> *Bases de Datos 2, Resumen 1: Data Warehousing on AWS*
> *Miguel Ku Liang - 2019061913*

Introduction

An enterprise must:

- Store every relevant data point about their business
- Give data access to everyone who needs it
- Have the ability to analyze the data in different ways
- Distill the data down to insights

Introducing Amazon Redshift

Cloud data warehouses like Amazon Redshift changed how enterprises think about data warehousing by dramatically lowering the cost and effort from deploying data warehouse systems without affecting its performance.

Amazon Redshift is a data warehousing solution that makes it simple and cost-effective to analyze large volumes of data using business intelligence tools.

Modern Analytics and Data Warehousing Architecture

Data warehouses are optimized for batched write operations and reading high volumes of data. **Online Transaction Processing databases** are optimized for continuous write operations and high volumes of small read operations. Data warehouses generally employ denormalized schemas because of high data throughput requirements while OLTP databases employ highly normalized schemas that are more suited for high transaction throughput requirements.

To benefit from using a data warehouse as a separate data store with a source OLTP, you should build an efficient **data pipeline**. The pipeline extracts data from the source system, converts it into a schema for data warehousing and loads it into the data warehouse.

AWS Analytics Services

This services help enterprises quickly convert their data to answers by providing analytics services. **Lake Formation** provides on-demand access to specific resources that fit the requirements of each analytics workload. AWS provides many analytic services that are integrated with the infrastructure layers.

Analytics Architecture

Analytics pipelines are made to process large volumes of data from different sources. The pipelines collect data, store data, process data, analyze and visualize data.

Data Collection

AWS provides solutions for data storage for different types of data:

- Transactional Data: this data is stored in relational database management systems or NoSQL database systems. A **NoSQL** database is used when data cannot fit into a defined schema or the schema changes often. A **RDBMS** is used when transactions happen across many table rows and queries need complex joins.

- Log Data: logs help fix issues, conduct audits and perform analytics from the information that is stored in it.
- Streaming Data: this data is collected, stored and processed from web applications, mobile devices and software apps and services.
- Internet of Things Data: AWS IoT is used to leverage AWS services to build apps that gather, process, analyze and act on IoT data.

Data Processing

Analyzing the data collected can help grow the business. There are two types of processing workflows:

- Batch Processing: **Extract Transform Load** is the process of pulling data from the sources to load into data warehousing systems. **Extract Load Transform** loads the extracted data into the target system. Transformation occurs after the data is loaded into the data warehouse. **Online Analytical Processing** store data in multidimensional schemas. It can extract data and spot trends on multiple dimensions. Online analytic processing uses batch processing.
- Real-Time Processing: when processing data sequentially and incrementally on a record-by-record basis, this processing is called real-time processing. OLTP uses real-time processing.

Data Storage

A **lake house** is an architectural pattern that combines the best elements of data warehouses and data lakes. It can query data across your data warehouse, data lake and operational databases to gain faster and deeper insights. A **data warehouse** can run fast analytics on large volumes of data and find hidden patterns in the data using BI tools. A **data mart** is a data warehouse focused on a specific functional area or subject matter.

Analysis and Visualization

You can analyze data using the tools used for processing data. Amazon Redshift works well with third-party BI solutions. AWS offers services to implement an end-to-end analytics platform.

Data Warehouse Technology Options

- Row-Oriented Databases: they store whole rows in a physical block. High performance for read operations is achieved through secondary indexes.
- Column-Oriented Databases: they organize each column in its own set of physical blocks. This allows them to be more input/output efficient for read-only queries because they only read the columns accessed by the query.
- Massively Parallel Processing Architectures: it can use all the resources available in the cluster for processing data, increasing performance of data warehouses. MPP data warehouses can improve performance by adding more nodes to the cluster.

Amazon Redshift Deep Dive

Amazon Redshift offers key benefits for performant, cost-effective data warehousing. It delivers fast query and I/O performance for any data size by using columnar storage and parallelizing and distributing queries across multiple nodes.

Integration with Data Lake

Amazon Redshift provides a feature called **Redshift Spectrum** that makes it easier to query data and write data to the data lake in open file formats.

Performance

Amazon Redshift offers features to achieve an industry-leading performance. This includes:

- High performing hardware
- Advanced Query Accelerator
- Efficient storage and high-performance query processing
- Materialized views
- Auto workload management to maximize throughput and performance
- Result caching

Durability and Availability

Amazon Redshift automatically detects and replaces any failed node in the data warehouse cluster. It makes the replaced node available immediately and loads the most frequently accessed data first.

Elasticity and Scalability

Amazon Redshift gets the elasticity and scalability needed for the data warehousing workloads. It provides two forms of compute elasticity:

- Elastic resize
- Concurrency Scaling

Amazon Redshift Managed Storage enables to scale and pay for compute and storage independently to size the cluster based on the compute needs.

Operations

Ideal Usage Patterns

Enterprises use Amazon Redshift to:

- Run 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

Amazon Redshift is not suited for:

- OLTP
- Unstructured data
- BLOB data