Data Warehouse Architecture Design

A lakehouse is a new data platform architecture paradigm that combines the best features of data lakes and data warehouses.

Required Data Platform Stack(Services)

S3, AWS EMR, AWS Glue, AWS Redshift, Airflow, Athena

Read Data from Source using Pyspark running on AWS EMR to create Delta Lake Table stored in s3.

The data lake house will be structured using the Medallion Architecture where we will have 3 layers for data storage

Bronze Layer - in the layer we will store the raw data gotten from data sources without any transformations. This layer would allow for data versioning, historical data storage. After the initial load, new data for this layer would only read recently updated records from the source and perform an UPSERT on the data already stored. This layer will be updated using incremental loads (Merge statement).

Silver Layer - this layer stores cleansed data. In this layer basic data transformations and joins would occur such as dropping unused columns, merging customer information into a single table etc

Gold Layer - this is the layer that would hold data marts that would be provided to end users (analytics, management etc).

The gold layer would be broken down into THREE major categories: analytics, monitoring and historical data sources.

Analytics Layer -

Granularity: lowest

Refresh rate: Once end of day

Monitoring Layer -

Granularity: aggregated/lowest (depending on request)

Refresh rate: once an hour between 8 am - 10 pm

Historical Layer

Granularity: aggregated

Refresh rate: Once a week/ON demand

Delivery of Data to BI tools

For Data usage on powerBi we can create a direct connection between the data stored in the gold layer and powerBI, using delta sharing.

For BI tools that may not support delta sharing, the gold layer data here would be written to Amazon redshift data warehouse and a direct connection to Redshift would be provided to Redshift from the BI tool.

Appendix

Further Readings

- Delta Lake Acid Transactions
- Medallion Architecture

Estimated Cost(excluding airflow):

- Cost Estimate

Architecture Diagram