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

Airflow Assignment-1 Documentation

This file details and describes all the attached files for the Airflow Assignment -1

Tools Used:

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

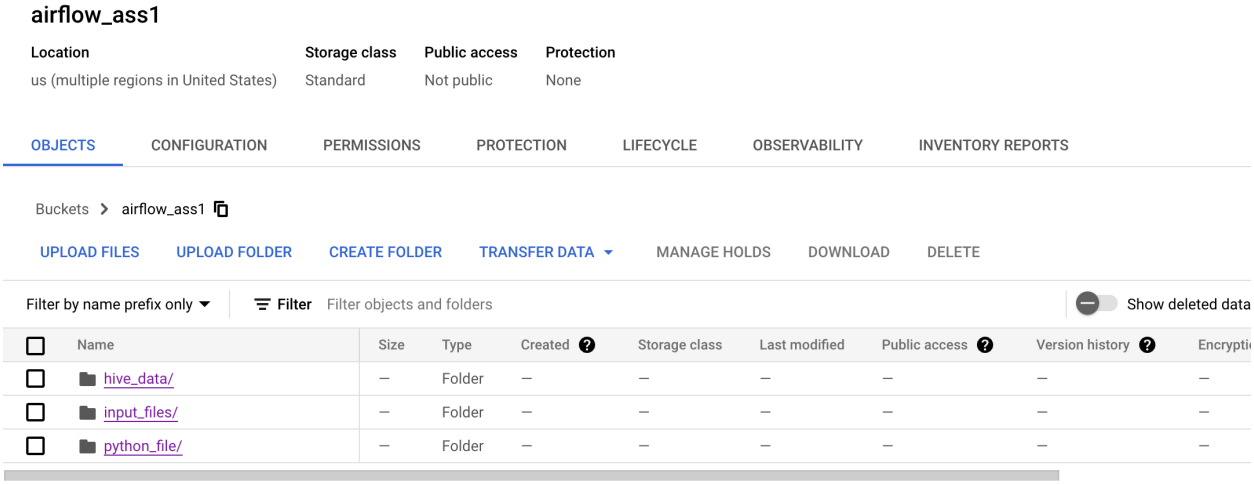
Files Attached:

1. Airflow_assignment_1.pdf – This file
2. Employee_batch.py – Pyspark job that filters employees with salary >=60,000 and stores it in a GCS location.
3. Airflow_ass1_job.py – Airflow job that details the DAGs
4. Employee.csv – Csv file used for the exercise

Process and File Descriptions:

Step 1:

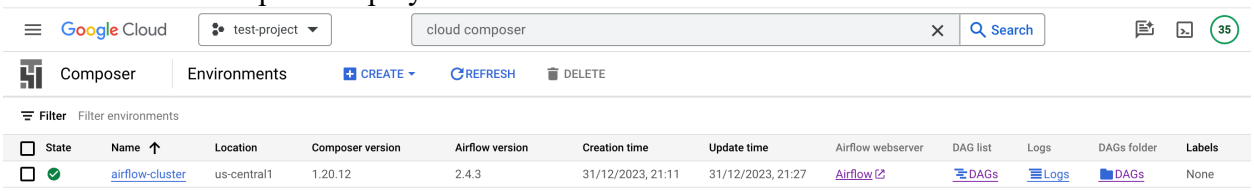
I created a bucket called 'airflow_ass1' and placed employee.csv under input_files folder. This file will be picked up by the spark job which is a part of the DAG

