Amazon Redshift

Fully Managed Petabyte-size Data Warehouse.
Analyze (Run complex SQL queries) on massive amounts of data
Columnar Store database.

What is a Data Warehouse?

What is a Database Transaction?
A transaction symbolizes a unit of work performed within a database management system
eg. reads and writes

Database

Online **Transaction** Processing (OLTP)

A database was built to store current transactions and enable **fast access to specific transactions** for ongoing business processes

Adding Items To Your Shopping List

Single Source

short transactions (small and simple queries) with an emphasis on writes.

VS

Data Warehouse

Online **Analytical** Processing (OLAP)

A data warehouse is built to store large quantities of historical data and **enable fast, complex queries across all the data**

Generating Reports

Multiple Sources

Long transactions (long and complex queries) with an emphasis on reads.

Introduction of Redshift

AWS Redshift is the AWS managed, petabyte-scale solution for **Data Warehousing**.

Pricing starts at just $0.25 per hour with no upfront costs or commitments.
Scale up to petabytes for $1000 per terabyte, per year.
Redshift is price is less than 1/10 cost of most similar services.

Redshift is used for Business Intelligence.

Redshift uses OLAP (Online Analytics Processing System)

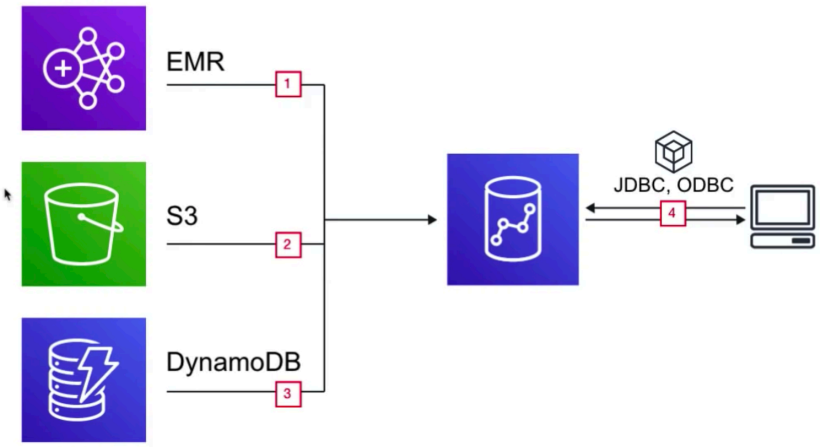
Redshift is **Columnar Storage** Database

Columnar storage for database tables is an important factor in optimizing analytic query performance because it drastically reduces the overall disk I/O requirements and reduces the amount of data you need to load from disk.

Redshift – Use Case

We want to continuously COPY data from
1. EMR,
2. S3 and
3. DynamoDB
to power a custom Business Intelligence tool.

Using a third-party library we can connect and query Redshift for data.



Redshift – Columnar Storage

Columnar Storage stores data together as columns instead of rows.

Name	Rank	Species
John-Luc Picard	Captain	Human
Worf	Lieutenant	Klingon
Data	Lieutenant commander	Android

John-Luc Picard | Worf | Data

BLOCK 1

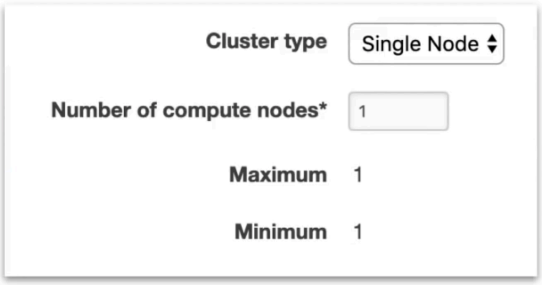
OLAP applications look at multiple records at the same time. You save memory because you fetch just the columns of data you need instead of whole rows

Since data is stored via column, that means all data is of the same data-type allowing for easy compression.

Redshift – Configurations

Single Node

Nodes come in sizes of **160 GB**. You can launch a single node to get started with Redshift.



