ASSIGNMENT 04 - Data Product - Starburst

Trainer: Udayakumar M

Data Engineering

Assignment Overview

Upload Supply Chain Data to S3: Prepare and upload a CSV file with supply chain data to an S3 bucket.

Create a Catalog in Starburst Galaxy: Set up a Starburst Galaxy catalog pointing to the S3 bucket.

Query Data from Jupyter Notebook: Use Python in a Jupyter Notebook to query the data from Starburst Galaxy.

Prerequisites

Starburst Galaxy Account: Ensure you have access to Starburst Galaxy.

AWS S3 Bucket: Have an S3 bucket with necessary permissions.

Jupyter Notebook: Set up a Jupyter Notebook environment with Python.

Step 1: Download and Upload Supply Chain Data to S3

Upload CSV to S3: (https://www.kaggle.com/datasets/amirmotefaker/supply-chain-dataset)

Use AWS CLI or AWS Management Console to upload the CSV file:

supply_chain_data.csv

Step 2: Create a Catalog in Starburst Galaxy

Log in to Starburst Galaxy: Navigate to the Starburst Galaxy web console and log in.

Create a New Catalog:

Go to the "Catalogs" section.

Click "Add Catalog".

Choose "AWS S3" as the data source.

Fill in the necessary details:

Catalog Name: supply_chain_catalog

S3 Bucket Name: Your S3 bucket name.

IAM Role: Specify an IAM role with access to the S3 bucket.

Authentication Method: Choose appropriate method (e.g., access keys or IAM role).

Create Schema and Table:

Use Starburst Galaxy's SQL interface to create a schema and table:

Step 3: Access Starburst Galaxy Catalog from Jupyter Notebook

Install Required Libraries:

Ensure you have pystarburst and pandas installed in your Jupyter environment:

Connect to Starburst Galaxy and Query Data:

Create a Jupyter Notebook and write it as a Python code to connect and query data from Starburst Galaxy.

Step 4: Create a Bar chart using Matplotlib to display the count of Products group by Location

Summary

This assignment demonstrates how to:

Upload supply chain data in CSV format to an S3 bucket.

Create a catalog in Starburst Galaxy that points to the S3 bucket.

Access and query the data from Starburst Galaxy using Python in a Jupyter Notebook.

This workflow helps in managing and analyzing supply chain data efficiently using modern data tools and cloud services.