Databricks: Flight Data Project (Bronze Layer Ingestion)

Data Understanding:

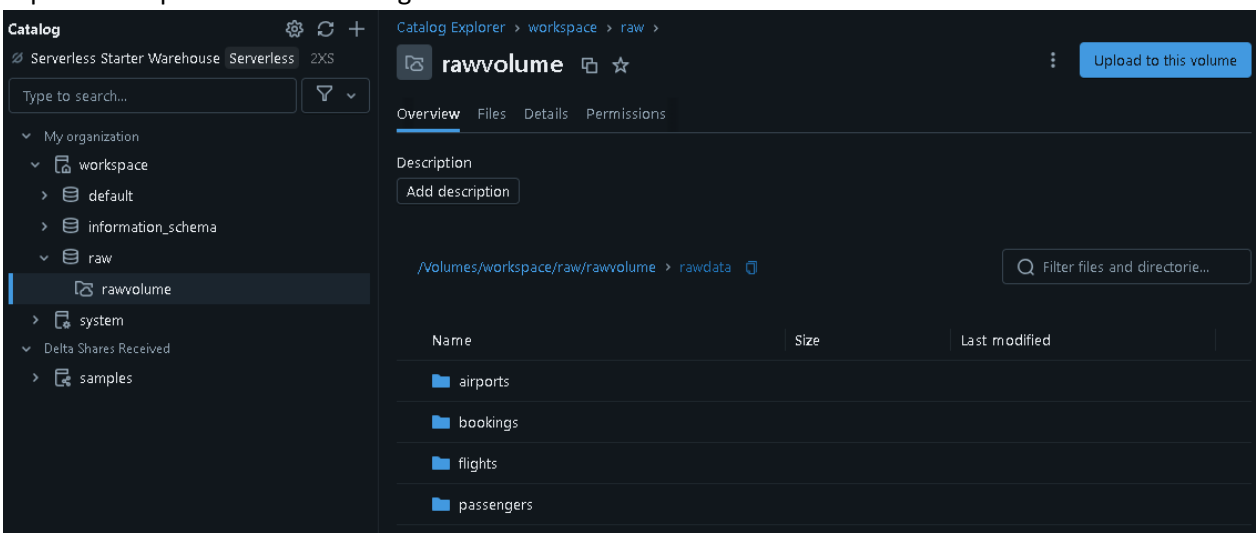
We have the following CSV files along with their data description:

1. Bookings Data (fact_bookings.csv) - Contains booking id, passenger id, flight id, airport id, amount and the booking date information.
2. Airport Data (dim_airports.csv) – Contains airport id, airport name, city and country details.
3. Flights Data (dim_flights.csv) – Contains flight id, airline, origin, destination, flight date details.
4. Passengers Data (dim_passengers.csv) – Contains passenger id, name, gender and nationality details.

Steps Followed:

Note: We are using Databricks Free Edition for this project

1. In our unity catalog, we create a schema named: raw (within the workspace catalog), from where we will be getting our csv files into the bronze layer using Autoloader. Raw schema is a managed one, hence will not be adding any external location while creating it.
2. Within this raw schema, we will create volume and folders within these volumes to stores our csv files (ex: airport folder to store our dim_airport.csv files). This will help us create our raw data layer. We will use the **Setup Notebook** to create the volume and folders.
   Expected output as shown in image below:



3. We will also create bronze, silver and gold layer schemas and volumes inside each schema.
4. We will now create a **BronzeLayer Notebook**, and use autoloader to load our data from the raw volume into our bronze volume, and this notebook will be parametrized to get each file from the raw volume and put it in separate folders in the bronze volume.
5. Now we will create a notebook called **SrcParamters**, which will contain the source parameter details, which can be passed through our BronzeLayer Notebook to automate the

bronze data ingestion from the raw layer. (SrcParamters will act as input for BronzeLayer Notebook)

6. Now in Databricks, we go to Workflows, and create a Job. Now we create a Task called **Parameters** using our SrcParamters Notebook. Then we click on add task (Name it as **Incremental Ingestion)**, and add our BronzeLayer Notebook

7. On our Incremental Ingestion task, we will select the Loop over this task option (same as ForEach activity in ADF) and it will get one parameter at a time from the Parameters task and run the notebook every time, there by loading the data for airports, passengers, flights and bookings into the respective volumes and folders of the bronze layer.