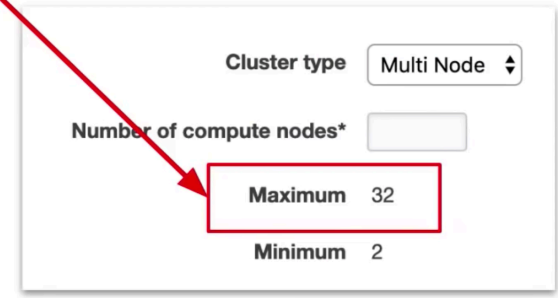
Multi-Node

You can launch a cluster of nodes with Multi Node mode

Leader Node - manages client connections and receiving queries

Compute Node stores data and performs queries up to **128 compute nodes**

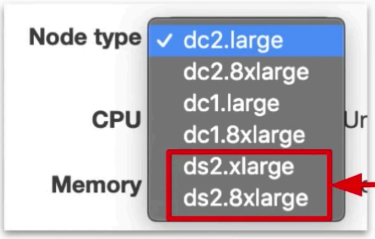
Ask for AWS service limit increase



Redshift – Node Types and Sizes

There are ✌️ two types of Nodes

The smallest node you can select is **dc2.large**



Dense Compute (dc)
best for high performance, but they have less storage

Dense Storage (ds)
clusters in which you have a lot of data

Redshift – Compression

Redshift uses **multiple compression techniques** to achieve significant compression relative to traditional relational data stores.

Similar data is **stored sequentially** on disk.

Does **not** require **indexes** or **materialized views**, which saves a lot of space **compared to traditional systems**.

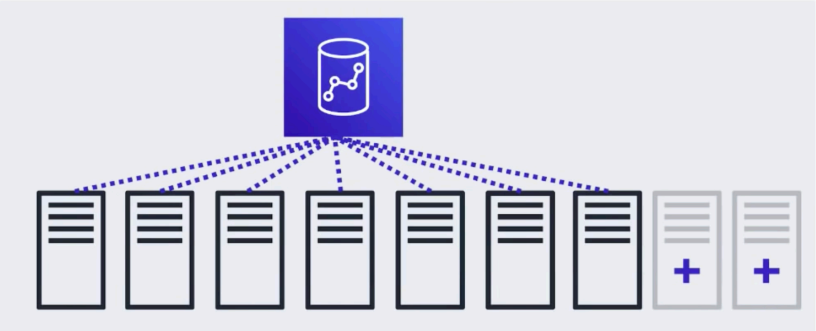
When loading data to an **empty table**, data is **sampled** and the most appropriate compression **scheme is selected automatically**.

Redshift – Processing

Redshift uses **Massively Parallel Processing (MPP).**

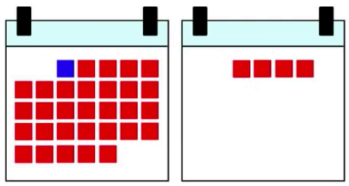
Automatically distributes data and query loads **across all nodes**.

Lets you easily **add new nodes** to your data warehouse while still **maintaining fast query performance**.



Redshift – Backups

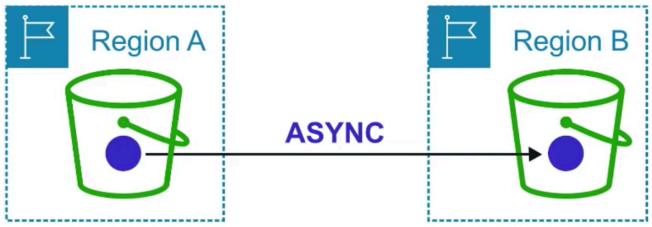
Backups are **enabled by default** with a **1 day** retention period. Retention period can be modified **up to 35 days**.



Redshift always attempts to maintain at least **3 copies of your data**.

