Resumen 1 (R1)
Estudiante: Celina Madrigal Murillo
Carné: 2020059364

Introduction

Data is an enterprise's most valuable asset. Most large enterprises have data warehouses for reporting and analytics purposes.They use data from a variety of sources, including their own transaction processing systems, and other databases.

Introducing Amazon Redshift

Cloud data warehouses like Amazon Redshift changed how enterprises think about data warehousing by dramatically lowering the cost and effort associated with deploying data warehouse systems, without compromising on features, scale, and performance. Amazon Redshift is a fast, fully managed, petabyte-scale data warehousing solution.

Modern Analytics and Data Warehousing Architecture

Differences between data warehouses and OLTP databases.

- **Data warehouses** are optimized for batched write operations and reading high volumes of data.
- **OLTP databases** are optimized for continuous write operations and high volumes of small read operations.

AWS Analytics Services

AWS gives you: An easy path to build data lakes and data warehouses, a secure cloud storage, a fully integrated analytics stack, the best performance, the most scalability, and the lowest cost for analytics.

Analytics Architecture

A typical analytics pipeline has the following stages: Collect data, Store the data, Process the data and Analyze and visualize the data.

Data Collection

Transactional Data

- A **NoSQL** database is suitable when the data is not well-structured to fit into a defined schema, or when the schema changes often.
- An **RDBMS** solution is suitable when transactions happen across multiple table rows and the queries require complex joins.

Log Data

Reliably capturing system-generated logs helps you troubleshoot issues, conduct audits, and perform analytics using the information stored in the logs.

Streaming Data

Web applications, mobile devices, and many software applications and services can generate staggering amounts of streaming data that need to be collected, stored, and processed continuously.

IoT Data

Devices and sensors around the world send messages continuously. Enterprises today need to capture this data and derive intelligence from it.

Data Processing

Batch Processing

Extract Transform Load (ETL), Extract Load Transform (ELT) and Online Analytical Processing (OLAP)

Real-Time Processing

Using the processed data for a wide variety of analytics, including correlations, aggregations, filtering, and sampling is called real-time processing.

Data Storage

Lake house, Data warehouse and Data mart

Analysis and Visualization

Amazon QuickSight is a fast, cloud-powered BI service that enables you to create visualizations, perform analysis as needed, and quickly get business insights from your data.

Analytics Pipeline with AWS Services

AWS offers a broad set of services to implement an end-to-end analytics platform.

Data Warehouse Technology Options

Row-Oriented Databases

They typically store whole rows in a physical block.

Column-Oriented Databases

They organize each column in its own set of physical blocks instead of packing the whole rows into a block.

Massively Parallel Processing (MPP) Architectures

An MPP architecture enables you to use all the resources available in the cluster for processing data, which dramatically increases performance of petabyte scale data warehouses.

Amazon Redshift Deep Dive

It offers key benefits for performant, cost-effective data warehousing.

Integration with Data Lake

Red Redshift Spectrum makes it easier to both query data and write data back to your data lake in open file formats.

Performance

Amazon Redshift offers multiple features to achieve superior performance, including: High performing hardware, AQUA (preview), Efficient storage and high-performance query processing, Materialized views, Auto workload management to maximize throughput and performance and Result caching.

Durability and Availability

Amazon Redshift automatically detects and replaces any failed node in your data warehouse cluster.

Elasticity and Scalability

Elastic resize and Concurrency Scaling

Amazon Redshift Managed Storage

It enables you to scale and pay for compute and storage independently so you can size your cluster based only on your compute needs.

Operations

Ideal Usage Patterns

Enterprises use Amazon Redshift to do the following: Running enterprise BI and reporting, Analyze global sales data for multiple products, Store historical stock trade data, Analyze ad impressions and clicks, Aggregate gaming data, Analyze social trends and Measure clinical quality, operation efficiency, and financial performance in health care

Anti-Patterns

OLTP, Unstructured data and BLOB data