

# PythonOperator Demonstration

This document will guide you through the demonstration of the PythonOperator in Session 2 of the Airflow module.

## Prerequisites:

transactions.csv (our data)

sample\_python.py(the code explained in the video)

## What are we doing?

In this demonstration, we need to create a DAG that consists of two tasks.

Bash task - checks whether a file exists or not

Python task - Once the bash task succeeds, the python task will read the file and analyze the data

## Please follow the instructions below:

1. Login to your EMR instance.
2. Activate the Python virtual environment using the following command:

**source /home/hadoop/airflow/bin/activate**

```
[hadoop@ip-172-31-33-82 ~]$ source /home/hadoop/airflow/bin/activate  
(airflow) [hadoop@ip-172-31-33-82 ~]$ |
```

3. Since the location the bash task is checking is /home/hadoop/transactions.csv, you need transactions.csv to be present in the /home/hadoop/ directory. If the file is not present in the correct location, the bash task will fail and the python task will never execute. Thus place the transactions.csv file in the /home/hadoop/ directory using WinSCP or create a new file and paste the contents in that file as shown in the video.
4. Run the following command to make sure that the file is in the correct directory.

**ls -l /home/hadoop/transactions.csv**

```
python3.7.4rc1 /home/hadoop/transactions.csv  
(airflow) [hadoop@ip-172-31-33-131 ~]$ ls -l /home/hadoop/transactions.csv  
-rwxrwxr-x 1 hadoop hadoop 349 Feb 18 22:57 /home/hadoop/transactions.csv  
(airflow) [hadoop@ip-172-31-33-131 ~]$ |
```

- Now you need to place the **sample\_python.py** file in the **/home/hadoop/airflow/dags/** directory. (You can use WinSCP or create a new file and paste the code in that file)
- To ensure that the file there are no issues/errors with the file is it considered good practice to compile the program using the following command:

**python sample\_python.py**

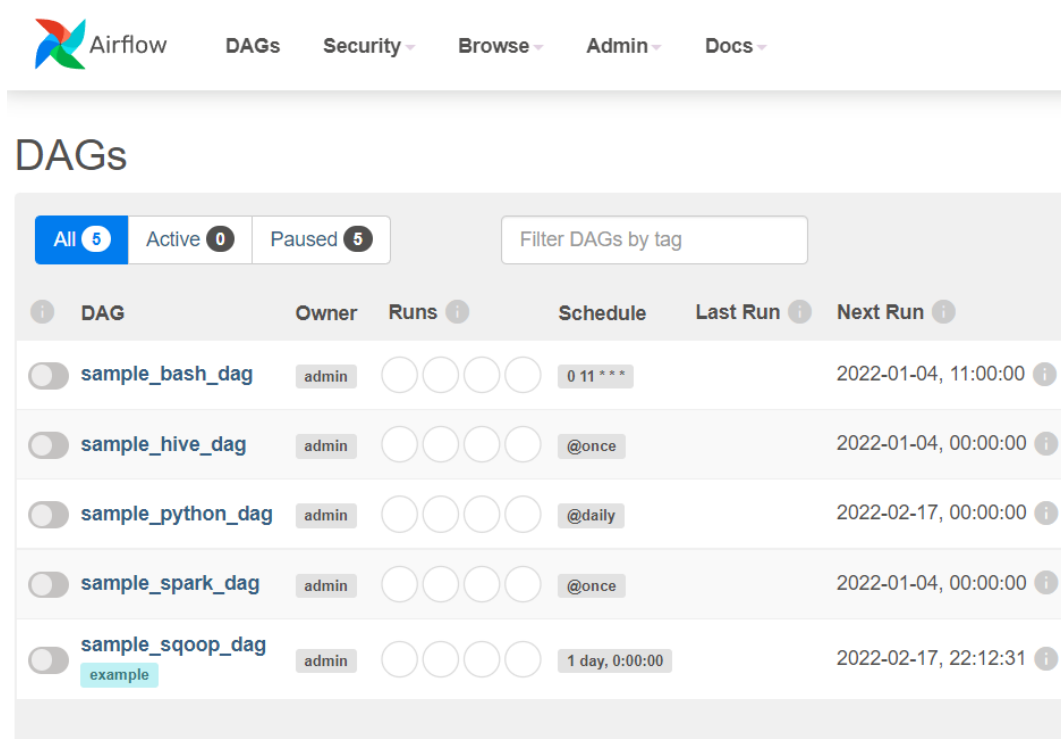
- You can also use the following command to list the dags in your instance:

