**Amazon Redshift**

Amazon Redshift is a fast, powerful, fully managed, petabyte-scale data warehouse service in the cloud.

Amazon Redshift is a relational database designed for OLAP scenarios and optimized for high-performance analysis and reporting of very large datasets  
based on industry-standard PostgreSQL, so most existing SQL client applications will work with only minimal changes.

Amazon Redshift manages the work needed to set up, operate, and scale a data warehouse, from provisioning the infrastructure capacity to automating ongoing administrative tasks such as backups and patching. Amazon Redshift automatically monitors your nodes and drives to help you recover from failures.

**Clusters and Nodes**

The key component of an Amazon Redshift data warehouse is a cluster. A cluster is composed of a leader node and one or more compute nodes. The client application interacts directly only with the leader node, and the compute nodes are transparent to external applications.

Cluster

* Leader node
* One or more compute nodes

six different node types and each has a different mix of CPU, memory, and storage

Grouped into 2:

Dense compute - The Dense Compute node types support clusters up to 326TB using fast SSD

Dense storage - Dense Storage nodes support clusters up to 2PB using large magnetic disks

Client  JDBC/ODBC  Leader  Compute

Each cluster contains one or more databases. User data for each table is distributed across the compute nodes.

The disk storage for a compute node is divided into a number of slices. The number of slices per node depends on the node size of the cluster and typically varies between 2 and 16. The nodes all participate in parallel query execution, working on data that is distributed as evenly as possible across the slices.

Amazon Redshift distributes and executes the query in parallel across all of a cluster’s compute nodes. Amazon Redshift also spreads your table data across all compute nodes in a cluster based on a distribution strategy that you specify

Whenever you perform a resize operation, Amazon Redshift will create a new cluster and migrate data from the old cluster to the new one. During a resize operation, the database will become read-only until the operation is finished

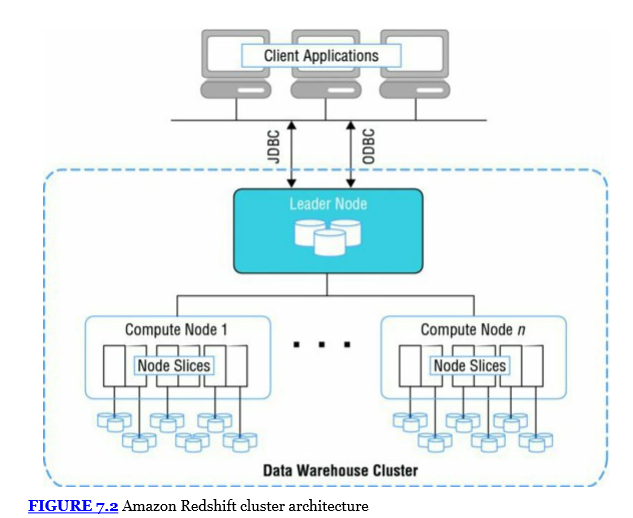


Table Design

Create table, column – specify data types

supports specifying compression encodings, distribution strategy, and sort keys.

Compressed data

The data distribution style that you select for your database has a big impact on query performance, storage requirements, data loading, and maintenance

three distribution styles:

EVEN,  - This is the default option and results in the data being distributed across the slices in a uniform fashion regardless of the data.

KEY, - With KEY distribution, the rows are distributed according to the values in one column. The leader node will store matching values close together and increase query performance for joins.

ALL. -  a full copy of the entire table is distributed to every node. This is useful for lookup tables and other large tables that are not updated frequently.

Loading Data

INSERT

UPDATE

COPY - - most efficient, read from flat file on S3 or from Dynamo DB

COPY command can read from multiple files at the same time.

After each bulk data load that modifies a significant amount of data, you will need to perform a VACUUM command to reorganize your data and reclaim space after deletes. It is also recommended to run an ANALYZE command to update table statistics

Data can also be exported out of Amazon Redshift using the UNLOAD command. This command can be used to generate delimited text files and store them in Amazon S3.