1. The original copy
2. Replica on the compute nodes
3. Backup copy in S3

Can asynchronously replicate your snapshots to S3 **in a different region**



Redshift – Billing

Compute Node Hours

- The total number of hours ran across all nodes in the billing period
- Billed for 1 unit per node, per hour.
- **Not charged for leader node hours**, only compute notes incur charges

Backup

- Backups are stored on S3 and you are billed the S3 storage fees

Data Transfer

- Billed for Only transfers within a VPN, not outside of it

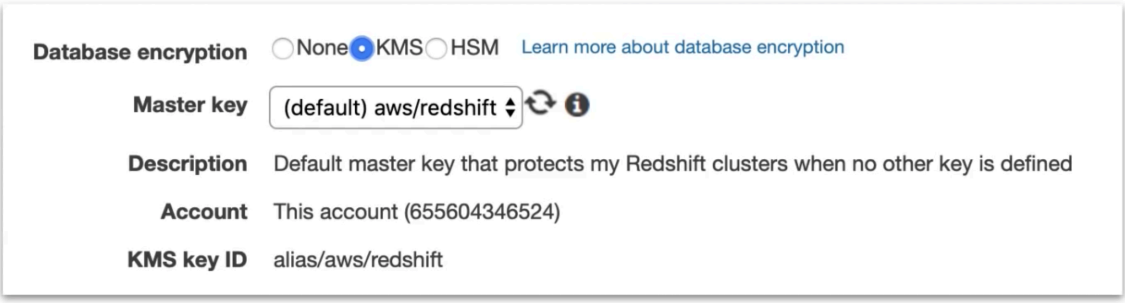
Redshift – Security

Data-in-transit - Encrypted using SSL

Data-at-rest - Encrypted using AES-256 encryption

Database Encryption can be applied using

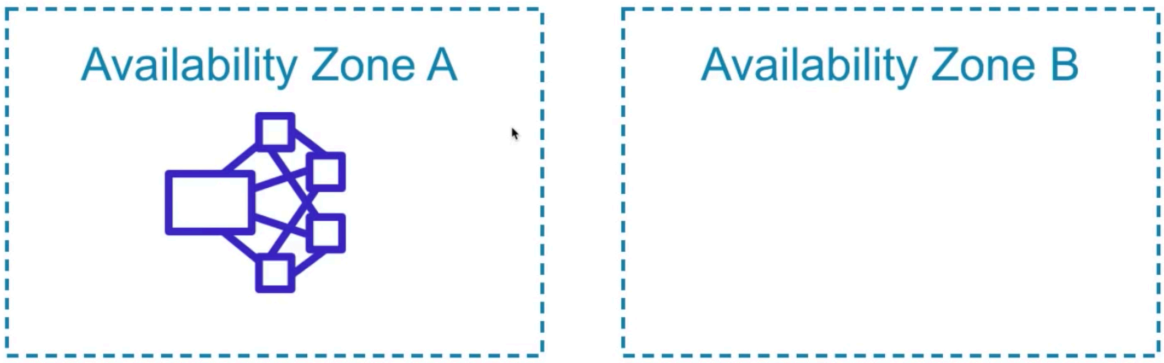
- Key Management Service (KMS) multi-tenant HSM
- CloudHSM single-tenant HSM



Redshift – Availability

Redshift is **Single-AZ**. To run in Multi-AZ you would have to run multiple RedShift Clusters in different AZs with same inputs.

Snapshots **can be restored to a different AZ** in the event an outage occurs



Redshift *CheatSheet*

- Data can be loaded from S3, EMR, DynamoDB, or multiple data sources on remote hosts.
- Redshift is Columnar Store database which can SQL-like queries and is an OLAP.
- Redshift can handle petabytes worth of data. Redshift is for Data Warehousing
- Redshift most common use case is Business Intelligence
- Redshift can only run in a 1 availability zone (**Single-AZ**)
- Reshift can run via a single node or multi-node (clusters)
- A single node is 160 GB in size
- A multi-node is comprised of a leader node and multiple compute nodes
- You are bill per hour for each node (excluding leader node in multi-node)
- You are not billed for the leader node
- You can have up to 128 compute nodes
- Redshift has two kinds of Node Type **Dense Compute** and **Dense Storage**
- Redshift attempts to backup 3 copies of your data, the original, on compute node and on S3
- Similar data is stored on disk sequentially for faster reads
- Redshift database can be encrypted via KMS or CloudHSM
- Backup Retention is default to 1 day and can be increase to maximum of 35 days
- Reshift can asynchronously back up your snapshot to Another Region delivered to S3
- Redshift uses Massively Parallel Processing (MPP) to distribute queries and data across all loads
- In the case of empty table, when importing Redshift will sample data to create a schema.