Amazon Redshift

Amazon Redshift

- Amazon Redshift is a fast, fully-managed, petabyte-scale data warehouse service that makes it simple and cost-effective to efficiently analyze all your data using your existing business intelligence tools.
- It is optimized for datasets that range from a few hundred gigabytes to a petabyte or more.
- Traditional data warehouses require significant time and resources to administer, especially for large datasets.
- The financial cost associated with building, maintaining, and growing self-managed, on-premises data warehouses is very high.

Amazon Redshift

- Amazon Redshift manages the work needed to
  - Set up,
  - Operate, and
  - Scale a data warehouse,
  - Provisioning the infrastructure capacity
  - Automating ongoing administrative tasks such as backups and patching

Ideal Usage Patterns

- Analyze global sales data for multiple products
- Store historical stock trade data
- Analyze ad impressions and clicks
- Aggregate gaming data
- Analyze social trends
- Measure clinical quality, operation efficiency, and financial performance in the health care space

Performance

- Amazon Redshift uses a variety of innovations to obtain very high query performance on datasets ranging in size from hundreds of gigabytes to a petabyte or more.
- It uses columnar storage, data compression, and zone maps to reduce the amount of I/O needed to perform queries.
- Amazon Redshift has a massively parallel processing (MPP) architecture that parallelizes and distributes SQL operations to take advantage of all available resources.
- The underlying hardware is designed for high performance data processing that uses local attached storage to maximize throughput.

Durability and Availability

- Amazon Redshift has multiple features that enhance the reliability of your data warehouse cluster.
- Amazon Redshift stores three copies of your data—all data written to a node in your cluster is automatically replicated to other nodes within the cluster, and all data is continuously backed up to Amazon S3. Snapshots are automated, incremental, and continuous.
- Amazon Redshift stores your snapshots for a user-defined period, which can be from one to thirty-five days.
- At any time, you can create one or more manual snapshots, which are retained until explicitly deleted.

Durability and Availability

- Amazon Redshift also continuously monitors the health of the cluster and automatically re-replicates data from failed drives and replaces nodes as necessary Amazon Redshift has multiple features that enhance the reliability of your data warehouse cluster.
- Amazon Redshift stores three copies of your data—all data written to a node in your cluster is automatically replicated to other nodes within the cluster, and all data is continuously backed up to Amazon S3.
- Snapshots are automated, incremental, and continuous. Amazon Redshift stores your snapshots for a user-defined period, which can be from one to thirty-five days. At any time, you can create one or more manual snapshots, which are retained until explicitly deleted.
- Amazon Redshift also continuously monitors the health of the cluster and automatically re-replicates data from failed drives and replaces nodes as necessary.

Cost Model

- With Amazon Redshift, you can pay as you go and there are no upfront costs. Amazon Redshift has three pricing components: data warehouse node hours, backup storage, and data transfer.
- Compute node hours are the total number of hours run across all compute nodes for the billing period.
- Backup storage is the storage associated with automated and manual snapshots for an Amazon Redshift data warehouse cluster.
- Increasing the backup retention period or taking additional snapshots increases the backup storage consumed by the Amazon Redshift data warehouse cluster.
- There is no additional charge for backup storage up to 100% of your provisioned storage for an active data warehouse cluster. There is no data transfer charge for data transferred to or from Amazon Redshift outside of Amazon Virtual Private Cloud (Amazon VPC)

Scalability and Elasticity

- Amazon Redshift provides “pushbutton scaling” of compute nodes within a data warehouse cluster. With a few clicks of the AWS Management Console or a simple API call, you can easily scale the number of nodes in your data warehouse cluster up or down as your performance or capacity needs change.
- An Amazon Redshift data warehouse cluster can be started with as little as a single 2 TB XL node and scale all the way to a hundreds 16 TB 8XL nodes for 1.6 PB of compressed user data.
- Amazon Redshift will place your existing cluster into read-only mode, provision a new cluster of your chosen size, and then copy data from your old cluster to your new one in parallel.
- Queries can continue running against the old cluster while the new one is being provisioned.
- Once the data has been copied to the new cluster, Amazon Redshift will automatically redirect queries to the new cluster and remove the old cluster

Interfaces

- The Amazon Redshift Query API provides a management interface to manage data warehouse clusters programmatically.
- Additionally, the AWS SDKs for Java, .NET, and other languages provide class libraries that wrap the underlying Amazon Redshift API to simplify your programming tasks.
- If you prefer a more interactive way of managing clusters, you can use the Amazon Redshift console and the AWS CLI.

Interfaces

- The Amazon Redshift APIs do not provide a data interface.
- Amazon Redshift is a SQL data warehouse and uses industry standard ODBC and JDBC connections and PostgreSQL drivers.
- Once you've provisioned your cluster, you can connect to it, start loading data, and run queries using the same SQL-based tools and business intelligence applications you use today.

Interfaces

- Data can be loaded into Amazon Redshift from a range of data sources including Amazon S3, Amazon DynamoDB, and AWS Data Pipeline.
- Amazon Redshift attempts to load data in parallel into each compute node to maximize the rate at which data can be ingested into the data warehouse cluster.

Anti-Patterns

- OLTP workloads—Amazon Redshift is a column-oriented database suited to data warehouse and analytics, where queries are typically performed over very large datasets. If your application involves online transaction processing, a traditional row-based database system, such as Amazon RDS, is a better match.
- BLOB data—If you plan on storing binary (e.g., video, pictures, or music), you'll want to consider Amazon S3.