Coaching - Databases in AWS

Cloud Infrastructure Engineering

**Nanyang Technological University
& Skills Union - 2022/2023**

Course Content

- Explore the different AWS database services
- Explore the different AWS analytics services

AWS DocumentDB

DocumentDB is the same for MongoDB (which is a NoSQL database)

MongoDB is used to **store, query, and index JSON data**

Similar "deployment concepts" as Aurora

Fully Managed, highly available with replication across 3 AZ

DocumentDB storage **automatically grows in increments of 10GB**, up to 64 TB.

Automatically scales to workloads with millions of requests per seconds

AWS DocumentDB

AWS Redshift

Redshift is based on PostgreSQL, but it's not used for OLTP, but OLAP – **online analytical processing (analytics and data warehousing)**

Load data **once every hour,** not every second

10x better performance than other data warehouses, **scale to PBs of data**

Columnar storage of data (instead of row based)

AWS Redshift

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL

101259797|SMITH|88|899 FIRST ST|JUNO|AL 892375862|CHIN|37|16137 MAIN ST|POMONA|CA 318370701|HANDU|12|42 JUNE ST|CHICAGO|IL

Block 1   Block 2   Block 3

AWS Redshift

Massively Parallel Query Execution (MPP), **highly available**

Pay as you go based on the instances provisioned, but **expensive**

SQL interface for **performing the queries from console**

Integration with **BI tools** such as AWS Quicksight or Tableau
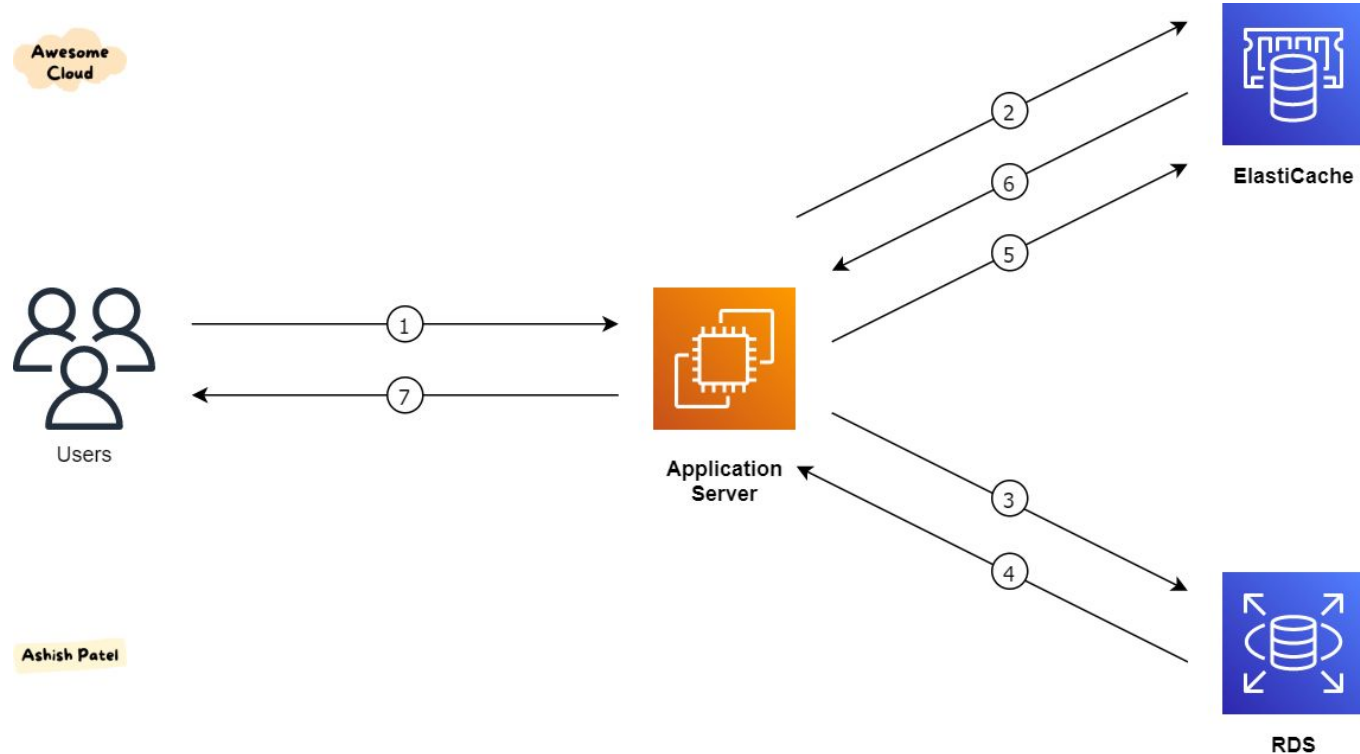
AWS Redshift

