Isaac Araya Solano | Carnet: 2018151703

Resumen #2

Introducing Amazon Redshift

In the past, when the data managed by a company grew, they had to choose between accepting slow performance or investing time, effort and money on an upgrade process. They were forced to update hardware sometimes and it made troublesome to have healthy relationships with database providers. Then cloud data warehouses like Amazon Redshift appeared and changed how enterprises think about data warehousing since it lowered costs and effort. Amazon Redshift is a fast, fully managed, petabyte-scale data warehousing solution that makes it simple and cost-effective to analyze large volumes of data using existing business intelligence (BI) tools.

Modern analytics and data warehousing architecture

Data typically flows into a data warehouse from transactional systems and other relational databases, and typically includes structured, semi-structured, and unstructured data.

Differences between Data warehouses and OLTP databases:

- Data warehouses are optimized for batched write operations and reading high volumes of data.
- OLTP databases are optimized for continuous write operations and high volumes of small read operations.

To get the benefits of using a data warehouse managed as a separate data store with your source OLTP or other source system, we recommend that you build an efficient data pipeline. It must extract data from the source system and converts its schema to be suitable for the warehouse and then load it to the warehouse.

AWS analytics services

This help enterprises convert their data to answers by providing analytics services between gload data warehouses to serverless data lakes. AWS helps you make everything work together by giving you:

- Facilitating building data lakes and warehouse
- A secure cloud storage, compute and network infrastructure adapted to specific needs
- An analytics stack with a set of tools
- The best performance, scalability and lowest cost.

Analytics architecture

Analytics pipelines are designed to handle large volumes of stream data from different sources.

Analytics pipeline has these stages:

- Collect data
- Store the data
- Process the data
- Analyze and visualize the data

Data collection

There are different types of data and AWS provides solutions for data storage for all of them.

Transactional Data

Usually stored in relational databases or NO SQL data bases.

Databases:

- Amazon DynamoDB(NO SQL)
- Amazon Aurora: compatible with MYSQL and PostgreSQL
- Amazon RDS

Log data

Amazon S3 is a popular storage solution for non-transactional data, such as log data used for analytics.

Streaming data

You can use Amazon MSK and Amazon Kisesis for this data.

IoT data

For connected devices you can use AWS IoT to interact easily with AWS cloud to process this data without having to manage any infraestructure.

Data Processing

Batch Processing

- Extract Transform Load (ETL): Is the process of pulling data from multiple sources to load into data warehousing systems. The data extracted is cleansed, enriched, transformed and loaded into a data warehouse.
- Extract Load Transform (ELT): Is a variant of ETL where the data is loaded into the target system first.
- Online Analytical Processing (OLAP): Storage aggregated historical data in multidimensional schemas.

Real-time Processing

You can process the data secuentially on a record-by-record basis or sliding time windows. Then use the processed data for a wide variety of analytics. It requires a highly concurrent and scalable processing layer.

Data Storage

- Lake house: is an architectural pattern that combines the best elements of data warehouses and data lakes.
- Data warehouse: using data warehouses, you can run fast analytics on large volumes of data and unearth patterns hidden in the data.
- Data mart: is a simple form of data warehouse focused on a specific functional area or subject matter.

Analysis and visualization

You need the right tools to analyze and visualize the process data. Tools such as MySQL Workbench, Amazon Redshift, Tableau and MicroStrategy. Amazon Quicksight is also a fast and cloud-powered solution. Amazon offers integration with Amazon Redshift, Amazon S3 and Amazon RDS.

Data warehouse technology options

Row-oriented databases

Row-oriented databases store whole rows in physical blocks and use secondary indexes to achieve high performance for read operations. They are typically used for transactional processing rather than analytics. Developers use various techniques to optimize performance for data warehousing, such as building materialized views and pre-aggregated rollup tables, implementing data partitioning, and performing index-based joins. Traditional row-based data stores are limited by resources, but data marts can alleviate this by using functional sharding. However, in a row-based data warehouse, every query has to read through all columns for all rows, causing a significant performance bottleneck, especially when the tables have more columns than the queries use.

Row-oriented databases

Column-oriented databases organize each column in its own set of physical blocks, making them more I/O efficient for read-only queries as they only have to read the accessed columns from disk or memory. This makes them a better choice for data warehousing than row-oriented databases. In contrast, row-oriented databases pack whole rows into a block.

Massively Parallel Processing (MPP) architectures

An MPP architecture enables you to use all the resources available in the cluster for processing data, which dramatically increases performance of petabyte scale data warehouses.

Amazon Redshift deep dive

Amazon Redshift is a fast and cost-effective columnar MPP technology for data warehousing. It offers efficient compression, reduced I/O, and lower storage requirements. Most administrative tasks are automated, and you can build petabyte-scale data warehouses in minutes. You can also run exabytes-scale queries using Amazon Redshift Spectrum and scale compute and storage separately using RA3 nodes with Redshift Managed Storage.

