Data Warehousing on AWS

Introduction to data Warehousing on AWS

Introduction

Data is an enterprise’s most valuable asset

- 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

They cannot separate cold and warm data, which results in bloated costs and wasted capacity. They limit the number of users and the amount of accessible data, which leads to anti-democratization of data.

Introducing Amazon Redshift

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 tools. Spectrum feature. With this setup, you can query data directly from files on Amazon S3 for as low as $5 per terabyte of data scanned.

Modern Analytics and Data Warehousing Architecture

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

AWS Analytics Services

AWS analytics services help enterprises quickly convert their data to answers by providing mature and integrated analytics services, ranging from cloud data warehouses to serverless data lakes. Many enterprises choose cloud data lakes and cloud data warehouses as the foundation for their data and analytics architectures. AWS is focused on helping customers build and secure data lakes and data warehouses in the cloud within days, not months. The data is curated and cataloged, already prepared for any type of analytics. This enables you to take advantage of features like intelligent tiering and Amazon Elastic Compute Cloud spot instances, to reduce cost and run analytics faster. AWS helps you in this process wiwth:

- An easy path to build data lakes and data warehouses.
- A secure cloud storage.
- A fully integrated analytics stack with a mature set of analytics tools.
- The best performance, the most scalability, and the lowest cost for analytics.

Data Collection

- [Transactional Data] Transactional data, such as e-commerce purchase transactions and financial transactions, is typically stored in relational database management systems or NoSQL database systems. A NoSQL database is suitable when the data is not well-structured to fit into a defined schema, or when the schema changes often. An RDBMS solution is suitable when transactions happen across multiple table rows and the queries require complex joins.

Service enable you to implement an SQL-based relational database solution for your application

    - Amazon Aurora is a MySQL and PostgreSQL-compatible relational database built for the cloud
    - Amazon RDS is a service that enables you to easily set up, operate, and scale relational databases on the cloud.

- [Log Data] Amazon S3 is a popular storage solution for non-transactional data, such as log data, that is used for analytics.
- [Streaming Data] Using Amazon Kinesis services, you can do that simply and at a low cost. Apache Kafka to process streaming data.
- [IoT Data] Enterprises today need to capture this data and derive intelligence from it.

Data Processing

The collection process provides data that potentially has useful information. The best practice to gather this intelligence is to load your raw data into a data warehouse to perform further analysis. OLTP system, you keep the data processing from affecting your OLTP workload. First, let's look at what is involved in batch processing.

- [Batch Processing]
    - Extract Transform Load (ETL) -> Pulling data from multiple sources to load into data warehousing systems.
    - Extract Load Transform (ELT) -> ETL variant, where the extracted data is loaded into the target system first.
    - Online Analytical Processing (OLAP) -> systems store aggregated historical data in multidimensional schemas.
- [Real-Time Processing] Amazon MSK as solutions to capture and store streaming data. You can process this data sequentially and incrementally on a record-by-record basis, or over sliding time windows. Use the processed data for a wide variety of analytics, including correlations, aggregations, filtering, and sampling. To process streaming data in real-time, use AWS Lambda. Lambda can process the data directly from AWS IoT or Amazon Kinesis Data Streams.
- [Data Storage] You can store your data in a lake house, data warehouse, or data mart, each one of those have their differences and atributes that makes them better in specific situations.
- [Analysis and Visualization] After processing the data and making it available for further analysis, you need the right tools to analyze and visualize the processed data.

Data Warehouse Technology Options

- Row-Oriented Databases -> These systems have been traditionally used for data warehousing, but they are better suited for transactional processing than for analytics.
- Column-oriented databases -> This functionality allows them to be more input/output (I/O) efficient for read-only queries.

- Massively Parallel Processing (MPP) Architectures -> MPP data warehouses allow you improve performance by simply adding more nodes to the cluster.

Amazon Redshift Deep Dive

Amazon Redshift delivers fast query and I/O performance for virtually any data size by using columnar storage, and by parallelizing and distributing queries across multiple nodes.

- [Integration with Data Lake] -> Amazon Redshift provides a feature called [Redshift Spectrum] that makes it easier to both query data and write data back to your data lake in open file formats
- [Performance] -> Amazon Redshift offers multiple features to achieve this superior performance, including:
  - High performing hardware
  - AQUA
  - Efficient storage and high-performance query processing
  - Materialized views
  - Auto workload management to maximize throughput and performance
  - Result caching
- [Durability and Availability] -> It makes your replacement node available immediately, and loads your most frequently accessed data first so you can resume querying your data as quickly as possible.
- [Elasticity and Scalability] -> Can easily run non-uniform and unpredictable data warehousing workloads.
  - Elastic resize
  - Concurrency Scaling

Operations:

As a managed service, Amazon Redshift completely automates many operational tasks, including:

- [Ideal Usage Patterns] -> Amazon Redshift is ideal for OLAP using your existing BI tools.
  - 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 ideally suited for the following usage patterns:
  - OLTP : Amazon Redshift is designed for data warehousing workloads delivering extremely fast and inexpensive analytic capabilities.
  - Unstructured data : Data in Amazon Redshift must be structured.
  - BLOB data : In this case is better to store the data in S3 and reference its location in Amazon Redshift.