Databases and Analytics

Relational Databases - like excel spreadsheets with link b/w them
- use SQL for lookups & queries

NoSQL DB : Non-Relational
- built for specific data models and have flexible schemas for building modern applications.
- flexible, scalable, high performance - optimized for specific data model, highly functional
eg : key-value, document graph, in-memory, search databases
- can be in JSON like.

```
{
    "name" : "John",
    "age" : 30,
    "cars" : [
            "Ford",
            "BMW",
            "Fiat"
        ],
    "address" : {
        type : "House",
        number : "23"
    }
}
```

- Data can be nested
- fields can change over time
- support for new types, arrays etc.

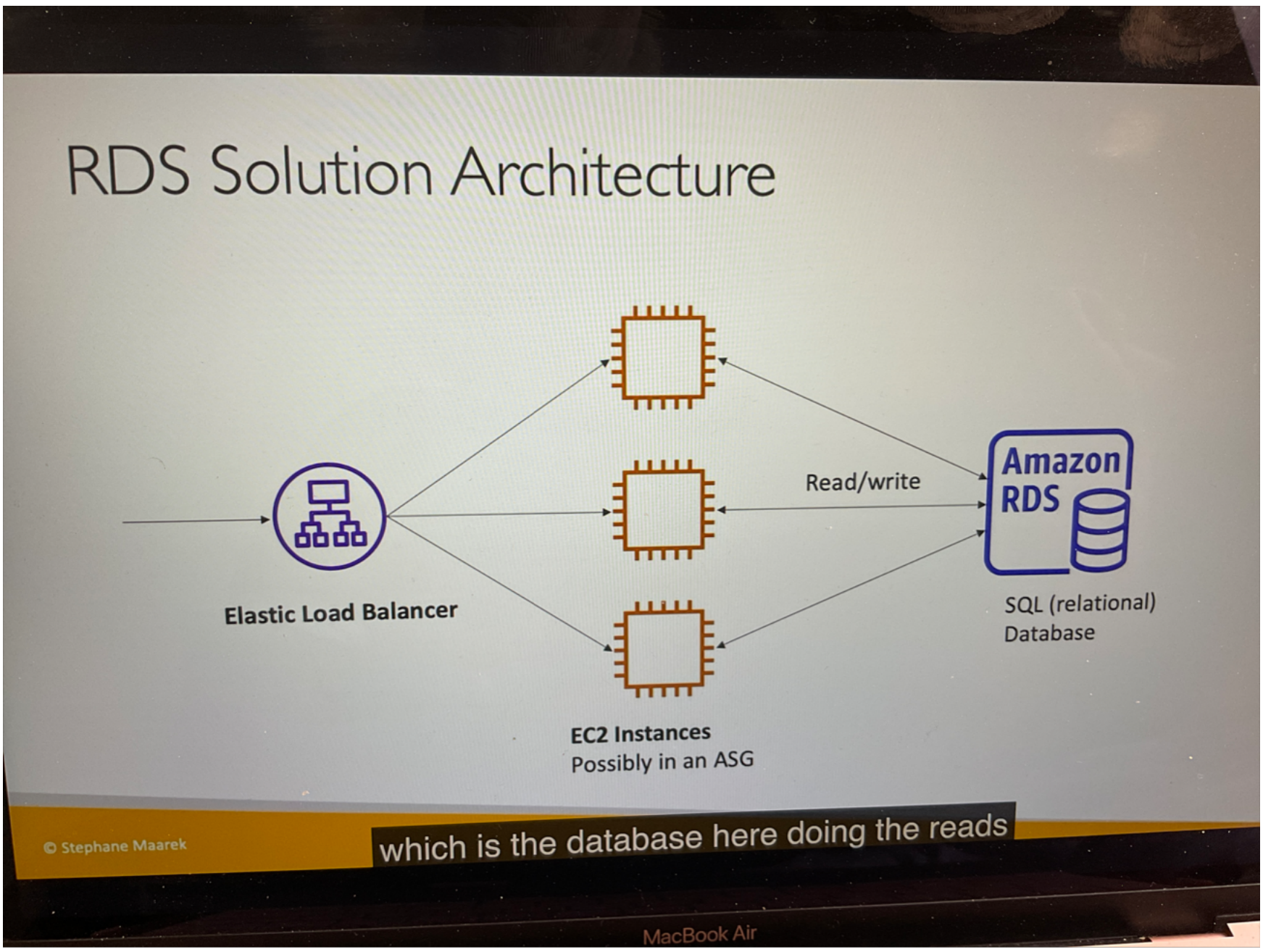
SRM & Databases:
- AWS offers use to manage different databases
Benefits include:
- Quick Provisioning, High availability, vertical and Horizontal Scaling
- Automated Backup, Restore, Operations and Upgrades
- Patching is handled by AWS.
- Monitoring, alerting.

