AMAZON AWS SOLUTIONS ARCHITECT

RDBMS OVERVIEW

OVERVIEW

WHAT IS AN RDBMS?

- RDBMS is a Relational Database Management System
  - Traditional Rows and Tables
  - Accessed using SQL (Structured Query Language)
  - A **column** will contain a field
  - A **row** will contain the values in the fields, for a particular record
- Types of RDBMS
  - Database
    - Few number of frequently accessed tables where data is read and written to
    - Data is from one source
    - Updated in Real-time
  - Data Warehouse
    - Large number of infrequently read tables, where data is only read from
    - Used primarily for reporting (known as **O**n**l**ine **A**nalytical **P**rocessing or OLAP)
    - Data comes from more than one source
    - Updated in batch jobs, typically daily or hourly

REDSHIFT

- Fast, powerful, fully managed data warehouse in the cloud
- Relational database designed for OLAP and High-performance analysis
- Fast query using SQL commands to support querying large datasets
- Integrates well with ODBC, JDBC for Data Loading, Reporting and Analytics
- Based on Industry-Standard PostgreSQL

REDSHIFT

- An Amazon Redshift data warehouse is a collection of computing resources called *nodes*, which are organized into a group called a *cluster*
- Each cluster runs an Amazon Redshift engine and contains one or more databases
- An Amazon Redshift cluster consists of nodes.
  - Each cluster has a leader node and one or more compute nodes.
  - The *leader node* receives queries from client applications, parses the queries, and develops query execution plans. The leader node then coordinates the parallel execution of these plans with the *compute* nodes and aggregates the intermediate results from these nodes. It then finally returns the results back to the client applications.
  - *Compute nodes* execute the query execution plans and transmit data among themselves to serve these queries. The intermediate results are sent to the leader node for aggregation before being sent back to the client applications.

REDSHIFT

- When you launch a cluster, one option you specify is the node type.
- The node type determines the CPU, RAM, storage capacity, and storage drive type for each node.
    - The *dense storage* (DS) node types are storage optimized.
    - The *dense compute* (DC) node types are compute optimized.
- DS2 node types are optimized for large data workloads and use hard disk drive (HDD) storage
- DC1 and DC2 nodes are optimized for performance-intensive workloads. Because they use solid state drive (SSD) storage, DC1 and DC2 node types deliver much faster I/O compared to DS node types, but provide less storage space.
- *Slices per Node* is the number of slices into which a compute node is partitioned.
- *Storage* is the capacity and type of storage for each node.

REDSHIFT COMPUTE NODES

Dense Storage Node Size	vCPU	ECU	RAM (GiB)	Slices Per Node	Storage Per Node	Node Range	Total Capacity
ds2.xlarge	4	13	31	2	2 TB HDD	1–32	64 TB
ds2.8xlarge	36	119	244	16	16 TB HDD	2–128	2 PB

Dense Compute Node Size	vCPU	ECU	RAM (GiB)	Slices Per Node	Storage Per Node	Node Range	Total Capacity
dc1.large	2	7	15	2	160 GB SSD	1–32	5.12 TB
dc1.8xlarge	32	104	244	32	2.56 TB SSD	2–128	326 TB
dc2.large	2	7	15.25	2	160 GB NVMe-SSD	1–32	5.12 TB
dc2.8xlarge	32	99	244	16	2.56 TB NVMe-SSD	2–128	326 TB

REDSHIFT TABLE DESIGN

- Data Distribution
  - **Even distribution:** The leader node distributes the rows across the slices in a round-robin fashion, regardless of the values in any particular column. EVEN distribution is appropriate when a table does not participate in joins or when there is not a clear choice between KEY distribution and ALL distribution. EVEN distribution is the default distribution style.
  - **Key distribution:** The rows are distributed according to the values in one column. The leader node will attempt to place matching values on the same node slice. If you distribute a pair of tables on the joining keys, the leader node collocates the rows on the slices according to the values in the joining columns so that matching values from the common columns are physically stored together.
  - **All distribution:** A copy of the entire table is distributed to every node. Where EVEN distribution or KEY distribution place only a portion of a table's rows on each node, ALL distribution ensures that every row is collocated for every join that the table participates in. ALL distribution multiplies the storage required by the number of nodes in the cluster, and so it takes much longer to load, update, or insert data into multiple tables. ALL distribution is appropriate only for relatively slow moving tables; that is, tables that are not updated frequently or extensively.