Elasticache

ElastiCache can be used for managed Redis or Memcached in AWS

Caches are **in-memory databases with high performance, low latency**

Helps **reduce load off databases for read intensive workloads**

AWS takes care of OS maintenance / patching, optimizations, setup, configuration, monitoring, failure recovery and backups

Elasticache

Memcached vs Redis

	Memcached	**Redis**
Sub-millisecond latency	Yes	Yes
Developer ease of use	Yes	Yes
Data partitioning	Yes	Yes
Support for a broad set of programming languages	Yes	Yes
Advanced data structures	-	Yes
Multithreaded architecture	Yes	-
Snapshots	-	Yes
Replication	-	Yes
Transactions	-	Yes
Pub/Sub	-	Yes
Lua scripting	-	Yes
Geospatial support	-	Yes

Memcached

NANYANG TECHNOLOGICAL UNIVERSITY | SINGAPORE

Redis

Elastic Map Reduce

EMR helps **creating Hadoop clusters (Big Data)** to **analyze and process vast amount of data**
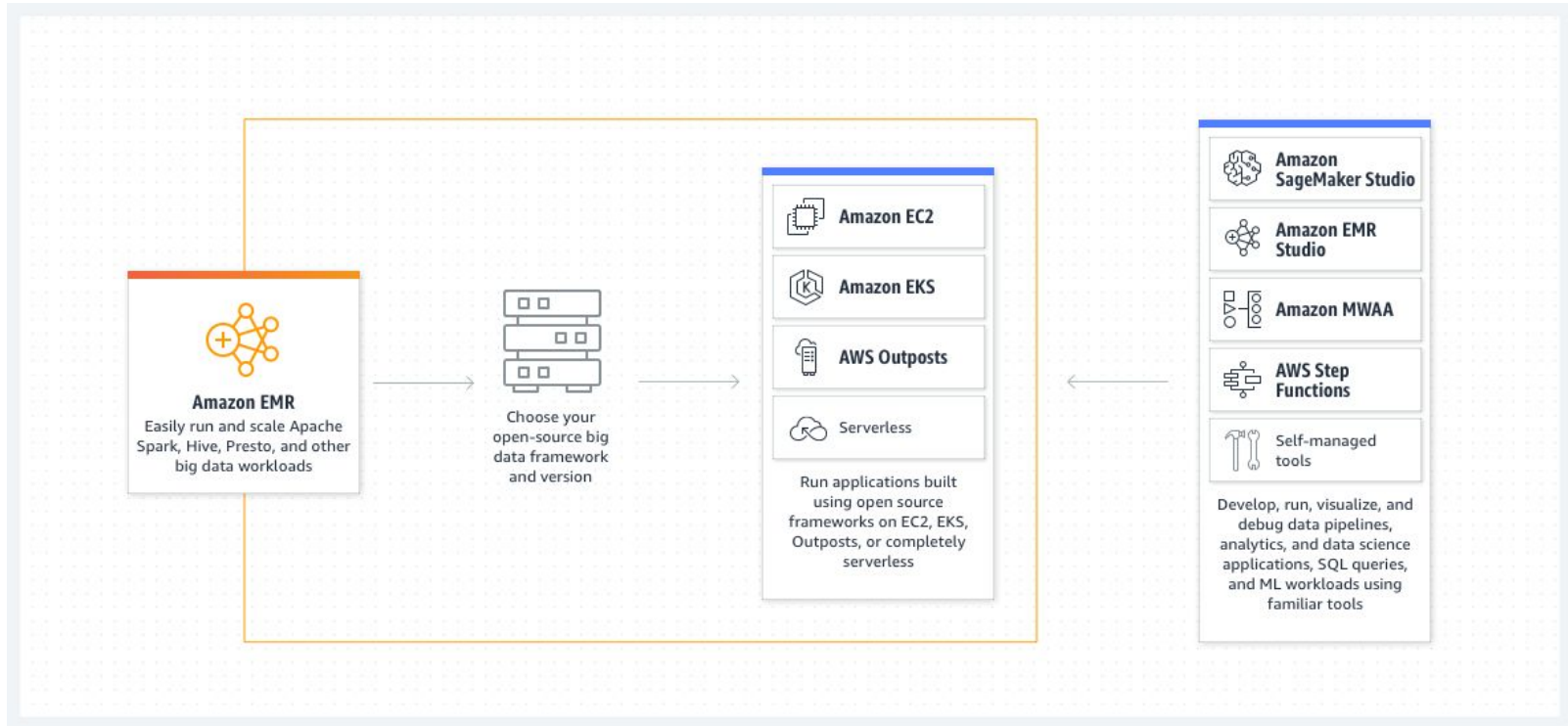
The clusters can be made of hundreds of EC2 instances