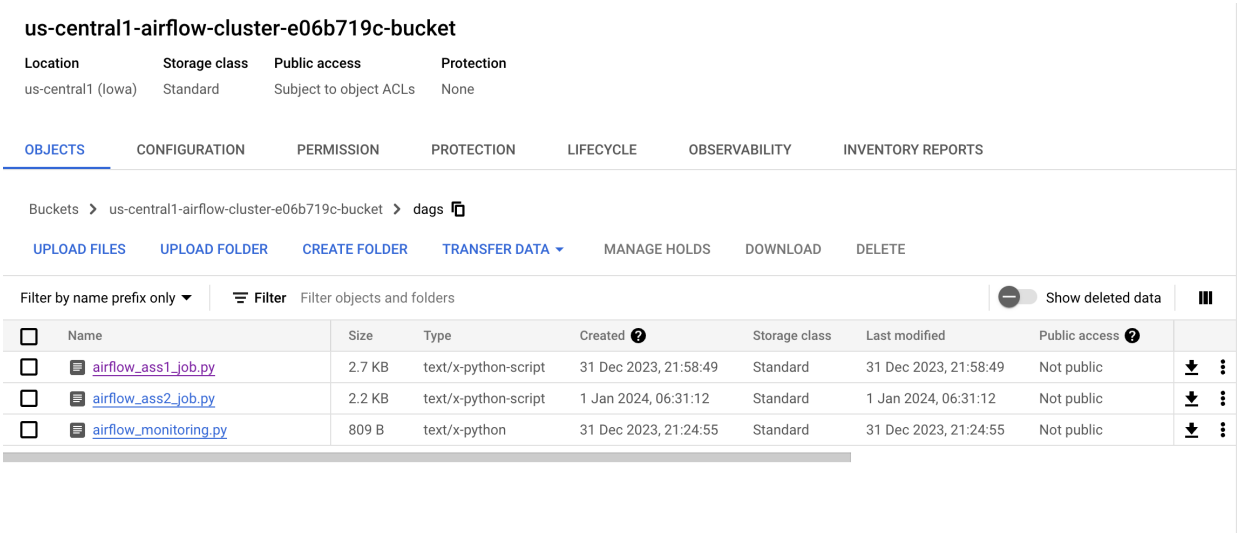


I also made sure to place the pyspark job 'employee_batch.py' in the python_file folder in the same GCS bucket.

Step 2:

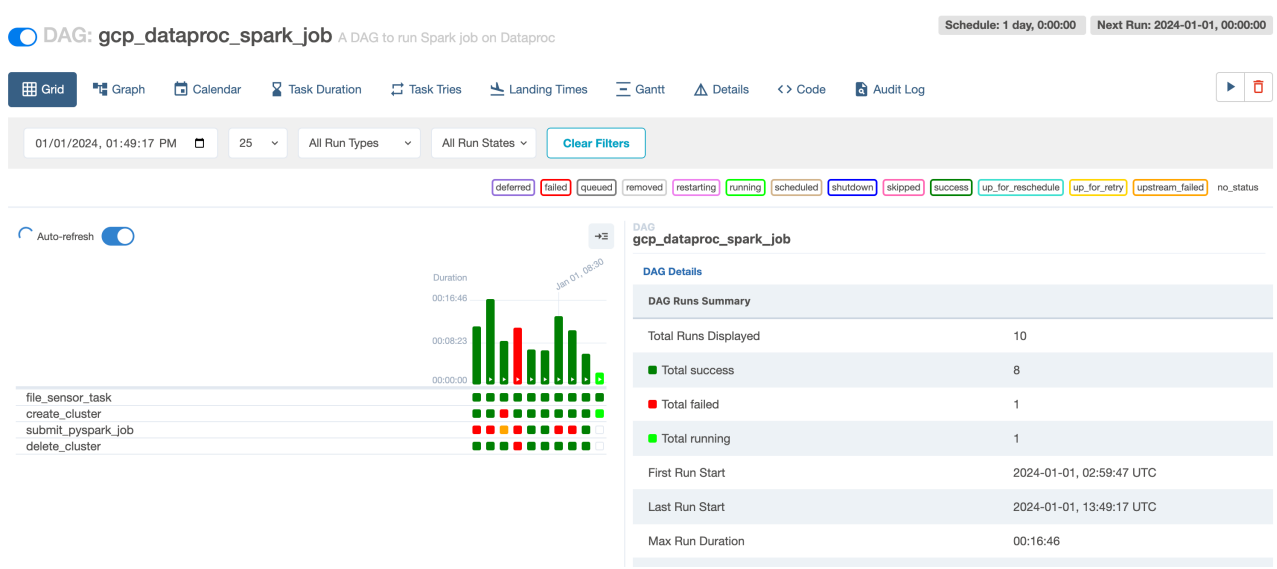
I create an airflow cluster and then proceeded to place the ‘airflow_ass1_job.py’ file in the DAG list so that it can be picked up by airflow

