Amazon RedShift

is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and existing Business Intelligence (BI) tools. Amazon RedShift is a clustered peta-byte scale data warehouse and is an SQL based data warehouse used for **analytics applications**.

Amazon RedShift is an Online Analytics Processing (OLAP) type of Database which can be used for running complex analytic queries against petabytes of structured data, using sophisticated query optimization, columnar storage on high-performance local disks, and massively parallel query execution. Amazon RedShift is also ideal for processing large amounts of data for business intelligence.

**Advantages of Amazon RedShift**

The benefits of Amazon RedShift are as follows:

Amazon RedShift is extremely cost-effective as compared to some other on-premises data warehouse platforms.

Amazon RedShift is PostgreSQL compatible with JDBC and ODBC drivers available; compatible with most Business Intelligence tools out of the box.

Features parallel processing and columnar data stores which are optimized for complex queries.

Option to query directly from data files on S3 via Amazon RedShift Spectrum.

Amazon RedShift is 10x faster than a traditional SQL DB.

Amazon RedShift can store huge amounts of data but cannot ingest huge amounts of data in real time.

Amazon RedShift uses columnar data storage:

* Data is stored sequentially in columns instead of rows.
* Columnar based DB is ideal for data warehousing and analytics.
* Requires fewer I/Os which greatly enhances performance.

Amazon RedShift provides advanced compression:

* Data is stored sequentially in columns which allows for much better performance and less storage space.
* Amazon RedShift automatically selects the compression scheme.

Amazon RedShift provides good query performance and compression.

Amazon RedShift provides Massively Parallel Processing (MPP) by distributing data and queries across all nodes.

**Availability and Durability**

Amazon RedShift uses replication and continuous backups to enhance availability and improve durability and can automatically recover from component and node failures.

Only available in one AZ but you can restore snapshots into another AZ.

Alternatively, you can run data warehouse clusters in multiple AZ’s by loading data into two Amazon RedShift data warehouse clusters in separate AZs from the same set of Amazon S3 input files.

Amazon RedShift replicates your data within your data warehouse cluster and continuously backs up your data to Amazon S3.

Amazon RedShift always keeps three copies of your data:

* The original.
* A replica of compute nodes (within the cluster).
* A backup copy on S3.

Amazon RedShift provides continuous/incremental backups:

* Multiple copies within a cluster.
* Continuous and incremental backups to S3.
* Continuous and incremental backups across regions.
* Streaming restore.

Amazon RedShift provides fault tolerance for the following failures:

* Disk failures.
* Node failures.
* Network failures.
* AZ/region level disasters.

For node failures the data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the DB.

High availability for Amazon RedShift:

* Currently, Amazon RedShift does not support Multi-AZ deployments.
* The best HA option is to use a multi-node cluster which supports data replication and node recovery.
* A single node Amazon RedShift cluster does not support data replication and you’ll have to restore from a snapshot on S3 if a drive fails.

Amazon RedShift can asynchronously replicate your snapshots to S3 in another region for DR.

Single-node clusters do not support data replication (in a failure scenario you would need to restore from a snapshot).

Scaling requires a period of unavailability of a few minutes (typically during the maintenance window).

During scaling operations Amazon RedShift moves data in parallel from the compute nodes in your existing data warehouse cluster to the compute nodes in your new cluster.

By default, Amazon RedShift retains backups for 1 day. You can configure this to be up to 35 days.

If you delete the cluster, you can choose to have a final snapshot taken and retained.

Manual backups are not automatically deleted when you delete a cluster.

**Security**

You can load encrypted data from S3.

Supports SSL Encryption in-transit between client applications and Amazon RedShift data warehouse cluster.

VPC for network isolation.

Encryption for data at rest (AES 256).

Audit logging and AWS CloudTrail integration.

Amazon RedShift takes care of key management, or you can manage your own through HSM or KMS.

**Charges**

Charged for compute nodes hours, 1 unit per hour (only compute node, not leader node).

Backup storage – storage on S3.

Data transfer – no charge for data transfer between Amazon RedShift and S3 within a region but for other scenarios you may pay charges.

HDD and SSD storage options.

The size of a single node is 160GB and clusters can be created up to a petabyte or more.

Multi-node consists of:

Leader node:

* Manages client connections and receives queries.
* Simple SQL endpoint.
* Stores metadata.
* Optimizes query plan.
* Coordinates query execution.

Compute nodes:

* Stores data and performs queries and computations.
* Local columnar storage.
* Parallel/distributed execution of all queries, loads, backups, restores, resizes.
* Up to 128 compute nodes.

Amazon RedShift Spectrum is a feature of Amazon RedShift that enables you to run queries against exabytes of unstructured data in Amazon S3, with no loading or ETL required.

**Use Cases of Amazon RedShift**

A data warehouse for enterprise operations: Many organizations work with data from multiple sources, such as advertising, customer relationship management, and customer support.

As a centralized repository, Redshift can be used to store data from multiple sources in a unified schema and structure. This can then feed enterprise-wide reporting and analytics.

In business intelligence and analytics, Redshift’s fast query execution against terabyte-scale data makes it an excellent selection. BI tools such as Tableau often use Redshift as the underlying database (which would otherwise struggle to perform queries and joins of large datasets).

Organizations may choose to monetize their data by exposing it to their customers through embedded analytics and analytics as a service. In these scenarios, Redshift’s data sharing, search, and aggregation capabilities make it ideal, as it allows customers to access only relevant subsets of data while keeping other databases, tables, or rows confidential.

As long as the cluster is adequately resourced, Redshift’s performance is consistent and predictable. It is therefore a popular choice for data-driven applications, such as reporting and calculations.

Database migration and change data capture: AWS Database Migration Service (DMS) can be used to replicate changes in an operational data store into Amazon Redshift. It is typically done to provide more flexibility in analysis, or when migrating from legacy data warehouses.

<https://digitalcloud.training/aws-application-integration/>

<https://digitalcloud.training/amazon-s3-and-glacier/>

<https://digitalcloud.training/aws-lambda/>