Instituto Tecnológico de Costa Rica

Bases de Datos - IC4302

Resumen 1

Estudiante:

Andrea María Li Hernández - 2021028783

Profesor:

Gerardo Nereo Campos Araya

Fecha de entrega: 9 de agosto del 2022
Segundo Semestre, 2022

Introduction

Data is an enterprise's **most valuable** asset. Most large enterprises have data warehouse for reporting and analytics purposes. In the past, building and running a data warehouse was complicated and expensive, because of the software and hardware expenses, as well as time consuming.

In this summary, you can find information about Modern Analytics and Data Warehousing Architecture, along with a deep dive into Amazon Redshift and other technology choices available .

Introducing Amazon Redshift

- It is a fast, fully managed, petabyte-scale data warehousing technology that makes it simple and cost-effective to analyze large volumes of data thorough existing business intelligence (BI) tools.
- It lowers cost and effort, without compromising on features, scale, and performance.

Modern Analytics and Data Warehousing Architecture

Data Warehouse	**Online Transaction Processing (OLTP) databases**
1. Optimized for: • Batched write operations. • Reading high volumes of data.	1. Optimized for: • Continuous write operations. • High volumes of small read operations.
2. Generally employ denormalized schemas like the Star schema and Snowflake schema.	2. Employ highly normalized schemas.
3. High data throughput requirements.	3. High transaction throughput requirements.

AWS Analytics Services

AWS analytics services help enterprises quickly convert their data to answers by providing:

1. An easy path to build data lakes and data warehouses.
2. A secure cloud storage and infrastructure for the analytic workloads.
3. Best performance and lowest cost for analytics.

Analytics Architecture

Analytics pipelines are designed to handle large volumes of incoming data from sources such as databases, applications, and devices. Stages:

1. **Data Collection and storage**: AWS provides solutions for data storage for different types of data. Some examples:
    - **Transactional Data**: Such as e-commerce purchase and financial transactions.

    - **Streaming Data**: High volumes of data coming from web applications, mobile devices, and many other.

2. **Data Storage**:

    - **Lake house**: You can query data cross your data watehouse, data lake, and operational databases.
    - **Data warehouse**: You can run fast analytics on large volumes of data.
    - **Data mart**: It is a simple form of data warehouse focused on a specific functional area or subject matter.

3. **Data Processing**: Load the raw data into a data warehouse to perform further analysis. Processing workflows:

    - **Batch Processing**: Commonly uses online analytic processing (OLAP).
    - **Real-Time Processing**: Processing used to process streaming data.

4. **Analysis and Visualization**: You need to analyze and visualized the processed data. Some tools:

    - **Amazon QuickSight**: Native integration with AWS data sources.
    - **Amazon Athena/QuickSight integration**: If you are using S3 as your primary storage.

Data Warehouse Technology Options

Row-Oriented Databases

- They typically store whole rows in a physical block. Some databases include Oracle Database Server, Microsoft SQL Server, MySQL and PostgreSQL. They are better suited for OLTP.
- Every query must read though **all of the columns for all of the rows** in the block that satisfy the predicate.

Column-Oriented Databases

- They organize **each column in its own set** of physical blocks. Some databases include Amazon Redshift, Vertica and Druid.
- It is more input/output efficient for read-only queries and improved compression.
- Better choice for data warehousing.

Massively Parallel Processing (MPP) Architectures

- This enables you to use **all the resources available** in the cluster for processing data. Some databases that use MPP include Amazon Redshift and Greenplum.
- Improve performance by simply **adding more nodes to the cluster.**

Amazon Redshift Deep Dive

- It offers efficient compression, reduced I/O and lower storage requirements.

- It is based on ANSI SQL.
- It automates most of the common administrative tasks associated with the managing of the data warehouse.

Integration with Data Lake

Amazon Redshift provides a feature called Redshift Spectrum that makes it easier to query and write data back to the data lake in open formats such as JSON and CSV.

Performance

- High performing hardware.
- Efficient storage and high-performance query processing.
- Auto workload management to maximize throughput and performance.

Durability and Availability

- It automatically detects and replaces any failed node in the data warehouse cluster.
- It attempts to maintain at least three copies of data.
- It has a robust disaster recovery environment that automatically takes incremental backups of the data every eight hours. Also, you can keep backups in multiple AWS Regions.

Elasticity and Scalability

Quickly resize your Amazon cluster by adding or removing nodes. Also, support virtually unlimited concurrent users and concurrent queries.

Amazon Redshift Managed Storage

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

Operations

Ideal Usage Patterns

To **analyze** global sales data for multiple products, social trends and ad impressions. Run analysis on **large volumes** of event data, offload infrequently accessed history data and join the external dataset with the data warehouse without loading it.

Anti-Patterns

Amazon Redshift is not ideally suited for this usage patterns:

- **OLTP**: There are better options such as Amazon Aurora to get a fast transactional system.
- **Unstructured data**: Amazon Redshift doesn't support an arbitrary schema structure.
- **BLOB data**: You might want to store this data in S3 and reference its location in Amazon Redshift.