**airflow dags list**

- Once you have made sure that your dag file has no issues you can go to the Airflow UI which is hosted in the URL : **your\_public\_dns:8082**

**Note:** You can find you\_publin\_ip in your AWS EMR dashboard (IPv4 Public DNS)

- Switch ON the DAG(sample\_python\_dag)



The screenshot shows the Airflow web interface. At the top, there's a navigation bar with links: Airflow, DAGs, Security, Browse, Admin, and Docs. Below this, the 'DAGs' section is active. It features a filter bar with 'All 5', 'Active 0', and 'Paused 5' buttons, along with a 'Filter DAGs by tag' input field. The main content is a table listing DAGs:

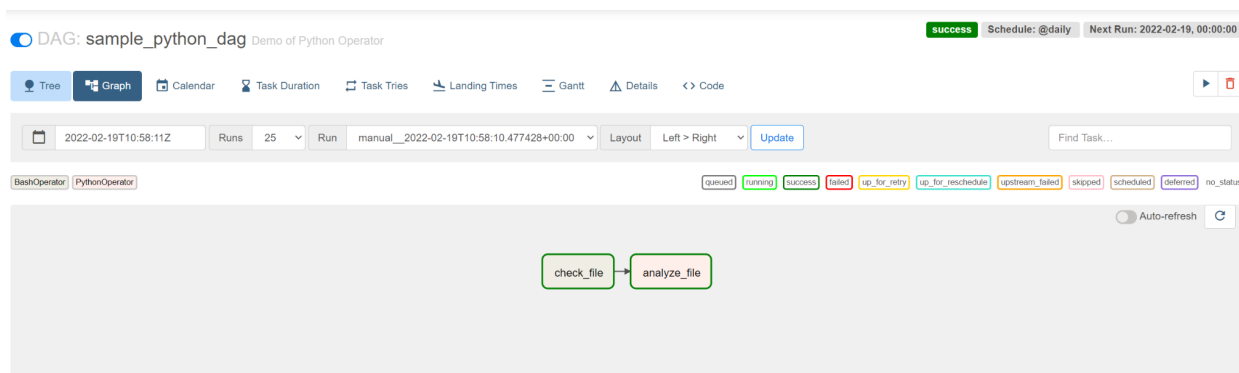
DAG	Owner	Runs	Schedule	Last Run	Next Run
<input type="checkbox"/> sample_bash_dag	admin	<div><div></div><div></div><div></div><div></div><div></div></div>	0 11 * * *		2022-01-04, 11:00:00
<input type="checkbox"/> sample_hive_dag	admin	<div><div></div><div></div><div></div><div></div><div></div></div>	@once		2022-01-04, 00:00:00
<input type="checkbox"/> sample_python_dag	admin	<div><div></div><div></div><div></div><div></div><div></div></div>	@daily		2022-02-17, 00:00:00
<input type="checkbox"/> sample_spark_dag	admin	<div><div></div><div></div><div></div><div></div><div></div></div>	@once		2022-01-04, 00:00:00
<input type="checkbox"/> sample_sqoop_dag example	admin	<div><div></div><div></div><div></div><div></div><div></div></div>	1 day, 0:00:00		2022-02-17, 22:12:31

10. After switching on the DAG, you need to start the DAG by clicking on the play button



11. Click on the sample\_python\_dag and go to the graph view

You will see the check\_file(Bash) task is running/completed



If the DAG is still running then click on the refresh button to check if the DAG run is finished.

12. Once the DAG has completed execution, the output will be generated in the /home/hadoop/output.json location

You can use the **vi /home/hadoop/output.json** command to see the same

```
"I301": 1,
"I302": 4,
"I305": 2,
"I307": 1,
"I303": 1,
"I304": 1
```

13. At any point, if a task fails you can click on it and click on the clear button to restart the task

Task Instance: check\_file ×  
at: 2022-02-19, 10:58:10 UTC

[Instance Details](#)
[Rendered](#)
[Log](#)
[All Instances](#)
[Filter Upstream](#)

Download Log (by attempts):  
1

Task Actions

[Ignore All Deps](#)
[Ignore Task State](#)
[Ignore Task Deps](#)
[Run](#)

[Past](#)
[Future](#)
[Upstream](#)
[Downstream](#)
[Recursive](#)
[Failed](#)
[Clear](#)

[Past](#)
[Future](#)
[Upstream](#)
[Downstream](#)
[Mark Failed](#)

[Past](#)
[Future](#)
[Upstream](#)
[Downstream](#)
[Mark Success](#)

[Close](#)

14. You can switch off your DAG if you don't want it to run anymore.