RDS - Relational Database Service
- a managed DB, uses SQL.
- Allows to create dbs in cloud that are managed by AWS.
- Postgres
- MySQL
- MariaDB
- Oracle
- Microsoft SQL server
- Aurora (AWS Proprietary database)

RDS is a managed service: (Advantages).

- Automated provisioning, OS patching.
- Continuous backup and restore to specific timestamp. (Point in Time Restore)!
- Monitoring Dashboards.
- Read replicas for improved read performances.
- Multi AZ setup for DR.
- Maintenance windows for upgrades.
- Scaling capability (Horizontal + vertical)
- Storage backed by EBS (gp2 or io1)

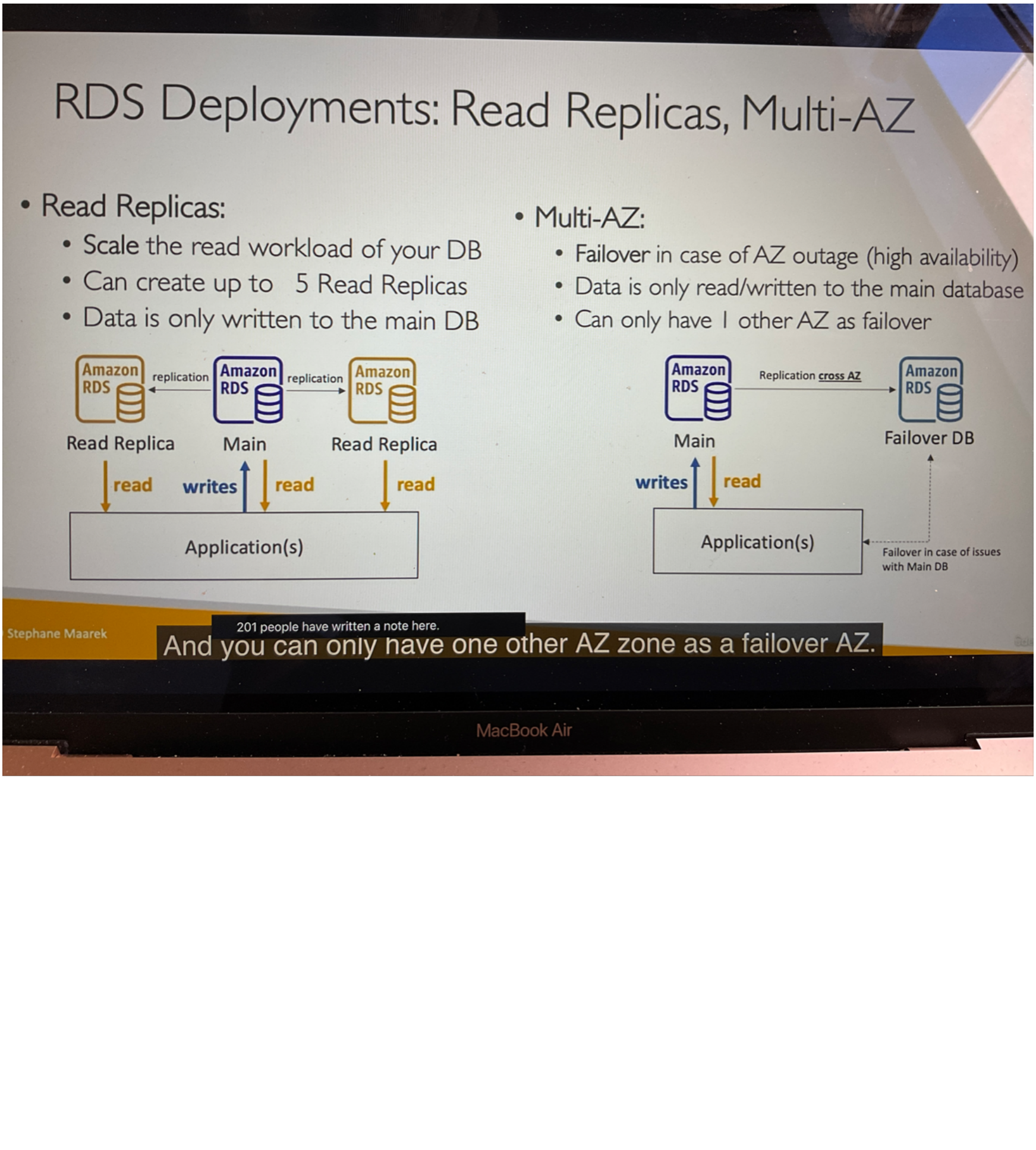
BUT, you cannot SSH into your instances.



