AWS Data Engineering

Capstone Project Synopsis

ETL pipelines to feed data to Data Marts to drive Customer Analytics

Industry Background :

ECommerce is one of the most successful businesses of modern times. The likes of Amazon, Flipkart, Myntra etc. have made multi-billion dollar businesses by making the whole process of searching, buying and returning of products extremely simple for the end user.These companies heavily rely on technology and specifically data to personalize the shopping experience of their customers. Retail over the years have become one of the most matured industries and has seen many applications of data driven decision making to squeeze in as much margin as possible and at the same time engaging the customers so well

Ecommerce has taken this to the next level by virtualizing the whole shopping experience through apps. As an industry heavily driven by data, there are a lots of interesting problem statements which ECommerce companies work on

Problem Statement :

GlobalMart is one of the leading E-Commerce giants with presence in the North America and Europe region. It has presence across 120 markets and primarily deals with 3 lines of business : Technology, Office Supplies and Furniture. With an increase in customers and expansion in geography, GlobalMart has developed tie-ups with a number of local vendors to help them deliver their products to the end customers.

The Vendor Management Team at Global Mart is tasked with the responsibility of ensuring that the vendors deliver the products on time and without any damage or wrong deliveries. The Business Analytics team prepares a number of reports to be shared with the VP, Vendor Operations on a week-on-week basis. Some of the metrics to be covered would be :

- On Time Delivery %
- Delivery Delay
- % of Damaged or Wrong Deliveries
- Average Customer Rating

Source Data stores :

- Customer and Order delivery details on **Amazon RDS**
- Product features and pricing details on **Amazon DYnamoDB**
- Customer Reviews data on **Amazon S3**

Destination Data Store :

- Storing Fact table in **Amazon Redshift**

ETL Tools :

- Amazon **GLUE ETL** + **Zeppelin Notebooks**
- Amazon **EMR** + **PySpark Notebooks**

Downstream Visualization:

- QuickSights

Tasks :

- Create a **Service Analysis** template and **Survey** template.
- Create an **Architecture Solution** which will represent the data movement of an **End-to-End ETL** solution for **batch processing.**
- **Copy data** from **S3** to base systems (**RDS** and **DynamoDB**) using **Glue, EMR, Data pipeline.**
- Perform **ETL** using **Glue** by extracting data from base systems and **stage fact tables** in **S3.**
- **Loading Fact tables** from staged s3 bucket to **Redshift** using **Copy Query/AWS data pipeline.**
- **Building Dashboards** using Amazon QuickSights

Project Deliverables :

- Architecture diagram of the ETL solution
- Data model and tables in RDS and DynamoDB
- Successful ETL by loading fact tables in Redshift.
- A dashboard explaining the story of KPI's built as a part of fact_tables.