AMRITA VISHWA VIDYAPEETHAM | AHEAD Online

21CSA523A Data Engineering for AI

Mid Review Report

Project Title

Data Engineering – Amazon Sales Dataset Analysis

Prepared by

Name : Ramya Mercy Rajan
Roll Number : AA.SC.P2MCA2107434
Degree & Semester : MCA (Artificial Intelligence) Semester 3
Email : ramyamr_mca2107434@ahead.students.amrita.edu

March 2023

Tools Used

1. Amazon S3
2. Crawler
3. AWS Glue Data Catalog
4. Athena
5. QuickSight

Dataset details

https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset

Plan to Execute

1. Load dataset to Amazon S3 Bucket
2. A Crawler will be created which will connect to S3 Datastore which will infer or determine the structure of the csv file using built-in classifier.
3. It will create the metadata table in the AWS Glue Data Catalog.
4. After populating the Data Catalog and ETL job will be created to transform the csv into Parquet file. The data source of the ETL Job will be AWS Glue Data Catalog table.
5. As a part of transformation, we will apply the mappings.
6. Post transformation the data will be stored in S3 bucket as a Parquet file.
7. Querying of Parquet file will be done with the help of Amazon Athena.
8. Visualization will be done with the help of Amazon QuickSights.

Objective

Amazon Sales Dataset have been downloaded from Kaggle in order to perform Extract , Transform and Load process in AWS Cloud Storage and also to perform analysis using Amazon Athena and Amazon QuickSight , a cloud-scale business intelligence (BI) service that can be used to deliver easy-to-understand insights.

Block Diagram