Amazon Aurora

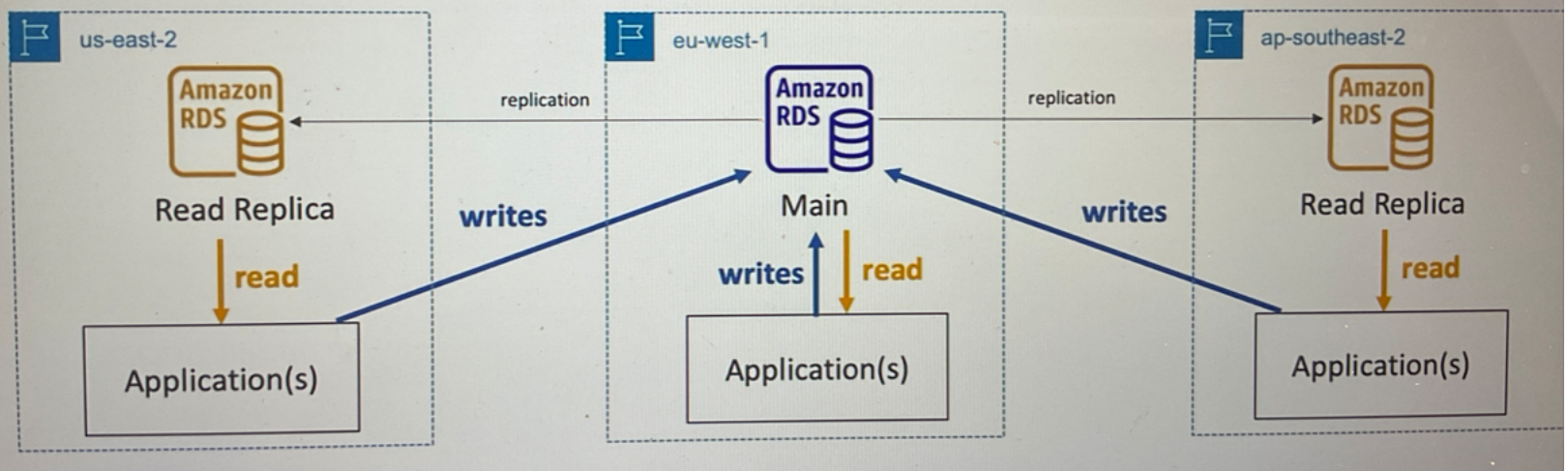
- not open source, more cloud native.
- created by AWS.
- works same as RDS
- not in free tier
- "AWS optimized", claims 5x performance improvement over MySQL on RDS, over 3x for Postgres on RDS.
- Aurora storage automatically grows in increments of 10GB up to 64 TB.
- costs more than RDS (20% more), but more efficient.

RDS Deployment Options:



RDS Deployments: Multi-Region

- Multi-Region (Read Replicas)
  - Disaster recovery in case of region issue
  - Local performance for global reads
  - Replication cost



153 people have written a note here.