Also supports **Apache Spark, HBase, Presto, Flink**

EMR takes care of all the provisioning and configuration

Auto-scaling and integrated with Spot instances

Use cases: data processing, machine learning, web indexing, big data

Elastic Map Reduce

AWS Athena

Serverless **query service** to **analyze data** stored in **Amazon S3**

Uses standard **SQL language** to query the files

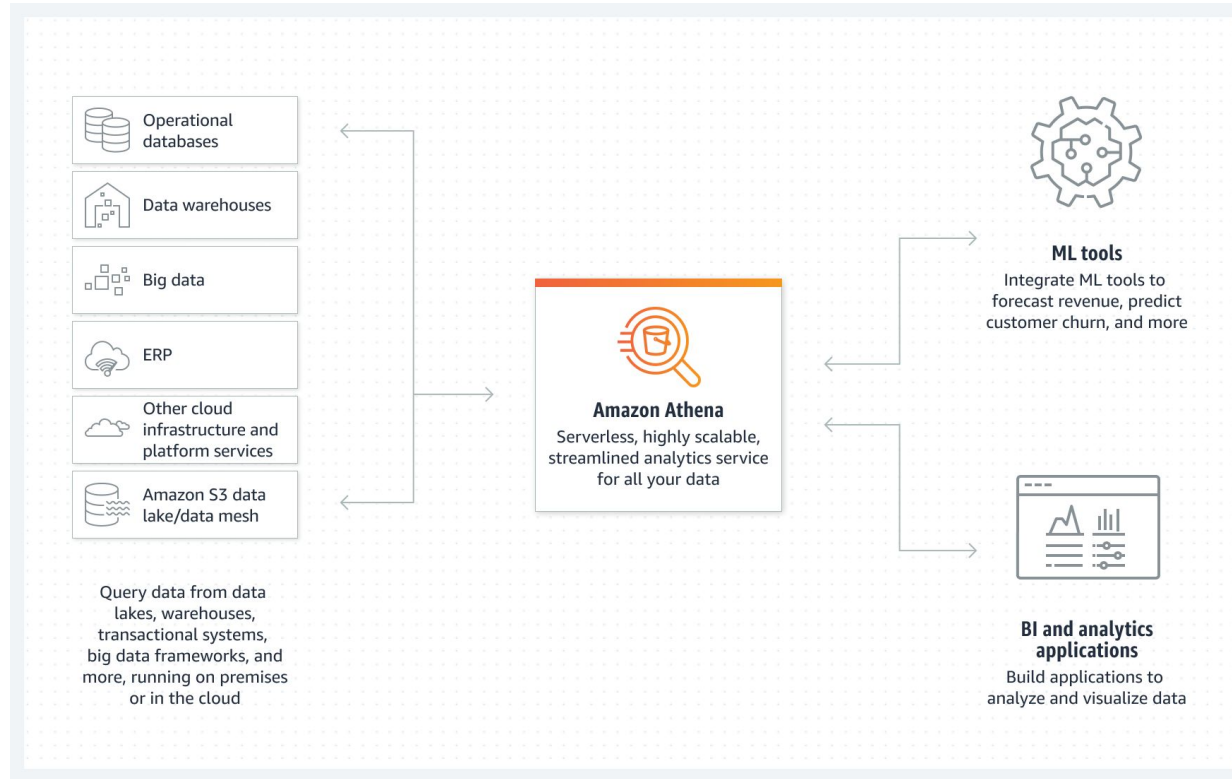Supports CSV, JSON, ORC, Avro & Parquet (built on Presto)

Pricing: $5.00 per TB of data scanned

Use **compressed or columnar data for cost-savings** (less scan)

Use cases: Business intelligence / analytics / reporting, analyze & query VPC Flow Logs, ELB Logs, CloudTrail trails, etc.

