IntelliPaat

LAB 04 - Data Product - Starburst

To create a catalog for a Starburst Galaxy, store CSV and Apache Iceberg data in an S3 bucket, you can follow these steps. Below is a structured exercise that guides you through the entire process:

Prerequisites

Starburst Galaxy Account: Ensure you have an active Starburst Galaxy account.

AWS S3 Bucket: Have access to an AWS account with an S3 bucket created.

Permissions: Ensure you have the necessary permissions to read/write to the S3 bucket and access to the Starburst Galaxy.

Step-by-Step Exercise

Step 1: Set Up Starburst Galaxy Connection

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.

Configure the catalog with necessary details such as:

Catalog Name: e.g., s3_catalog

S3 Bucket Name: The name of your S3 bucket.

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

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

Step 2: Upload CSV Data to S3

Prepare Your CSV Data: Download a supply chain CSV file from Kaggle. For example, supply_chain_data.csv with some sample data:

Upload to S3:

Use AWS Management Console, AWS CLI, or an SDK to upload the CSV file to your S3 bucket.

Example using AWS CLI:

aws s3 cp sample_data.csv s3://your-bucket-name/path/to/supply_chain_data.csv

Step 3: Query CSV Data in Starburst Galaxy

Connect to the Catalog: Ensure your Starburst Galaxy is connected to the S3 catalog.

Create Schema and Table:

Using Starburst Galaxy's SQL interface, create a schema and table that points to the CSV data:

CREATE SCHEMA s3_catalog.supply_chain_schema.

CREATE TABLE s3_catalog.supply_chain_schema.supplychain (

Specify the columns according to the supply chain csv file

)

WITH (

external_location = 's3://your-bucket-name/path/to/supply_chain_data.csv',

format = 'CSV'

);

Query the Data:

Run a simple query to ensure data is accessible:

SELECT * FROM s3_catalog. supply_chain_schema.supplychain;

Step 4: Create and Store Apache Iceberg Table

Create Iceberg Table:

Using Starburst Galaxy, create an Iceberg table in your S3 bucket:

```
CREATE SCHEMA s3_catalog.iceberg_schema;

CREATE TABLE s3_catalog.iceberg_schema.iceberg_table (
   id INTEGER,
   name VARCHAR,
   value INTEGER
)
WITH (
   format = 'ICEBERG',
   location = 's3://your-bucket-name/path/to/iceberg_table/'
);
```

Insert Data into Iceberg Table:

Insert data from the CSV table or directly add new data:

```
INSERT INTO s3_catalog.iceberg_schema.iceberg_table
VALUES (1, 'alpha', 10), (2, 'beta', 20), (3, 'gamma', 30);
```

Query the Iceberg Data:

Run a query to verify data in the Iceberg table:

SELECT * FROM s3_catalog.iceberg_schema.iceberg_table;

Summary

By following these steps, you have:

Created a catalog in Starburst Galaxy connected to an S3 bucket.

Uploaded CSV data to the S3 bucket and created a table to query this data in Starburst Galaxy.

Created an Apache Iceberg table in the S3 bucket, inserted data, and queried it through Starburst Galaxy.

This exercise helps in understanding the integration between Starburst Galaxy, AWS S3, and the use of different data formats like CSV and Apache Iceberg.