Integration with data lake

Redshift Spectrum allows querying and writing data back to open file formats in S3 using ANSI SQL. You can export data using Redshift UNLOAD command with Parquet as the file format. You can query data in S3 by creating an external schema or table. Writing data is possible with CREATE EXTERNAL TABLE AS SELECT or INSERT INTO an external table. You can keep frequently accessed data in Redshift and up to exabytes of structured, semi-structured, and unstructured data in S3. Data exported from Redshift can be further analyzed with AWS services.

Performance

- High performing hardware: Amazon Redshift provides multiple node types to choose from, including the latest generation RA3 instances built on the AWS Nitro System. These instances have high bandwidth networking, performance similar to bare metal, and are ideal for performance-intensive workloads that require large compute capacity.
- AQUA (preview): Amazon Redshift's AQUA (Advanced Query Accelerator) is a hardware-accelerated cache that speeds up data-intensive tasks like filtering and aggregation by running them closer to the storage layer, allowing Redshift to run up to 10 times faster than other cloud data warehouses. It uses AWS-designed processors to accelerate queries in parallel across multiple nodes and automatically scales out to add more capacity as storage needs grow.
- Efficient storage and high-performance query processing: Amazon Redshift delivers fast query performance on datasets ranging in size from gigabytes to petabytes. Columnar storage, data compression, and zone maps reduce the amount of I/O needed to perform queries.
- Materialized views: Amazon Redshift materialized views enable you to achieve significantly faster query performance for analytical workloads such as dashboarding, queries from BI tools, and ELT data processing jobs.
- Auto workload management to maximize throughput and performance: Amazon Redshift uses machine learning to tune configuration to achieve high throughput and performance, even with varying workloads or concurrent user activity. You can set the priority of your most important queries, even when hundreds of queries are being submitted.
- Result caching: Amazon Redshift uses result caching to deliver sub-second response times for repeated queries. Dashboard, visualization, and business intelligence tools that execute repeated queries experience a significant performance boost.

Durability and availability

Amazon Redshift automatically replaces failed nodes, loads frequently accessed data first, and maintains at least three copies of data for durability. It offers options for creating a Multi-AZ setup and setting up disaster recovery with incremental snapshots. You can keep copies of your backups in multiple AWS Regions.

Elasticity and scalability

Amazon Redshift offers elasticity and scalability for data warehousing workloads, allowing users to scale compute and storage independently and pay only for what they use. The service provides two forms of compute elasticity, including elastic resize and Concurrency Scaling. Elastic resize enables users to quickly resize their cluster by adding or removing nodes, while Concurrency Scaling supports virtually unlimited concurrent users and queries. Users can start with a single 160 GB node and scale up to multiple petabytes of compressed user data using many nodes.

Amazon Redshift managed storage

Amazon Redshift managed storage enables you to scale and pay for compute and storage independently so you can size your cluster based only on your compute needs.

Operations

As a managed service, Amazon Redshift completely automates many operational tasks, including: - Cluster Performance - Cost Optimization

Amazon Redshift Advisor

Amazon Redshift Advisor offers you recommendations about changes to make to improve the performance and lower costs.

Interfaces

Amazon Redshift offers custom JDBC and ODBC drivers, as well as standard PostgreSQL JDBC and ODBC drivers, allowing for a wide range of SQL client compatibility. The platform provides a built-in Query Editor in the web console, as well as validated integrations with popular BI and ETL vendors. Users can easily load streaming data into Amazon Redshift using Amazon Kinesis Data Firehose for near real-time analytics. Compute, memory, storage, and read/write traffic metrics can be found in the console or via the Amazon CloudWatch API.

Security

Amazon Redshift provides security features to ensure data security. It can be run inside a virtual private cloud using Amazon VPC, with firewall rules and SSL-enabled connections between the client application and data warehouse cluster. Enhanced VPC Routing can manage data flow, and AWS CloudTrail can audit API calls. Amazon Redshift also supports encryption with hardware-accelerated AES-256 encryption, and can manage keys through AWS KMS. Access management is controlled by managing user access, privileges, and column-level grants and revokes. Multiple authentication methods such as AWS IAM, federated authentication, and MFA are also provided.

Cost model

Amazon Redshift's pricing model is flexible, with charges based on the size and number of nodes in your cluster. There are no upfront costs or long-term commitments. Amazon Redshift-managed storage (RMS) with an RA3 instance is billed separately for the amount of compute and RMS used. Backup storage up to 100% of your provisioned storage is provided at no additional charge, and there is no data transfer charge between S3 and Amazon Redshift.

Ideal usage patterns

Enterprises use Amazon Redshift to do the following:

- 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

Amazon Redshift is not suitable for OLTP, unstructured data, or BLOB data. For these use cases, other database systems such as Amazon Aurora or DynamoDB might be more appropriate. If BLOB data needs to be stored, it's recommended to use S3 and reference the location in Amazon Redshift.