**** Exam Tip: if you need to analyze data in S3 using serverless SQL, use Athena

AWS Athena

AWS Athena

AWS QuickSight

Serverless machine learning-powered business intelligence service to create **interactive dashboards**

Fast, automatically scalable, embeddable, with per-session pricing

Use cases:

- Business analytics
- Building visualizations
- Perform ad-hoc analysis
- Get business insights using data

Integrated with RDS, Aurora, Athena, Redshift, S3

AWS QuickSight

AWS Neptune

Fully managed **graph database,** a popular example would be a graph dataset of a social network
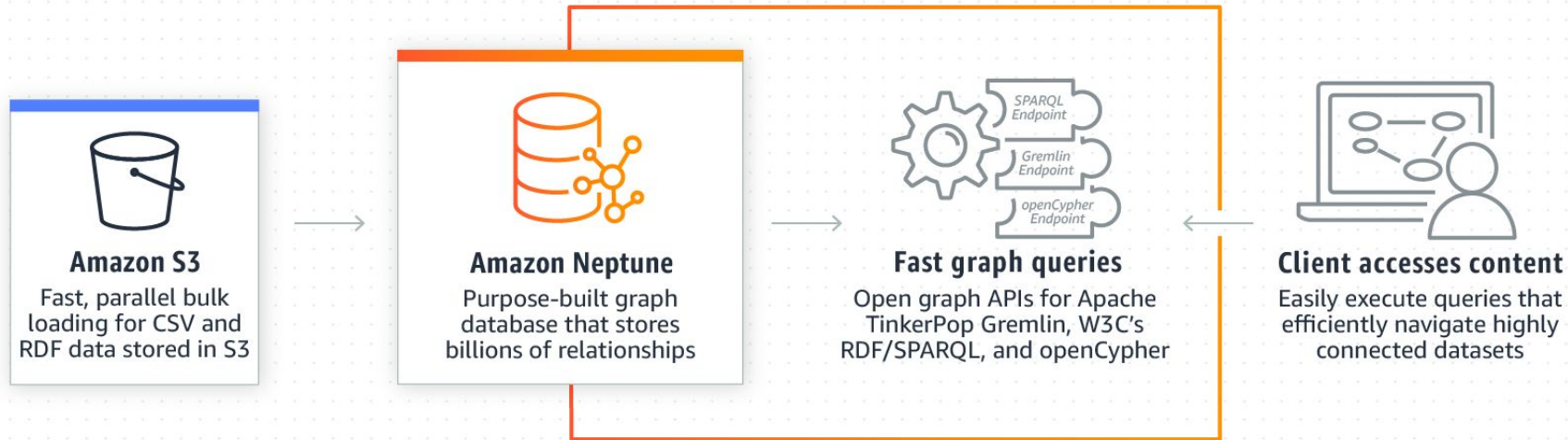
Highly available across 3 AZ, with up to 15 read replicas