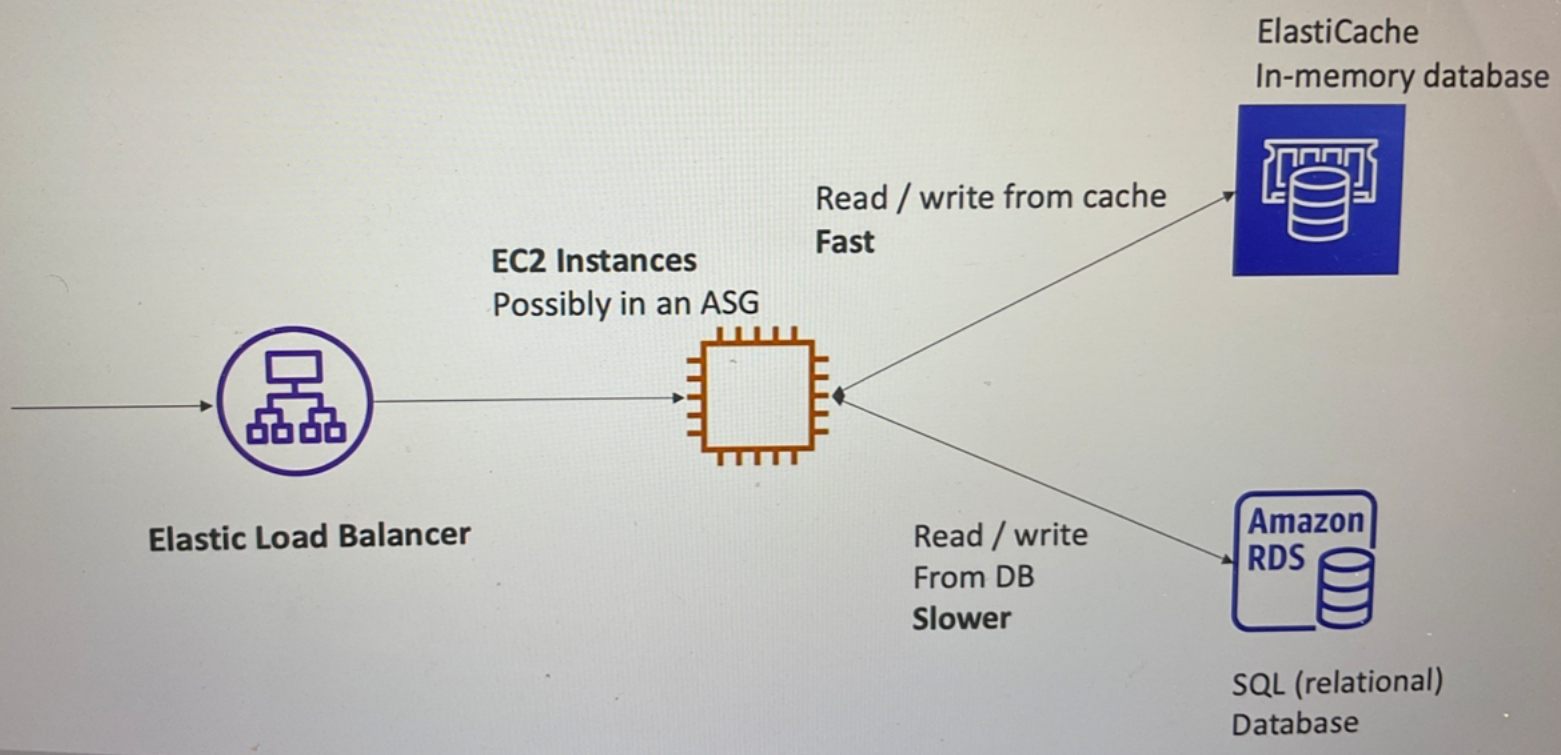
with a network transfers of data between regions.

MacBook Air

Elasticache

- Same as RDS is used to get managed Relational Db..
- Elasticache is to get managed Redis or Memcached
- Caches are in memory db. with high performance, low latency.
- Helps reduce load off db for load intensive workloads
- AWS handles OS maintenance/ patching, optimizations, setup, config, monitoring, failure recovery and backups

ElastiCache
Solution Architecture - Cache



383 people have written a note here.

into the exam and I will see you in the next lecture.

