Name: Abhishek Obla Hema
Email: oh_abhishek@yahoo.com

Airflow Assignment-2 Documentation

This file details and describes all the attached files for the Airflow Assignment -2

Tools Used:

1. Python3 – Microsoft VScode
2. Apache Airflow
3. GCP services – DataProc/GCS
4. Apache Hive

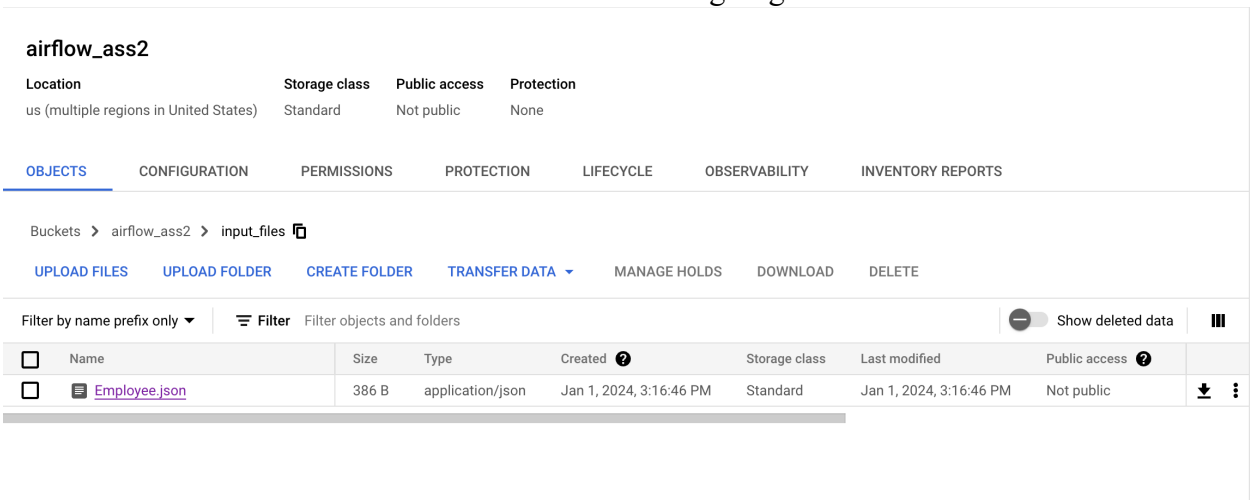
Files Attached:

1. Airflow_assignment_2.pdf – This file
2. Airflow_ass2_job.py – Airflow job that details the DAGs
3. Employee.json – JSON file used for the exercise

Process and File Descriptions:

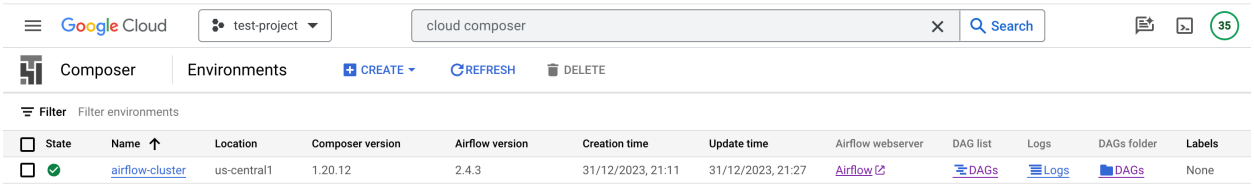
Step 1:

