> Bases de Datos 2, Resumen 2
> Miguel Ku Liang - 2019061913

Introducing Amazon Redshift

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.

- AWS Analytics Services: This services help enterprises quickly convert their data to answers by providing analytics services.
- Analytics Architecture: 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.
        - 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: 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. Online Analytical Processing store data in multidimensional schemas.
        - Real-Time Processing: when processing data sequentially and incrementally on a record-by-record basis.
    - Data Storage: A lake house is an architectural pattern that combines the best elements of data warehouses and data lakes. 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.
    - Analytics pipeline with AWS services: 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.
- Column-Oriented Databases: they organize each column in its own set of physical blocks.
- Massively Parallel Processing Architectures: it can use all the resources available in the cluster for processing data, increasing performance of data warehouses.

Amazon Redshift Deep Dive

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

- Integration with Data Lake: Amazon Redshift provides a feature called Redshift Spectrum.
- 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 and result caching.
- Durability and Availability: Amazon Redshift automatically detects and replaces any failed node in the data warehouse cluster.
- Elasticity and Scalability: Amazon Redshift provides two forms of compute elasticity: elastic resize and 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

Amazon Redshift automates operational tasks like cluster performance and cost optimization.

- Amazon Redshift Advisor: It offers recommendations about changes to make by analyzing workload and usage metrics.
- Interfaces: Amazon Redshift has custom Java Database Connectivity and Open Database Connectivity drivers. It also provides a built-in Query Editor in the web console.
- Security: to provide data security, run Amazon Redshift inside a virtual private cloud based on the Amazon Virtual Private Cloud service.
- Cost Model: Amazon Redshift requires no long-term commitments or upfront costs.
- 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 and measure clinical quality, operation efficiency and financial performance in health care
- Anti-Patterns: Amazon Redshift is not suited for OLTP, unstructured data and BLOB data