Build and run applications working with highly connected datasets – optimized for these complex and hard queries
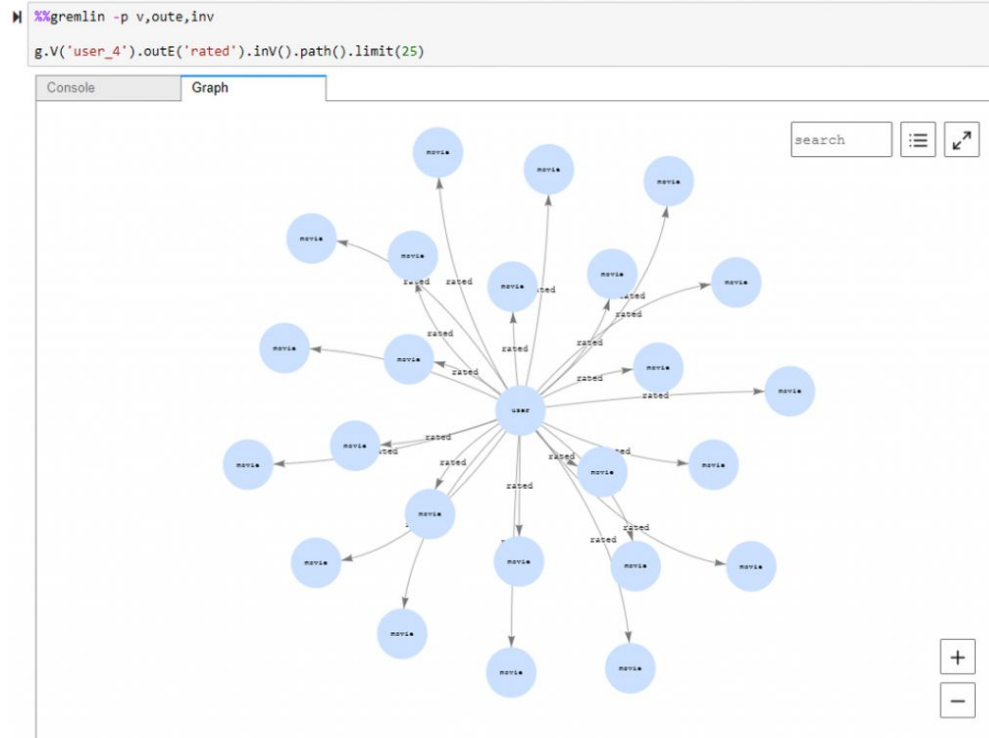
Can store up to billions of relations and query the graph with milliseconds latency

Great for **knowledge graphs (Wikipedia), fraud detection, recommendation engines, social networking**

AWS Neptune

Amazon S3
Fast, parallel bulk loading for CSV and RDF data stored in S3

Amazon Neptune
Purpose-built graph database that stores billions of relationships

Fast graph queries
Open graph APIs for Apache TinkerPop Gremlin, W3C's RDF/SPARQL, and openCypher

SPARQL Endpoint
Gremlin Endpoint
openCypher Endpoint

Client accesses content
Easily execute queries that efficiently navigate highly connected datasets
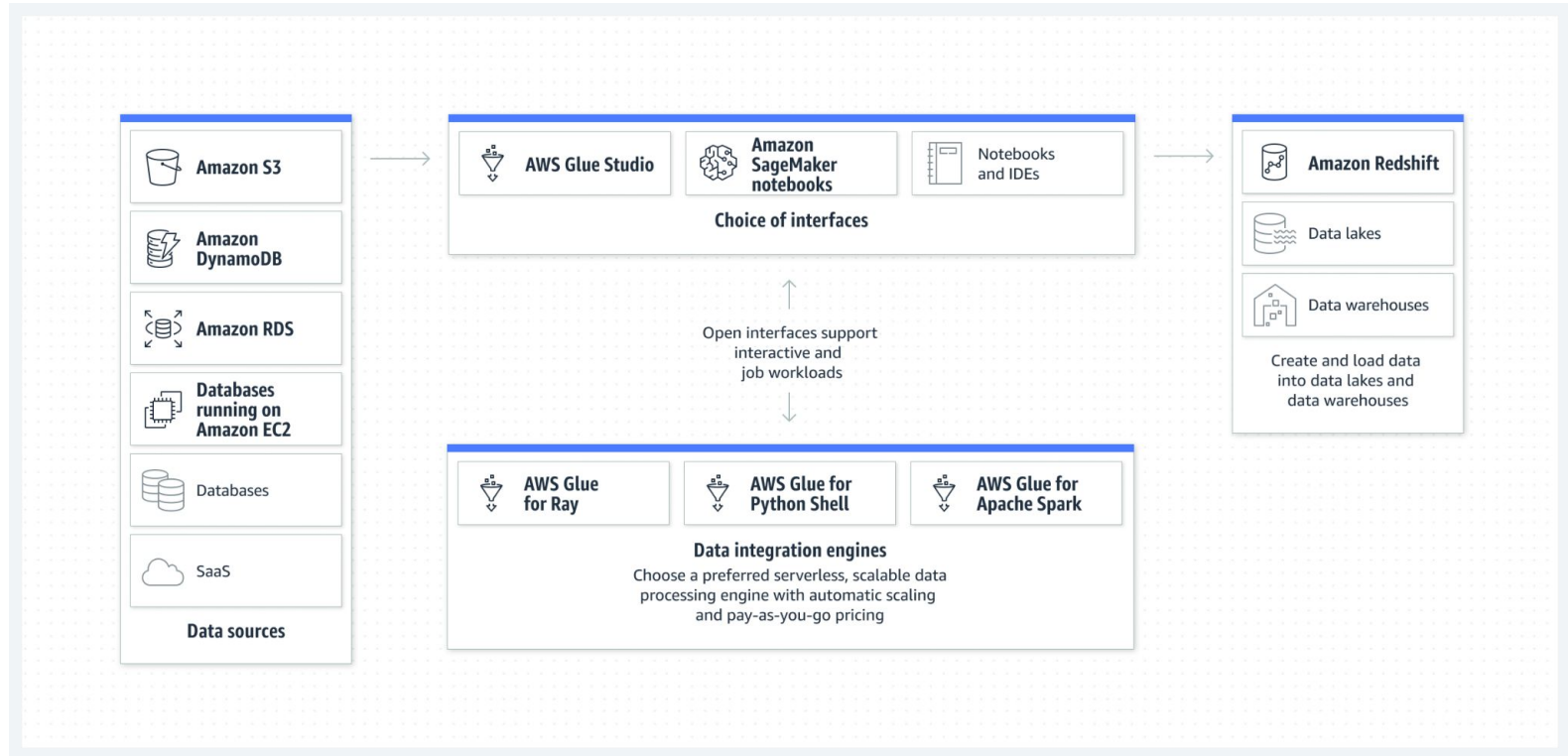
AWS Neptune