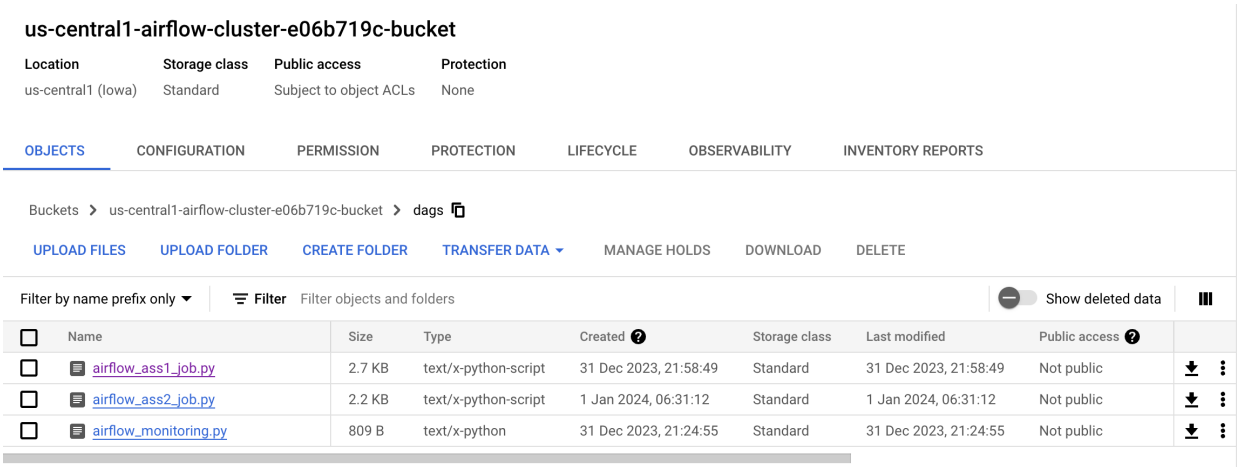
I created a bucket called 'airflow_ass2' and placed Employee.json under input_files folder. This file will be downloaded to the airflow local before being staged for an external hive table.



Step 2:

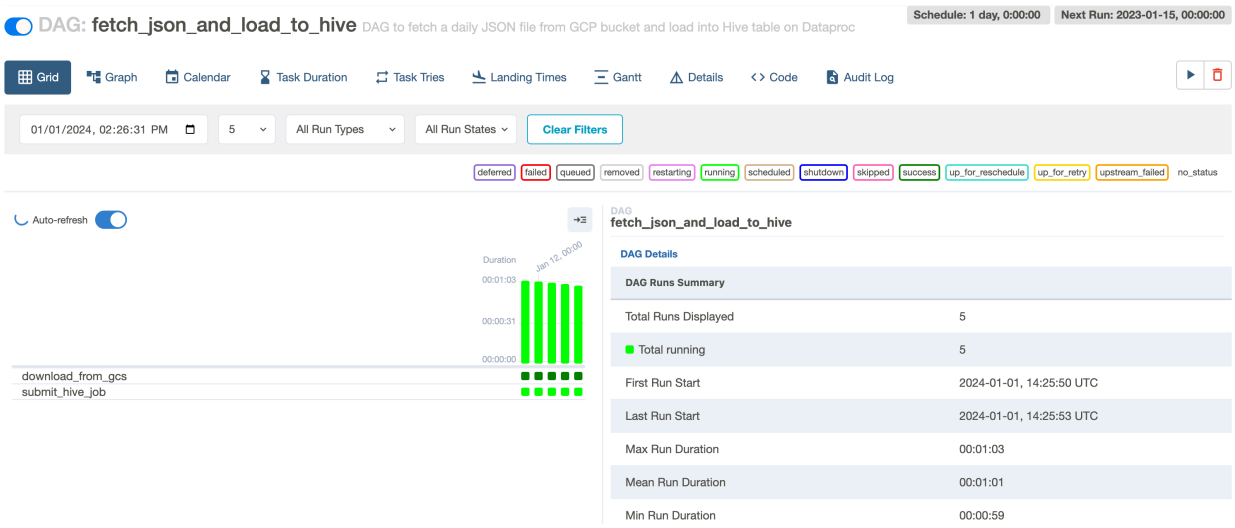
I create an airflow cluster and then proceeded to place the 'airflow_ass2_job.py' file in the DAG list so that it can be picked up by airflow





Step 3:

Now we can see the various stages in the DAG. The first task is about downloading the json file from the GCS bucket to the airflow local. It then stages this json file in a local directory on the airflow cluster itself over which external tables can be built.



Challenges

1. The HiveOperator is deprecated and hence had to use the new DataProcSubmitJobOperator
2. Learnt the hard way that file_names are case sensitive so make sure to refer to the path and files with correct case.
3. Difficult overall to troubleshoot the leading cause when airflow jobs fail