

Name: Abhishek Obla Hema  
Email: [oh\\_abhishek@yahoo.com](mailto: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.

**airflow\_ass2**

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

**OBJECTS**   CONFIGURATION   PERMISSIONS   PROTECTION   LIFECYCLE   OBSERVABILITY   INVENTORY REPORTS

Buckets > airflow\_ass2 > input\_files

UPLOAD FILES   UPLOAD FOLDER   CREATE FOLDER   TRANSFER DATA   MANAGE HOLDS   DOWNLOAD   DELETE

Filter by name prefix only   Filter   Filter objects and folders   Show deleted data

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	
<input type="checkbox"/>	Employee.json	386 B	application/json	Jan 1, 2024, 3:16:46 PM	Standard	Jan 1, 2024, 3:16:46 PM	Not public	

#### 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

Google Cloud

test-project

cloud composer

Search

35

Composer

Environments

CREATE

REFRESH

DELETE

Filter

Filter environments

State	Name	Location	Composer version	Airflow version	Creation time	Update time	Airflow webserver	DAG list	Logs	DAGs folder	Labels
<input checked="" type="checkbox"/>	<a href="#">airflow-cluster</a>	us-central1	1.20.12	2.4.3	31/12/2023, 21:11	31/12/2023, 21:27	<a href="#">Airflow</a>	<a href="#">DAGs</a>	<a href="#">Logs</a>	<a href="#">DAGs</a>	None

us-central1-airflow-cluster-e06b719c-bucket

Location

Storage class

Public access

Protection

us-central1 (Iowa)

Standard

Subject to object ACLs

None

OBJECTS

CONFIGURATION

PERMISSION

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

Buckets

us-central1-airflow-cluster-e06b719c-bucket

days

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

DOWNLOAD

DELETE

Filter by name prefix only

Filter

Filter objects and folders

Show deleted data

Name	Size	Type	Created	Storage class	Last modified	Public access
<a href="#">airflow_ass1_job.py</a>	2.7 KB	text/x-python-script	31 Dec 2023, 21:58:49	Standard	31 Dec 2023, 21:58:49	Not public
<a href="#">airflow_ass2_job.py</a>	2.2 KB	text/x-python-script	1 Jan 2024, 06:31:12	Standard	1 Jan 2024, 06:31:12	Not public
<a href="#">airflow_monitoring.py</a>	809 B	text/x-python	31 Dec 2023, 21:24:55	Standard	31 Dec 2023, 21:24:55	Not public

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.

DAG: fetch\_json\_and\_load\_to\_hive

DAG to fetch a daily JSON file from GCP bucket and load into Hive table on Dataproc

Schedule: 1 day, 0:00:00

Next Run: 2023-01-15, 00:00:00

Grid

Graph

Calendar

Task Duration

Task Tries

Landing Times

Gantt

Details

Code

Audit Log

01/01/2024, 02:26:31 PM

5

All Run Types

All Run States

Clear Filters

deferred

failed

queued

removed

restarting

running

scheduled

shutdown

skipped

success

up\_for\_reschedule

up\_for\_retry

upstream\_failed

no\_status

Auto-refresh

fetch\_json\_and\_load\_to\_hive

Duration

00:01:03

00:00:31

00:00:00

Jan 12, 00:00

download\_from\_gcs

submit\_hive\_job

DAG Details

DAG Runs Summary

Total Runs Displayed	5
Total running	5
First Run Start	2024-01-01, 14:25:50 UTC
Last Run Start	2024-01-01, 14:25:53 UTC
Max Run Duration	00:01:03
Mean Run Duration	00:01:01
Min Run Duration	00:00:59

## **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