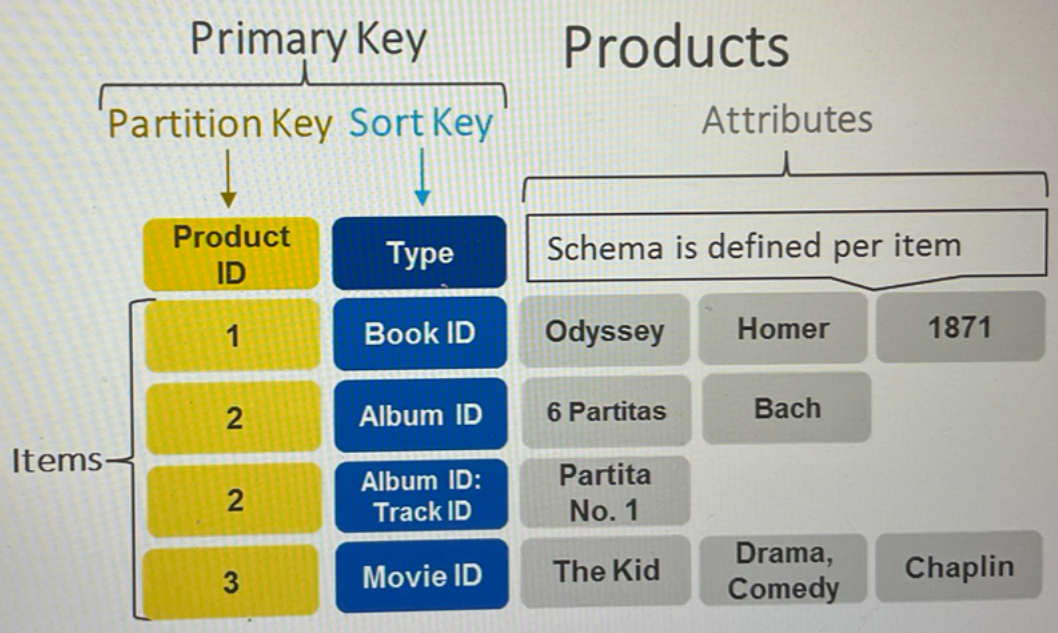
MacBook Air

DynamoDB

- Fully managed, highly available with replication across 3AZ.
- NoSQL DB.
- Scales to massive workloads, distributed "serverless" database
  ⇒ we dont provision any server
  (RDS, Elasticache - need to provision instance)
- millions of requests/sec, billions of rows, 100s of TB of storage.
- fast and consistent in performance.
- single digit millisecond latency - low latency level.
- Integrated with IAM for security, authorization & administration.
- low cost and auto scaling capabilities.
- key value db.

DynamoDB – type of data

- DynamoDB is a key/value database



https://aws.amazon.com/nosql/key

586 people have written a note here.

and you also get access to a serverless database.

DynamoDB Accelerator – DAX.
- fully managed in memory cache for DynamoDB.
- 10x performance improvement => microseconds latency
- secure, highly available & scalable
- only works with DynamoDB, ElastiCache can be used for others too

DynamoDB Global Tables:
- Make a DynamoDB table accessible with low latency in multiple regions.
- Active active replication (read/write to any AWS Region)

RedShift.

Redshift Overview

- Redshift is based on PostgreSQL, but it's not used for OLTP
- It's 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)
- Massively Parallel Query Execution (MPP), highly available
- Pay as you go based on the instances provisioned
- Has a SQL interface for performing the queries
- BI tools such as AWS Quicksight or Tableau integrate with it

I hope you liked it,

EMR - Elastic mapReduce

Amazon EMR

- EMR stands for "Elastic MapReduce"
- 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…

think no more, it's going to be Amazon EMR.

Athena

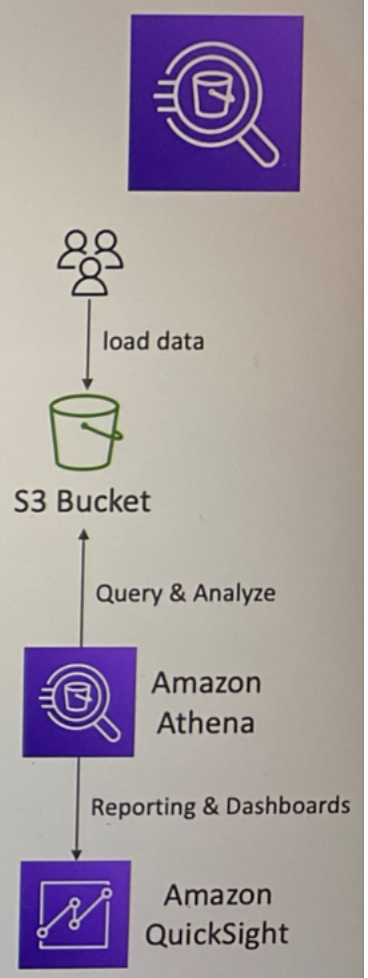
Amazon Athena

- Serverless query service to perform analytics against S3 objects
- Uses standard SQL language to query the files
- Supports CSV, JSON, ORC, Avro, and 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: analyze data in S3 using serverless SQL, use Athena