Querying Data

Amazon Redshift allows you to write standard SQL commands to query your tables

Integration with cloudwatch to monitor performance

Snapshots

Similar to Amazon RDS, you can create point-in-time snapshots of your Amazon Redshift cluster. A snapshot can then be used to restore a copy or create a clone of your original Amazon Redshift cluster.

Snapshots are durably stored internally in Amazon S3 by Amazon Redshift.

Amazon Redshift supports both automated snapshots and manual snapshots.

Security

Like other DB , refer above section

####

Redshift Spectrum

* Redshift Spectrum is a feature of Amazon Redshift that enables you to run queries against exabytes of unstructured data in [Amazon S3](https://aws.amazon.com/s3/), with no loading or ETL required

[Amazon Athena](https://aws.amazon.com/athena/) is the simplest way to give any employee the ability to run ad-hoc queries on data in Amazon S3. Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing your data immediately.

You should use [Amazon EMR](https://aws.amazon.com/emr/) if you use custom code to process and analyze extremely large datasets with big data processing frameworks such as Apache Spark, Hadoop, Presto, or Hbase.

Amazon EMR gives you full control over the configuration of your clusters and the software you install on them.

Amazon EMR is a managed service that lets you process and analyze extremely large data sets using the latest versions of popular big data processing frameworks, such as Spark, Hadoop, and Presto, on fully customizable clusters.

With Amazon EMR you can run a wide variety of scale-out data processing tasks for applications such as machine learning, graph analytics, data transformation, streaming data, and virtually anything you can code.

RDS –

Redshift – O

Storage –

            Dense compute – SSD + More CPU + RAM

            Dense storage – Magnetic Disk/HDD

Leader and compute Node

            A leader node receives queries from client applications, parses the queries and develops execution plans, which are an ordered set of steps to process these queries. The leader node then coordinates the parallel execution of these plans with the compute nodes, aggregates the intermediate results from these nodes and finally returns the results back to the client applications.

Compute nodes execute the steps specified in the execution plans and transmit data among themselves to serve these queries. The intermediate results are sent back to the leader node for aggregation before being sent back to the client applications.

Data Loading

* S3 , DynamoDB, EMR, Data Pipeline, Glue, SSH enabled Host

How  COPY command, INSERT TO, Data Pipeline , AWS Glue (fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics)

I have a lot of data for initial loading into Amazon Redshift. Transferring via the Internet would take a long time. How do I load this data?

* You can use [AWS Import/Export](https://aws.amazon.com/snowball/) to transfer the data to Amazon S3 using portable storage devices.
* use [AWS Direct Connect](https://aws.amazon.com/directconnect/) to establish a private network connection between your network or datacenter and AWS

Security

* Amazon Redshift takes care of key management but you can choose to manage your keys using your own [hardware security modules (HSMs)](http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-HSM.html) or manage your keys through [AWS Key Management Service](https://aws.amazon.com/kms/).
* Redshift Spectrum supports Amazon S3’s Server Side Encryption (SSE) using your account’s default key managed used by the AWS Key Management Service (KMS).

you can use Amazon Redshift as part of your VPC configuration.

Enable Enhanced VPC rounding to connect to S3 within region

Redshift spectrum do not support EPVC rounting

Only Leader node can access compute nodes

**What happens to my data warehouse cluster availability and data durability if my data warehouse cluster's Availability Zone (AZ) has an outage?**

* you will not be able to use your cluster until power and network access to the AZ are restored
* Data is preserved
* you can also choose to restore any existing snapshots to a new AZ in the same Region
* Amazon Redshift will restore your most frequently accessed data first so you can resume queries as quickly as possible.

Currently, Amazon Redshift only supports Single-AZ deployments.

Backup

* Amazon Redshift replicates all your data within your data warehouse cluster when it is loaded and also continuously backs up your data to S3.
* Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery
* By default, Amazon Redshift enables automated backups of your data warehouse cluster with a 1-day retention period. You can configure this to be as long as 35 days.

To increase performance - you can increase the number of nodes within your data warehouse cluster

**Massively Parallel Processing**

## Columnar Data Storage

## 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.

## Column wise storage

## cid:image001.png@01D42B32.197E8C30

## Row wise storage

## cid:image002.png@01D42B32.197E8C30

## Data Compression

## Result Caching