Summary 1

> *Gabriela Gutiérrez Valverde - 2019024089*

> *Book: Data Warehousing on AWS (2021)*

Data enterprise:

- Store relevant data.
- Access needed.
- Analyze the data.
- Data insights.

Traditional data warehouse architecture challenges:

- Difficult to scale.
- High overhead costs for administration.
- Costly and complex to access, refine, and join data from different sources.
- Cannot separate infrequently used and frequently used data.
- Limit the number of users and amount of accessible data.

Amazon Redshift

Dramatically lowers the cost and effort associated with deploying data warehouse systems, without compromising on features, scale, and performance. Fast, fully managed, petabyte-scale data warehousing solution. Simple and cost-effective to analyze large volumes of data.

Modern Analytics and Data Warehousing Architecture

Data warehouses: optimized for batched write operations and reading high volumes of data.

Online Transaction Processing (OLTP) databases: optimized for continuous write operations and high volumes of small read operations.

AWS Analytics Services

- Easy path to build data lakes and warehouses.
- Secure cloud storage, compute, and network infrastructure
- Analytics stack
- Best performance, the most scalability, and lowest cost for analytics.

Analytics Architecture

Designed to handle large volumes of incoming streams of data. Stages:

1. Collect data
2. Store the data
3. Process the data
4. Analyze and visualize the data

Data Collection

Transactional Data:

- **NoSQL** suitable when the data is not well-structured to fit into a defined schema.

- **Relational Database Management Systems** suitable when transactions happen across multiple table rows and the queries require complex joins.

Log Data: Capturing system-generated logs: 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 generate streaming data that needs to be collected, stored, and processed continuously.

IoT Data: Devices and sensors around the world send messages continuosly

Data Processing

Batch Processing:

- **Extract Transform Load (ETL)** pulling data from multiple sources to load into data warehousing systems.
- **Extract Load Transform (ELT)** extracted data is loaded into the target system first.
- **Online Analytical Processing (OLAP)** store aggregated historical data in multidimensional schemas.

Real-Time Processing

Record-by-record basis, or over sliding time windows. Real-time processing requires highly concurrent and scalable processing layer.

Data Storage

- **Lake house** enable you to query data across your data warehouse, data lake, and operational databases to gain faster and deeper insights that are not possible otherwise.
- **Data warehouse** run fast analytics on large volumes of data and unearth patters hidden in your data by leveraging BI tools.
- **Data mart** simple form of data warehouse focused on specific functional area or subject matter.

Data Warehouse Technology Options

Row-Oriented Databases

Typically store whole rows in a physical block. High performance for read operations is achieved through secondary indexes. **Optimize techniques:**

- Building materialized views
- Creating pre-aggregated rollup tables
- Building indexes on every possible predicate combination
- Implementing data partitioning to leverage partition pruning by query optimizer
- Performing index-based joins

Limited by the resources available on a single machine. Every query has to read through all the columns for all of the rows in the blocks.

Column-Oriented Databases

Organize each column in its own set of physical blocks instead of packing the whole rows into a block. More input/output (I/O) efficient for read only queries. Need less storage compared to a row-oriented database.

Amazon Redshift Deep Dive

Performance

- **High performing hardware** maximize speed for performance-intensive workloads.
- **AQUA** Advanced Query Accelerator, runs up to ten times faster than any other cloud data warehouse.
- **Efficient storage and high-performance query processing** columnar storage, data compression, and zone maps reduce the amount of I/O needed to perform queries.
- **Materialized views** enable you to achieve significantly faster query performance for analytical workloads.
- **Auto workload management to maximize throughput and performance** machine learning to tune configuration to achieve high throughput and performance, even with varying workloads or concurrent user activity.
- **Result caching** deliver sub-second response times for repeated queries.

Durability and Availability

Attempts to maintain at least three copies of data: the original and replica on the compute nodes, and a backup in S3.

Elasticity and Scalability

- **Elastic resize** quickly resize your cluster by adding nodes to get the resources needed for demanding workloads, and to remove nodes when the job is complete to save cost.
- **Concurrency Scaling** support virtually unlimited concurrent users and concurrent queries, with consistently fast query performance.

Amazon Redshift Managed Storage

Scale and pay for compute and storage independently so you can size your cluster based only on your compute needs.

Operations

- Cluster Performance
- Cost Optimization

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** choose a relational database system or a NoSQL database.
- **Unstructured data** data in Amazon Redshift must be structured by a defined schema.
- **BLOB data** to store binary large object (BLOB) files, store the data in S3 and reference its location in Amazon Redshift.