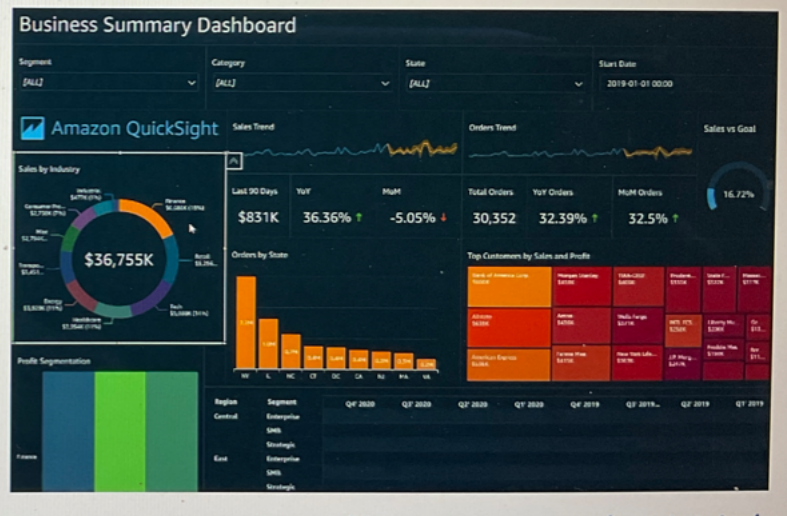
518 people have written a note here.

use SQL, then think Amazon Athena.

MacBook Air

Amazon 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…



https://aws.amazon.com/quicksight/

192 people have written a note here.

So QuickSight is your go-to tool for DI in AWS.

MacBook Air

DocumentDB

- Aurora is an "AWS-implementation" of PostgreSQL / MySQL …
- 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
- Aurora storage automatically grows in increments of 10GB, up to 64 TB.

- Automatically scales to workloads with millions of requests per seconds

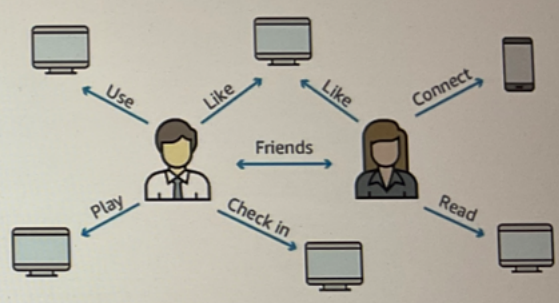
223 people have written a note here.

I hope you liked it,

MacBook Air

Amazon Neptune

- Fully managed **graph** database
- A popular **graph dataset** would be a **social network**
  - Users have friends
  - Posts have comments
  - Comments have likes from users
  - Users share and like posts…
- 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
- Highly available with replications across multiple AZs
- Great for knowledge graphs (Wikipedia), fraud detection, recommendation engines, social networking

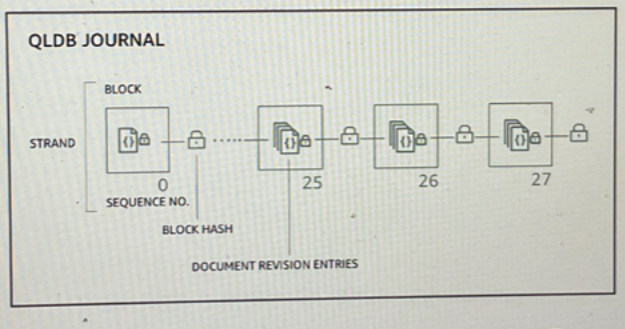


219 people have written a note here.

That's it.

Amazon QLDB

- QLDB stands for "Quantum Ledger Database"
- A ledger is a book **recording financial transactions**
- Fully Managed, Serverless, High available, Replication across 3 AZ
- Used to **review history of all the changes made to your application data** over time
- **Immutable** system: no entry can be removed or modified, cryptographically verifiable



- 2-3x better performance than common ledger blockchain frameworks, manipulate data using SQL
- Difference with Amazon Managed Blockchain: **no decentralization component**, in accordance with financial regulation rules

https://docs.aws.amazon.com/qldb/latest/developerguide/ledger-stru

193 people have written a note here.

