Fiorella Zelaya Coto - 2021453615

Introducing Amazon Redshift

Amazon Redshift is a cloud data warehouse that reduces the cost and efforts associated with deploying data warehouse systems, without compromising on features, scale and performance. Besides, Amazon Redshift is a good data warehousing solution that analyzes large volumes of data using Business Intelligence (BI) tools. It's very fast, providing the performance of a columnar warehousing.

Modern analytics and data warehousing architecture

The data flows into a data warehouse from transactional systems and other relational databases, then, the data is processed, transformed and ingested. The data can be accessed through BI tools, SQL clients and other tools.

Data warehouses are optimized for **batched write operations** and reading high volumes of data employing denormalized schemas, while Online Transaction Processing (OLTP) Databases are optimized for **continues write operations** and high volume of **small** read operations employing highly normalized schemas.

AWS analytics services

AWS analytics services convert data to answers, providing mature and integrated analytics services, ranging from cloud data warehouses to serverless data lakes. With AWS it's easier to build data lakes and datawarehouses, gives you a mature set of analytic tools, the best performance, the most scalability and the lowest cost for analytics. Amazon Elastic Compute Cloud (Amazon EC2) spot instances to reduce cost and run analytics faster.

Analytics architecture

Analytics pipelines handle large volumes of incoming streams of data. It has 4 stages: Collect, store, process and analyze and visualize the data.

For Data Collection, AWS provides solutions for data storage for many types of data.

- **Transactional data:** Typically stored in relational database management systems or NoSQL database systems. **Database services:** Amazon DynamoDB (noSQL), Amazon Aurora (mySQL and PostgreSQL compatible) and Amazon RDS.
- **Log data:** Amazon S3 is a solution for log data, which helps you troubleshoot issues and perform analytics by using the information stored in the logs.
- **Streaming data:** Amazon Kinesis or Amazon MSK can be used for this type of data.
- **IoT data:** AWS IoT can be used for connected devices to interact easily and securely with the AWS cloud.

For Data Processing, there are two types of processing workflows.

- **Batch processing:**

    - **Extract Transform Load (ETL):** Pulls data and loads it into data warehouses systems.
    - **Extract Load Transform (ELT):** Pulls data and loads it into the target system first.
    - **Online Analytical Processing (OLAP):** Store aggregated historical data in multidimensional schemas.
- **Real-time processing:** The data can be processed on a record-by-record basis or over sliding time windows. It requires a highly concurrent and scalable processing layer.
    - AWS Lambda, Amazon KCL, Amazon KDF, Amazon MSK or AWS Glue can be used to process the data.

For Data Storage, the data can be stored in:

- **Lake house:** Combines elements of data warehouse and data lakes. Store data in open file formats in your data lake and query it in place while joining with data warehouse data.
- **Data warehouse:** Run fast analytics on large volumens of data and unearth patterns hidden in the data by leveraging BI tools.
- **Data mart:** Focused on specific functional areas.

Analysis and Visualization

The right tools are needed to analyze and visualiza data. Amazon QuickSight creates visualizations, perfoms analysis and gets business insights from the data. It offers native integration with AWS sources like Amazon Redshift, Amazon S3 and Amazon RDS. With S3, Amazon Athena/QuickSight Integrations can be used to perform analysis and visualization.

Data warehouse technology options

- **Row-oriented databases:** Store whole rows in a physical block. High performance for read operations is achieved through secondary indexes. Better suited for OLTP than for analytics. Limited by the resources available on a single machine. Datamarts can alleviate the problem a little by using functional sharding, but if data marts grow large, data processing slows down.
- **Column-oriented databases:** Organize each column in its own set of physical blocks. This allows them to be more I/O efficient for read-only queries and improves compression, so you need less storage compared to a row-oriented database. This is better for data warehousing.
- **Massively Parallel Processing (MPP) Arquitectures:** This arquitecture enables the use of all resources available in the cluster for processing data, which increases performance of petabyte scale data warehouses and improves performance by adding more nodes to a cluster. Hadoop and Spark support MPP.