AWS Glue

Fully serverless service that manages **extract, transform, and load (ETL) service**

Useful to **prepare and transform data for analytics**

AWS Glue Data Catalog - catalog of datasets that can be used by Athena, Redshift, EMR

AWS Glue Studio - graphical interface that makes it easy to create, run, and monitor ETL jobs in AWS Glue
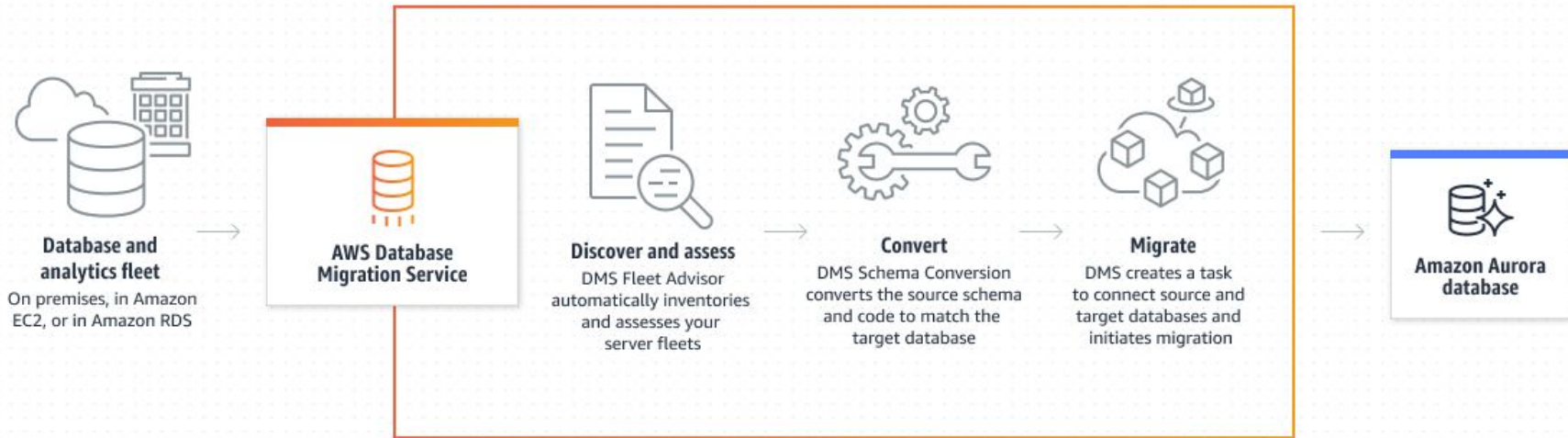
AWS Glue

AWS Database Migration Service

Database migration can be a tough process. With AWS DMS, you can **quickly and securely migrate databases to AWS**, with resilient and self healing capabilities

The source database **remains available during the migration**

Supports:

- Homogeneous migrations: ex Oracle to Oracle
- Heterogeneous migrations: ex Microsoft SQL Server to Aurora

AWS Database Migration Service

What's Next?