© Stephane Maarek

and ledger, think QLDB.

Amazon Managed Blockchain

- Blockchain makes it possible to build applications where multiple parties can execute transactions **without the need for a trusted, central authority.**
- Amazon Managed Blockchain is a managed service to:
  - Join public blockchain networks
  - Or create your own scalable private network
- Compatible with the frameworks Hyperledger Fabric & Ethereum

HYPERLEDGER FABRIC

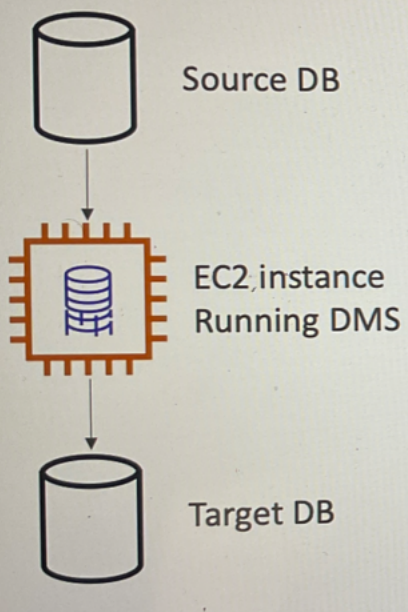
ethereum

222 people have written a note here.

© Stephane Maarek

So that's it, I hope you liked it,

DMS – Database Migration Service



- Quickly and securely migrate databases to AWS, resilient, self healing
- The source database remains available during the migration
- Supports:
  - Homogeneous migrations: ex Oracle to Oracle
  - Heterogeneous migrations: ex Microsoft SQL Server to Aurora

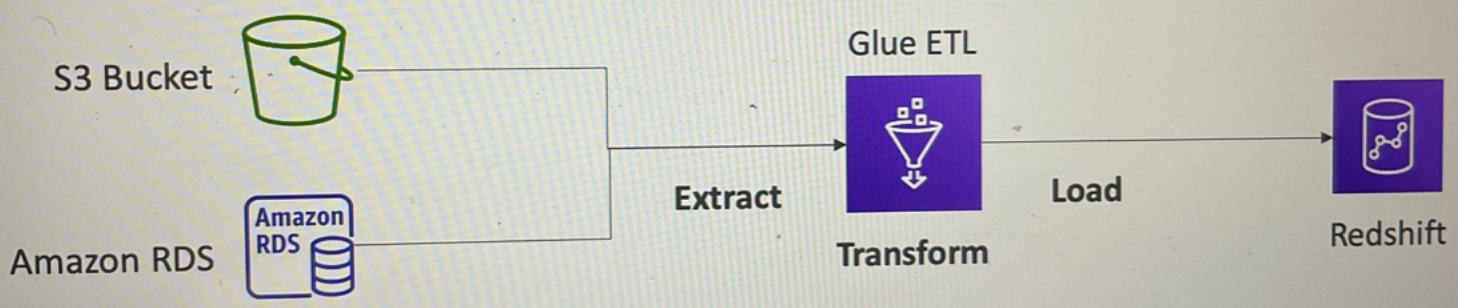
397 people have written a note here.

I hope that was helpful

MacBook Air

AWS Glue

- Managed **extract, transform, and load** (ETL) service
- Useful to prepare and transform data for analytics
- Fully **serverless** service



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

335 people have written a note here.

and build the proper schemas for them, okay?

Databases & Analytics Summary in AWS

- Relational Databases - OLTP: RDS & Aurora (SQL)
- Differences between Multi-AZ, Read Replicas, Multi-Region
- In-memory Database: ElastiCache
- Key/Value Database: DynamoDB (serverless) & DAX (cache for DynamoDB)
- Warehouse - OLAP: Redshift (SQL)
- Hadoop Cluster: EMR
- Athena: query data on Amazon S3 (serverless & SQL)
- QuickSight: dashboards on your data (serverless)
- DocumentDB: "Aurora for MongoDB" (JSON – NoSQL database)
- Amazon QLDB: Financial Transactions Ledger (immutable journal, cryptographically verifiable)
- Amazon Managed Blockchain: managed Hyperledger Fabric & Ethereum blockchains
- Glue: Managed ETL (Extract Transform Load) and Data Catalog service
- Database Migration: DMS
- Neptune: graph database