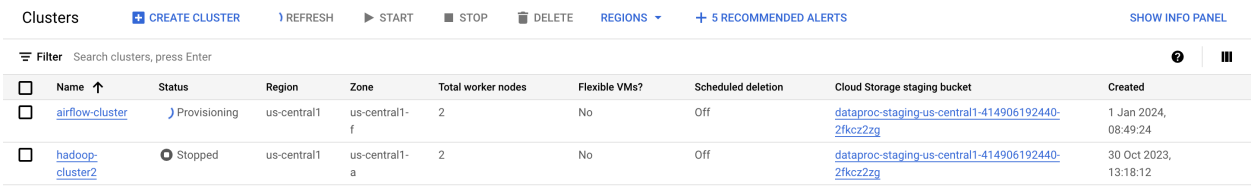


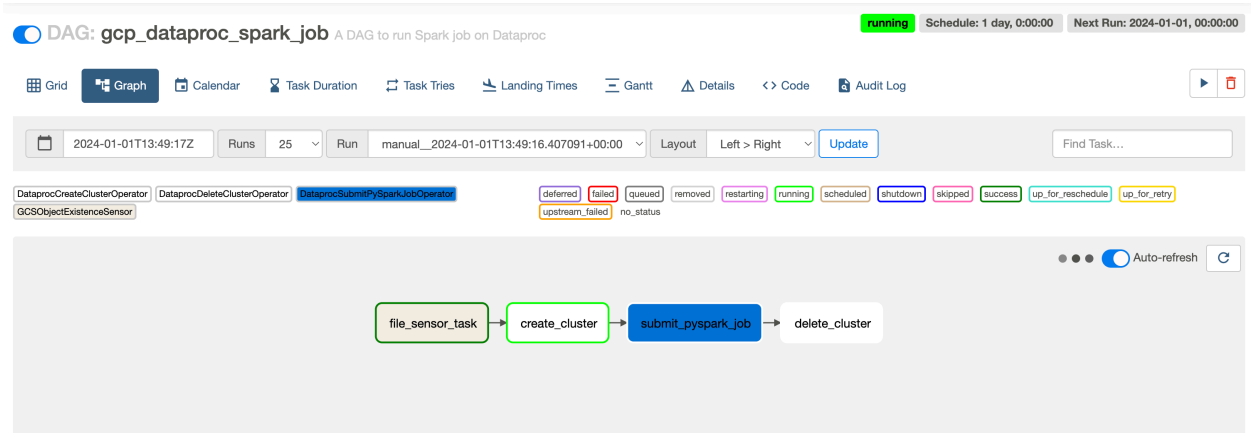


Step 3:

Now we can see the various stages in the DAG. A file sensor checks every 5 mins in the input_file location, once it detects employee.csv a new cluster is created followed by which the spark job is launched which then filters employees with salary >=60,000 and then places the output in another GCS location in the bucket airflow_ass1 under the hive_data folder.