Amazon Redshift deep dive

Amazon Redshift offers key benefits for performant, cost-effective data warehousing, including efficient compression, reduces I/O and lower storage requirements. Delivers fast query and I/O performance by using columnar storage, and by parallelizing and distributing queries across multiple nodes. It automates many administrative taks.

Integration with data lake

Redshift Spectrum: Makes it easier to query data and write data back to the data lake in open file formats. With Spectrum you can query open files like Parquet, JSON, ORC, SCV, Avro, etc and more directly using ANSI SQL. Exporting the data to the data lake can be donne using Redshift UNLOAD command in SQL code and specifying the Parquet file format.

Performance

Amazon Redshift offers fast, industry-leading performance with flexibility. To achieve this, it offers these features:

- **High performing hardware**
- **AQUA:** Advanced Query Accelerator. Distributed and hardware-accelerated cache than enables Amazon Redshift to run faster than any cloud data warehouse.
- **Efficient storage and high-performance query processing:** Columnar storage, data compression, zone maps.
- **Materialized views:** Dashboarding, queries from BI tools, ELT data processing jobs.
- **Auto workload management to maximize througput and performance:** Uses machine learning to predict and classify queries to manage resources and concurrency while prioritizing business critical workloads and detect opportunities to improve performance.
- **Result caching:** Deliver sub-second response for repeated queries, giving a significant performance boost.

Durability and availability

Amazon Redshift automatically detects and replaces any failed node in the data warehouse cluster and loads the most frequently accessed data first to resume the querying of the data as soon as posible. It tries to mantain at least three copies of data (original, replica and backup). Amazon Redshift clusters reside within one Avalilability Zone. A mirror to self-manage replication and failover can be created. A robust disaster recovery (DR) environment can be set up as well.

Elasticity and scalability

Amazon Redshift provides the capacity of scaling compute and storage independently and pay only for what is used. It offers two forms of compute elasticity:

- **Elastic resize:** Quickly resize the Amazon Cluster by adding nodes to ger the resources needed for demanding workloads and removing them when the job is complete.
- **Concurrency Scaling:** This feature offers support for unlimited concurrent users and queries with fast performance. User always see the most current data.

Amazon Redshift managed storage

Amazon Redshift managed storage offers the capacity of scaling and paying for compute and storage independently so the size of the cluster can be based only on the compute needs.

Operations

Amazon Redshift automates many operational tasks such as Cluster Performance and Cost Optimization.

Amazon Redshift Advisor

Helps to improve performance and decrease operating costs for the cluster. It offers specific recommendations about changes to make by analyzing the workload and using metrics for the cluster.

Interfaces

A wide range of familiar SQL clients can be used thanks to Amazon Redshift's custom Java Database Connectivity (JDBC) and Operan Database Connectivity (ODBC) drivers. It provides a built-in Query Editor in the web console so DBAs or users run queries as needed. With other integrations loads and unloads execute in parallel on each compute node and loading streaming data can be easily done with Amazon Kinesis Data Firehose.

Security

To provide data security, Amazon Redshift can be run inside a virtual private cloud based on the Amazon Virtual Private Cloud (Amazon VPC) service. The data can be accessed only from the cluster's leader node, providing another layer of security. It supports encryption, the I/O subsystem encrypts everything written to disk. Backups are encrypted. It also supports means of authentication.

Cost model

Charges are bases on the size and number of nodes in the cluster. Besides, there is no additional charge for backup storage. Lastly, there is no additional charge for communication between S4 and Amazon Redshift.

Ideal usage patterns

Amazon Redshift is ideal for OLAP using BI tools, like running enterprise BI and reporting, analyze global sales data, stores historical stock trade data, analyze ad impressions and clicks, aggregating gaming data, analyze social trends, etc. Semi-structured data is supported by using Redshift Spectrum.

Anti-Patterns

Amazon Redshift is not ideally suited for:

- **OLTP:** Relational Database System such as Amazon Aurora or Amazon RDS or noSQL database like DynamoDB is a better choice.
- **Unstructured data:** ETL on Amazon EMR is a better choice.
- **BLOB data:** For binary large objects, storing data in S3 and referencing the location on Amazon Redshift is a better choice.

Resumen de: Data Warehousing on AWS: AWS Whitepaper