
Read/ Write
From DB
Slower
amazon

Dynamo DB- NoSQL Database

- DynamoDB is a key-value Database(Primary Key and Attributes)
- It is fully manages highly available with replication across 3 AZ
- It is a server less database. DynamoDB is a serve rless service that automatically scales up and down to adjust for capacity and maintain performance
- Millions of requests per seconds, trillions of rows, 100s of TB of storage
- Fast and consistent in performance
- Single-digit millisecond latency- low latency retriveal
- low cost and auto-scaling capabilities

DynamoDB Accelerator(DAX)

- Fully managed in-memory cache for DynamoDB
- 10x performance improvement- from millisecond to microsecond latency when accessing DynamoDB tables
- Secure, highly scalable and available
- DAX is only used for and integrated with DynamoDB, while ElasticCache can be used with other databases

Choosing Amazon DynamoDB vs Amazon RDS :

DynamoDB is well-suited for handling large amounts of data with a flexible schema, making it a good choice for applications that need to store and retrieve unstructured data quickly.

On the other hand, is a good choice for traditional relational databases that require structured data and support for SQL queries.

RedShift

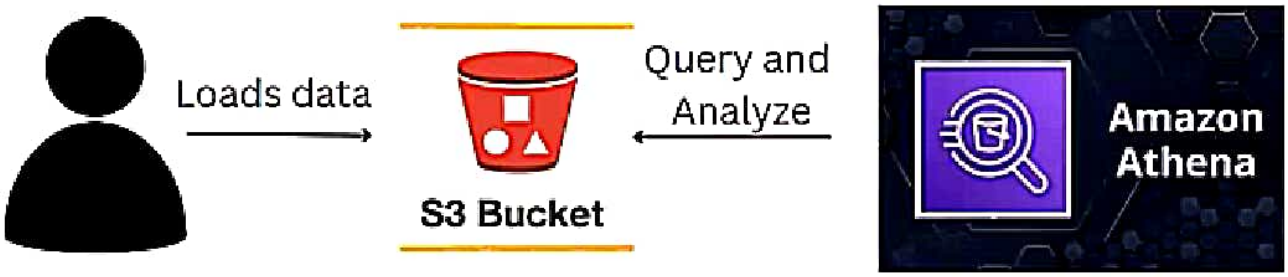
- Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud.
- it's OLAP- **Online Analytics Processing System (analytics and data warehousing)**
- Loads data once every hour
- 10x better performance than other data warehouse
- Columnar storage, instead of rows
- Massive parallel processing
- SQL interface to perform queries

Amazon Timestream

- Amazon timestream is used to handle the time series data that assesses how events change over time.
- The data which changes according to time which is used by the IoT and the other operational applications gather, maintain, and query with the help of Amazon Timestream
- It is a serverless database.
- It can handle the trillions of the request every single day which helps in the reduction of the cost and also it is much faster than the other databases.

AWS Athena

- Serverless query service to perform analytics against S3 objects
- Use SQL to query the files
- Supports CSV, JSON, Parquet files



- **Relational Databases- OLTP:** RDS and Aurora(SQL)
- **In-Memory Database:** ELasticCache
- **Key/Value Database:** DynamoDB(serverless) and DAX(cache for DynamoDB)
- **Warehouse-OLAP:** Redshift
- **Real time data:** Timestream
- **Athena:** query data on Amazon S3(serverless and SQL)
- **QuickSight:** dashboard on your data(serverless)