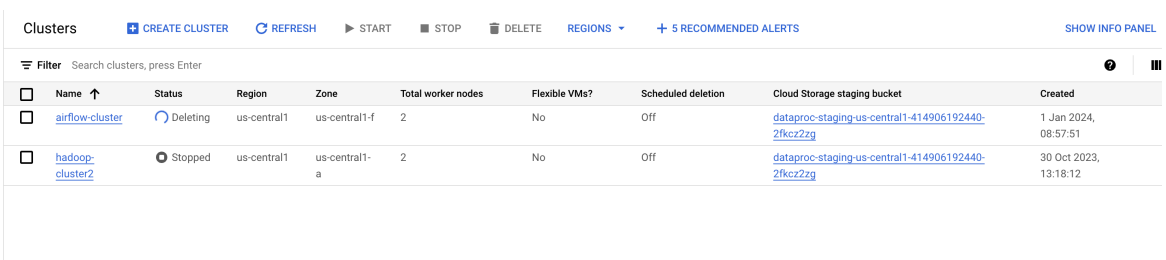


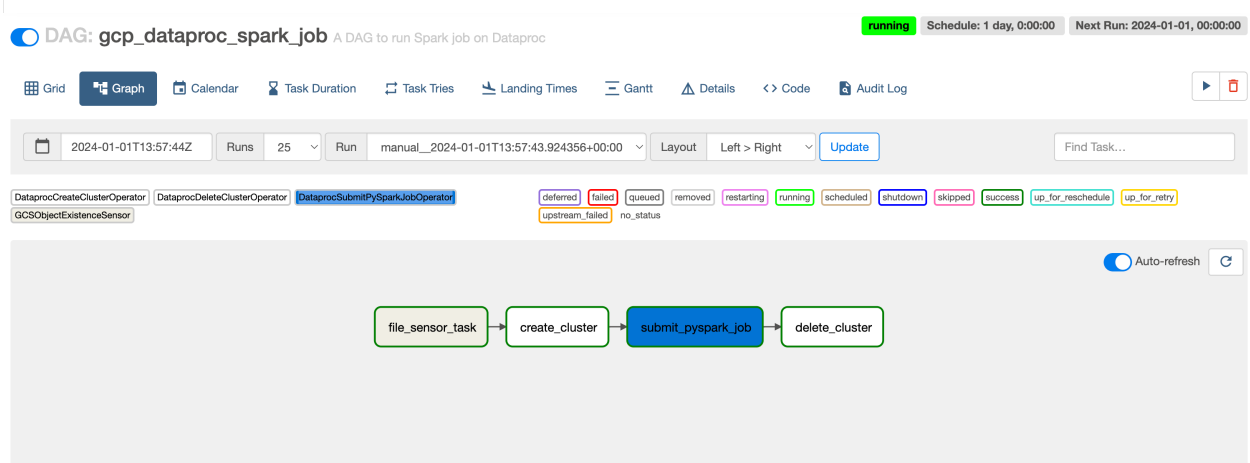




Step 4:

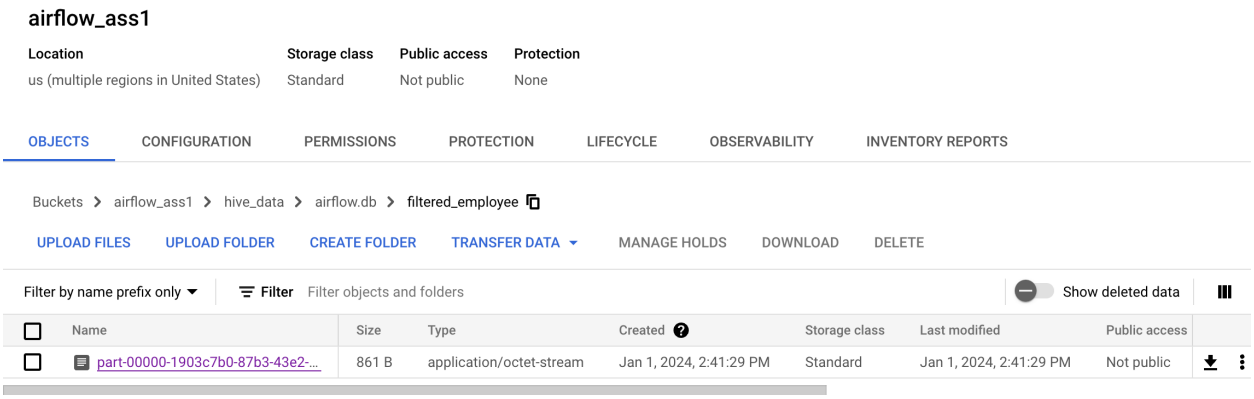
We can see that all stages have run successfully followed by deletion of the cluster as well





Step 5:

We can see the resultant data saved in a parquet file format saved in hive_data below (over which we can build external hive tables to query the data)



Challenges

1. Issues with the spark job being picked up since we had to define a location to save the output.
2. We can't put the output on the local of the cluster since with the deletion phase of the cluster this data would disappear as well. Hence it makes sense to place the output in a GCS bucket
3. Make sure to specify the format like "hive"/"parquet" while saving the output.