Redshift

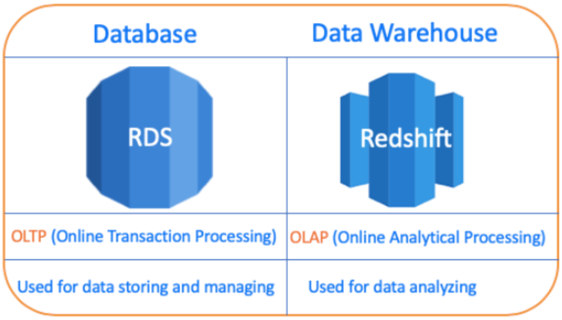
What is Amazon Redshift?

Amazon Redshift is a fully managed, cloud-based, petabyte-scale **data warehouse** service by Amazon Web Services (AWS).

Amazon Redshift is an efficient solution to **collect and store** all your data to **analyze**. It gives you fast querying capabilities over structured data using familiar SQL-based clients and business intelligence (BI) tools.

Amazon Redshift also includes Amazon Redshift Spectrum, allowing you to directly run SQL queries in Amazon S3 data lakes.

Why Amazon Redshift?


Database
Data Warehouse
RDS
Redshift
OLTP (Online Transaction Processing)
OLAP (Online Analytical Processing)
Used for data storing and managing
Used for data analyzing

RDS vs. Redshift

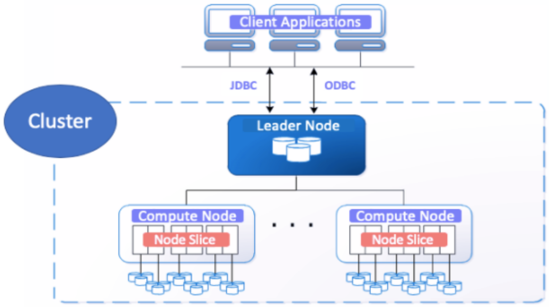
But we can also use a database such as RDS to store our data and analyze them with SQL queries. So, what is the reason to use Amazon Redshift/Data Warehouse instead of the database?

The answer is hidden in the category of data processing. There are two categories of data processing, OLTP(Online Transaction Processing) and OLAP(Online Analytical Processing). While Conventional databases like RDS are in the OLTP category, Data Warehouses like Redshift are in the OLAP Category.

In fact, it's possible to make an analysis on the OLTP database (like RDS) but processing of both storing and analyzing at the same time increases the workload of the database. So, your queries start taking a lot of time and the size of data becomes unmanageable on conventional databases.

Thus, we prefer data warehousing which is in OLAP class and can make data easily accessible for **reporting and analytics**. OLAP data warehouse is designed for work together with SQL query and business intelligence (BI) tools like Microsoft Power BI

Redshift System Architecture and Components



Redshift System Architecture

Client Applications:

Amazon Redshift integrates with various data loading and ETL (extract, transform, and load) tools and business intelligence (BI) reporting, data mining, and analytics tools.

Business Intelligence (BI) refers to technologies, applications, and practices for the collection, integration, analysis, and presentation of business information. The purpose of Business Intelligence is to support better business decision making. For example, Microsoft Power BI or Microsoft Excel.

Connections:

Amazon Redshift communicates with client applications by using industry-standard JDBC and ODBC drivers for PostgreSQL.

Clusters:

The **core infrastructure component** of an Amazon Redshift data warehouse is a cluster.

A cluster is **composed of one or more compute nodes**. Your client application interacts directly only with the leader node in the cluster. The compute nodes are transparent to external applications.

Each cluster runs an Amazon Redshift engine and contains one or more databases.

Leader Node:

The leader node **manages communications with client programs** and all **communication with compute nodes**.

The leader node distributes SQL statements to the compute nodes only when a query references tables that are stored on the compute nodes. All other queries run exclusively on the leader node. Amazon Redshift is designed to implement certain SQL functions only on the leader node. A query that uses any of these functions will return an error if it references tables that reside on the compute nodes.

Compute Nodes:

The compute nodes execute the compiled code and send intermediate results back to the leader node for final aggregation.

Each compute node has its own dedicated CPU, memory, and attached disk storage, which is determined by the node type. As your workload grows, you can increase the computing capacity and storage capacity of a cluster by increasing the number of nodes, upgrading the node type, or both.

Amazon Redshift provides three node types;

- RA3 node types enable you to scale storage and compute independently to fit your needs.
- DS2 node types are optimized for large data workloads and use hard disk drive (HDD) storage.
- DC2 node types are optimized for performance-intensive workloads and use SSD storage.

Node Slices:

A compute node is partitioned into slices. Each slice is allocated a portion of the node's memory and disk space, where it processes a portion of the workload assigned to the node. The number of slices per node is determined by the node size of the cluster.

Complementary Lesson about AWS Redshift

aws
Introduction to Data Warehousing on AWS with Amazon Redshift
Watch later
Share
Redshift