Bases II GR 1 - Resumen 2 - 07/03/2023

Gerald Núñez Chavarría - 2021023226

Data Warehousing on AWS

1. Introducing Amazon Redshift

In the past, when enterprises wanted to increase data volume or available information for more users, they had to accept slow queries or go through a costly system upgrade process. Cloud warehouses like Amazon Redshift store data without the need to deploy expensive systems and maintaining the features, scale and performance.

Amazon Redshift is a fast, fully managed, petabyte-scale data warehousing solution that makes it simple and cost-effective to analyze large volumes of data using existing business intelligence (BI) tools.. Since launching in February 2013, it has been one of the fastest AWS service, companies like NTT DOCOMO, FINRA, Johnson & Johnson, McDonalds... has migrated to Amazon Redshift.

2. Modern Analytics and data warehousing architecture

AWS analytics services help enterprises to convert their data to answers providing integrated analytics services. It provides a easy path to make data lakes and date warehouses, a secure cloud, storage and networks infrastructure, a fully integrated analytics stack with a mature set of analytics tools and the best performance, most scalability and lowest price. AWS Lake Formation enables secured, self-service discovery and access for users.

Analytics pipelines are designed to handle large volumes of incoming streams of data from heterogeneous sources such as databases, applications, and devices. Pipeline has to: 1- Collect different data types like transactional data, log data, streaming data, loT. 2- Store data in lakes house, warehouses and data mart. 3- It has two types of process to process data, batch process and real-time processing. 4- You can analysis data and create visualizations with a variety of tools, such as MySQL Workbench, Amazon QuickSight, Amazon Redshift, Amazon S3, Amazon RDS...

3. Data warehouse technology options

There are three process to build a data warehouse:

1. Row oriented-database: Store whole rows in a physical block. To increase the performance developers used same techniques like: Building materialized views, creating pre-aggregated roll up tables, Building indexes on every possible predicate combination. The queries in this system has to read all the columns and all the rows where the query predicate condition is correct.
2. Column-oriented databases: Column-oriented databases organize each column in its own set of physical blocks instead of packing the whole rows into a block, with this are more I/O efficiently in read-only queries, instead of read all the columns in a disk, it reads only those column that query need. It is better than row oriented-database.
3. Massively Parallel Processing (MPP) architectures: MPP let you use all the resources in a cluster for processing data, which dramatically increases performance of petabyte scale data warehouses. MPP data warehouses allow you improve performance by simply adding more nodes to the cluster.

4. Amazon Redshift deep dive

Performance: The Redshift Spectrum feature in Amazon Redshift that enables querying and writing data to a data lake in open file formats such as Parquet, ORC, JSON, Avro, and CSV using ANSI SQL. To export data to the data lake, one can use the Redshift UNLOAD command with Parquet as the file format. To query data in the data lake, an external schema or table can be created.

Amazon Redshift offers a industry-leading performance with flexibility, and this is for its multiple features such as: High performing hardware, AQUA (Advanced Query Accelerator), efficient storage and high-performance query processing, materialized views, auto workload management to maximize throughput and performance, and result caching.

Elasticity and scalability: Amazon provide two forms to get this features: 1- Elastic resize: With the elastic resize feature, you can quickly resize your Amazon cluster by adding nodes to get the resources needed for demanding workloads, and to remove nodes when the job is complete to save cost. 2- Concurrency scaling: With the Concurrency Scaling feature, you can support virtually unlimited concurrent users and concurrent queries, with consistently fast query performance. When concurrency scaling is enabled, Amazon Redshift automatically adds additional compute capacity when you need it to process an increase in concurrent read queries.

Amazon Redshift managed storage: You can pay for the compute and storage independently base on what you need. It automatically use high performance.

5. Operations

Amazon redshift automates operational task, including cluster performance and cost optimization.

The amazon redshift advisor tool make recommendations about changes to make, it uses metrics for your cluster and the changes recommendation analysis you should make is in order of impact.

Amazon Redshift provides custom JDBC and ODBC drivers. The platform also offers a Query Editor in the web console for running SQL. Amazon Kinesis Data Firehose enables loading of streaming data for near real-time analytics with existing BI tools. Metrics for compute, memory, and storage utilization can be accessed through the console or Amazon CloudWatch API.

Amazon Redshift offers a flexible pricing approach, without long-term commitments or upfront costs, based on the size and number of nodes in the cluster. Charges apply for backup storage beyond provisioned storage and backups stored after cluster termination. For Redshift Spectrum, users pay for query costs based on data scanned.

Amazon redshift is not recommended for the following patterns: 1- OLTP: If you require a fast transactional system, you might want to choose a relational database system. 2- Unstructured data: You must have a defined data schema, because redshift does not support an arbitrary schema.