TEC | Tecnológico de Costa Rica

Instituto Tecnologico de Costa Rica

Sede Central de Cartago

Escuela de Computación

Resumen 1 (R1)

IC-4302 Bases De Datos II GR 1

Estudiantes: Deyan Sanabria Fallas #2021046131

Profesor: Gerardo Nereo Campos Araya

Fecha de Entrega: 09/08/2022

Segundo Semestre 2022

Indice

Introduction

Data is a very valuable resource for innovation and growth. In the past, most enterprises had to build, run, and maintain their own data warehouses for reporting and analytics, a very complex and expensive job with all kinds of challenges.

That's why, this document provides information for taking advantage of Amazon Web Services (AWS).

Introducing Amazon Redshift

In the past, expanding data warehouse services had two possible options, making queries slow, or upgrading hardware, a very time consuming and expensive job. With Amazon Redshift is easier than ever, allowing enterprises to scale without sacrificing performance or features at a very low cost.

Modern Analytics and Data Warehousing Architecture

Why to build a data warehouse instead of using an OLTP?

> **Data warehouse**: Batched Writing/Reading high volumes of data. High data throughput and denormalized schemas.

> **OLTP (OnLine Transaction Processing) databases**: Continuous Writing/ High volume of small read operations. High transaction throughput and highly normalized schemas.

To use the benefit of data warehouse as a separate storage with a OLTP is recommended to build an efficient pipeline.

AWS Analytics Services

AWS provides an easy way to convert data to answers with their analytics services, lets you build data lakes and data warehouses to run diverse analytics workloads while giving the best performance, scalability, and the lowest cost.

Analytics Architecture

Analytics pipelines handle large volumes of incoming streams of data, typically it has these stages:

> **Data Collection**: AWS provides data collection for these types of data
>
> - *Transactional Data* -> RDBMS and NoSQL databases
> - *Log data*
> - *Streaming data*
> - *IoT Data*

> **Data storage**:
>
> - *Lake House*: Combines Data warehouses and Data lakes

> - *Data warehouse*
> - *Data mart*: like a data warehouse but specialized

> **Data processing**:
>
> - *Batch Processing*: used by OLAP’s
> - *Real-Time processing*: used by OLTP’s

Data Warehouse Technology Options

Row-Oriented Databases

MySQL, SQL Server, PostgreSQL. Store rows in a physical block. High performance read operations through secondary indexes. Better suited for OLTP than analytics. Queries reads through all columns and rows in the predicate, producing bottleneck.

> **Optimization**:
>
> - *Materialized views*
> - *Pre-aggregated rollup tables*
> - *Indexes on predicate combination*
> - *Data partitioning*
> - *Index-based joins*

Column-Oriented Database

Store columns in a physical block. More I/O efficient for read-only. Better for data warehouse. Improve compression: Every block uses same data types

Massively Parallel Processing (MPP) Architectures

Allows you to use all resources of the cluster for processing data. Increases performance of petabyte scale data houses.

Amazon Redshift Deep Dive

Integration with Data Lake

Redshift Spectrum makes query data and write data back to data lakes easier in open file formats like Parquet, ORC, JSON, Avro, CSV

Performance

High performing hardware, AQUA, Efficient storage and high-performance query processing, Materialized views, Auto workload management and Result caching.

Durability and Availability

Automatically detects and replaces any failed node. Attempts to maintain at least three copies of data: the original and replica on the compute nodes, and a backup

Elasticity and Scalability

- **Elastic resize**: Resize cluster by adding nodes and remove nodes to save cost. Minimal disruption.
- **Concurrency Scaling**:Automatically adds additional compute capacity when you need it to process an increase in concurrent read queries. Write operations continue as normal on your main cluster. Users always see the most current data.

Manage Storage

Enables to scale and pay for compute and storage independently

Operations

Ideal Usage Patterns

- *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*
- *Measure clinical quality, operation efficiency, and financial performance in health care*

Anti-Patterns

- **OLTP**: Redshift is designed for data warehousing workloads delivering extremely fast and inexpensive analytic capabilities.
- **Unstructured data**: Data in Redshift must be structured by a defined schema.
- **BLOB data**: to store binary large object (BLOB) files you might want to store the data in S3 and reference its location in Amazon Redshift.