REDSHIFT SNAPSHOTS

- Snapshots are point-in-time backups of a cluster.
- There are two types of snapshots: *automated* and *manual*.
- Amazon Redshift stores these snapshots internally in Amazon S3 by using an encrypted Secure Sockets Layer (SSL) connection
- If you need to restore from a snapshot, Amazon Redshift creates a new cluster and imports data from the snapshot that you specify
- Amazon Redshift periodically takes incremental snapshots that track changes to the cluster since the previous snapshot
- Amazon Redshift provides free storage for snapshots that is equal to the storage capacity of your cluster until you delete the cluster

REDSHIFT SNAPSHOTS

- Automated Snapshots
  - Amazon Redshift periodically takes snapshots of that cluster, usually every eight hours or following every 5 GB per node of data changes, or whichever comes first
  - Enabled by Default
  - These snapshots are deleted at the end of a retention period
  - The default retention period is one day, but you can modify it by using the Amazon Redshift console or programmatically by using the Amazon Redshift API
  - Only Amazon Redshift can delete an automated snapshot; you cannot delete them manually.
  - Amazon Redshift deletes automated snapshots at the end of a snapshot’s retention period, when you disable automated snapshots, or when you delete the cluster
  - If you want to keep an automated snapshot for a longer period, you can create a copy of it as a manual snapshot
    - The automated snapshot is retained until the end of retention period, but the corresponding manual snapshot is retained until you manually delete it

REDSHIFT SNAPSHOTS

- Manual Snapshots
  - Regardless of whether you enable automated snapshots, you can take a manual snapshot whenever you want
  - Amazon Redshift will never automatically delete a manual snapshot. Manual snapshots are retained even after you delete your cluster.
  - Because manual snapshots accrue storage charges, it’s important that you manually delete them if you no longer need them.
    - If you delete a manual snapshot, you cannot start any new operations that reference that snapshot. However, if a restore operation is in progress, that restore operation will run to completion.
  - Amazon Redshift has a quota that limits the total number of manual snapshots that you can create; this quota is per AWS account per region

RELATIONAL DATABASE SERVICE (RDS)

- Service provided by AWS for your data needs
- Installation and provisioning of hardware and resources handled by AWS
- Supports open and commercial database engines
  - MySQL
  - PostgreSQL
  - MariaDB
  - M$SQL
  - Oracle
- Controlled by API or connecting to the instance
- Supports Cross Region Replication and Multi Availability Zone
- 99.95% SLA
- Read Replicas available for PostgreSQL, MaraiaDB, MySQL, Aurora

AWS DATABASE OFFERINGS

- EC2 Instances
  - Install Database Product on your own EC2 Instance
  - Self Managed
  - No Replication unless created by user
  - No Backups unless created by the user
  - Not Great for the cloud

HANDS-ON DEMO

- Amazon Aurora

COMPARISON OF RESPONSIBILITIES

Responsibility	Database On-Premise	Database on EC2	Database on RDS
App Optimization	You	You	You
Scaling	You	You	AWS
High Availability	You	You	AWS
Backups	You	You	AWS
DB Engine Patches	You	You	AWS
Software Installations	You	You	AWS
OS Patches	You	You	AWS
OS Installation	You	AWS	AWS
Server Maintenance	You	AWS	AWS
Rack and Stack	You	AWS	AWS
Power and Cooling	You	AWS	AWS

QUESTIONS?