Databases in AWS

Uday Manchanda

September 13, 2021

1 Choosing the right database

- Based on architecture
  - Read-heavy, write-heavy, or balanced workload?
  - Throughput needs
  - How much data to store and for how long? Will it grow?
  - Data durability?
  - Latency rqts
  - Schema type?
- Database types
  - RDBMS (SQL/OLTP): RDS, Aurora - great for joins
  - NoSQL: DynamoDB (JSON), ElastiCache (k/v pairs), Neptune (graph) - no SQL
  - Object store: S3 for big objects, Glacier for backups/archives
  - Data Warehouse (SQL Analytics): Redshift, Athena
  - Search: Elastisearch - free text, unstructured searches
  - Graphs: Neptune - displays relationships between data

2 RDS

- Managed SQL/NoSQL server
- Must provision EC2 and EBS behind the scenes
- Security thru IAM/SGs/KMS
- Backup and snapshots
- Monitoring thru Cloudwatch

- Use case: store relational datasets
- Multi AZ feature

3 Aurora

- Auto healing capability
- Can be global for DR/latency purposes
- Define EC2 instance for an instance
- Aurora serverless for unpredictable/intermittent workloads
- Like RDS with less maintence, more flexibility

4 ElastiCache

- Managed Redis/Memcached
- In-memory data store, sub-milisecond latency
- Must provision an EC2 instance
- Support for Clustering and sharding
- Maximum amount of replication and auto scaling capability
- Use cases: key/value store, frequent reads/less writes, store session data for websites

5 DynamoDB

- Managed NoSQL db
- Serverless provisioned capacity, auto scaling
- Could replace ElastiCache as a k/v store
- reads can be eventually or strongly consistent
- can only query on primary/sort key or indexes

6 S3

- A key/value store for objects
- Great for big objects
- Serverless, scales infinitely, max object size is 5 TB

7 Athena

- Provides a query engine on top of S3 with SQL capabilities
- Query data in S3, output results back to S3
- use Cases: one time SQL queries, serverless queries
- Uses presto

8 Redshift

- Based on postgres, not used for OLTP
- OLAP - online analytical processing (analytics and data warehousing)
- Columnar storage of data, massive parallel query execution
- Data is loaded from S3/DynamoDB
- Leader node, compute node
- No multi-AZ mode, snapshots are stored in S3, can restore a snapshot into a new cluster
- Loading data into Redshift
  - Amazon Kinesis, Data Firehouse
  - S3 using COPY command
  - EC2 instance via JDBC driver
- Redshift spectrum: query data that is already in S3 without loading it. Query is submitted to thousands of redshift spectrum nodes
- Faster than Athena because of indexes

9 Glue

- Managed ETL service, useful to prepare and transform data for analytics, fully serverless
- Glue Data Catalog - catalog of datasets
- Glue Data Crawler will crawl thru your datasets, writes metadata to the catalog, and then used by glue jobs

10 Neptune

- Fully managed graph database
- High relationship data (EX: social networking or knowledge graphs)

11 ElasticSearch

- EX: in DynamoDB you can only find data via primary key or indexes
- You can search any field, even partial matches
- good compliment to DynamoDB or other DBs
- Provision a cluster of instances and has built in integrations to other services
- ELK Stack: ElasticSearch, Kibana, and Logstash