AWS PROJECT

RAVEENA PRABAKARAN

TIGER ANALYTICS

**Batch Time Analysis of Transactional Data**

Project 1

DESCRIPTION

Lenodo is a multinational e-commerce organization that sells products directly to consumers. The database administrator exports the data every night in a CSV file, but this export functionality is unused. Lenodo wants to use this data to uncover insights about the most-sold item and the countries where customers have bought this item.

You are a data analytics consultant, and you're asked to provide valuable insights and statistics across products, brands, categories, segments to the marketing, product, sales, and procurement teams and inform them about which product has the highest amount of sales and which product and its marketing needs the most improvement. These statistics will help to run effective digital marketing campaigns. The scope of this project is limited to data engineering and analysis.

**Objective:**

To use AWS Big Data stack for data engineering to analyze transactions, uncover patterns, and share actionable insights

**Steps to perform:**

1. Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)
2. Create a crawler to crawl the CSV data and generate a metadata catalog
3. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries
4. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena
5. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

AWS SERVICES INVOLVED IN THE PROJECT:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Creating Role For Glue :-

A screenshot of a computer

Description automatically generated

1. Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)

Creating S3 Bucket :-

A screenshot of a computer

Description automatically generated

Create Source Folder and upload csv file :-

A screenshot of a computer

Description automatically generated

1. Create a crawler to crawl the CSV data and generate a metadata catalog

Creating Crawler for csv Data:-

Graphical user interface, text

Description automatically generated

Table for csv data :-

A screenshot of a computer

Description automatically generated

1. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries

Create destination folder:-

A screenshot of a computer

Description automatically generated

Create Source in ETL :-

A screenshot of a computer

Description automatically generated

Create Destination in ETL :-

A screenshot of a computer

Description automatically generated

Fill Job Details ( Attach the Created IAM Role) :-

Graphical user interface, application, Word

Description automatically generated

Save the Job:

A screenshot of a computer

Description automatically generated

Job Successfully Runned:-

A screenshot of a computer

Description automatically generated

Output in Destination Folder:

A screenshot of a computer

Description automatically generated

1. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena

Creating Crawler For parquet data

A screenshot of a computer

Description automatically generated

Run the Crawler

Graphical user interface, text, application

Description automatically generated

Table Created For Parquet:

A screenshot of a computer

Description automatically generated

In Athena

Select \* from destination ;

Graphical user interface, text, application

Description automatically generated

1. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

Query:

select sum(CAST(quantity AS INT)) as Quantity\_Sold , description as Best\_Selling\_Item from destination group by description order by 1 desc limit 1);

Graphical user interface, text, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Query:

select distinct(country) from destination where description in (select description as Best\_Selling\_Item from destination group by description having sum(CAST(quantity AS INT)) = ( select sum(CAST(quantity AS INT)) from destination group by description order by 1 desc limit 1));

Graphical user interface, text

Description automatically generated

Graphical user interface, application, table

Description automatically generated