AIrflow

Orchestration Service

JERIN SAM

Table of Contents

[Airflow 3](#_Toc181730863)

[Core Components 3](#_Toc181730864)

[Core Components 4](#_Toc181730865)

[Types of Airflow Architecture: 5](#_Toc181730866)

[How does Airflow work 7](#_Toc181730867)

[Execution Steps 7](#_Toc181730868)

[Important Points to Remember 10](#_Toc181730869)

[Different Views to Monitor DAG 11](#_Toc181730870)

[Code Data Pipeline 14](#_Toc181730871)

[Setup Environment to write Python code 15](#_Toc181730872)

[Define a DAG 15](#_Toc181730873)

[DAG 15](#_Toc181730874)

[DAG Python Script Details 17](#_Toc181730875)

[Define Operator 17](#_Toc181730876)

[Operator 17](#_Toc181730877)

[Providers 17](#_Toc181730878)

[Operator Python Script Details 18](#_Toc181730879)

[Test the Task 20](#_Toc181730880)

[Define Sensors 21](#_Toc181730881)

[Sensors 21](#_Toc181730882)

[Sensors Python Script Details 21](#_Toc181730883)

[Additional Operators 22](#_Toc181730884)

[Define Hook 23](#_Toc181730885)

[Hook 23](#_Toc181730886)

[Hook Python Script Details 24](#_Toc181730887)

[Define Dependency 25](#_Toc181730888)

[Dependency 25](#_Toc181730889)

[Dependency Python Script Details 25](#_Toc181730890)

[DAG and Tasks - Python Solution 26](#_Toc181730891)

[Test DAG and Its Tasks 26](#_Toc181730892)

[Scheduling DAGs using Datasets 27](#_Toc181730893)

[Dataset Schedule/ Trigger 27](#_Toc181730894)

[Important concepts while implementing Dataset Triggers 28](#_Toc181730895)

[Python Script Snippet 28](#_Toc181730896)

[Dataset - Python Solution 29](#_Toc181730897)

[Executor 30](#_Toc181730898)

[Appendix 30](#_Toc181730899)

[Important Links 30](#_Toc181730900)

# Airflow

Airflow is an open-source platform to programmatically author, schedule and monitor workflows.

***Benefits:***

* It is a python-based framework, which can help to create dynamic tasks, workflows and branching (execute tasks based on certain conditions).
* Airflow is highly scalable.
* Its Web UI provides good visibility on the workflows.
* Extensibility: Customize Web UI, create and add new connectors etc.

***What Airflow isn’t:***

Airflow is an orchestrator and not any processing framework or Realtime streaming solution or any storage solution. Process your gigabytes of data outside of Airflow (i.e. You have a Spark cluster, you use an airflow operator to execute a Spark job, and the data is processed in Spark).

***Use case where Airflow is not the Best:***

* High Frequency, sub-minute scheduling.
* Directly processing large datasets.
* Real-time data streaming.
* Simple workflow with less dependency – CRON jobs can be used instead of Airflow

## Core Components

1. ***Web Server***: Helps to view, manage and monitor the workflows using a Web UI.
2. ***Scheduler***: Schedules tasks; It ensures tasks run at the right time and in correct order.
3. ***Meta Database***: Stores the information of the workflows, users, tasks and status of the tasks. Airflow uses Meta DB to remember which tasks have been executed, when it was executed and its status.
4. ***Trigger***: Trigger is responsible for managing deferrable tasks – tasks that wait for external events.
5. ***Executor:*** It determines how the tasks will be executed, deciding whether to run in parallel or in sequence and on which system.
6. ***Queue:*** It’s a list of tasks waiting to be executed.
7. ***Worker***: It is a process which is executing the task, it does the work defined in the tasks.

## Core Components

1. ***DAG (Directed Acyclic Graph)***: Collection of all the tasks that need to be executed, it also defines organizing these tasks in a way that reflect the dependency.

**i.e.** It helps to define the entire structure of the workflow, defining which tasks (T1, T2, T3 and T4) need to be executed before others, as shown in below snippet.

A screenshot of a computer

Description automatically generated

It supports only non-cyclic graph, won’t support cyclic nature of the graph. As can be seen in the above snippet, task T1 depends on T4 and T4 depends on T1 causing cyclic graph, which is not supported by Airflow.

Since DAG is a graph, therefore ***Task*** is also called as ***Node*** and ***Dependency*** is also called as ***Edge***.

A diagram of a diagram

Description automatically generated

1. ***Operator***: Operators allow to break down the workflows in discrete, manageable piece of work (Taks). It defines a single, idempotent, task in the DAG

Follow below link to get the list of Providers which is used to create Operators:

1. [**https://registry.astronomer.io/providers**](https://registry.astronomer.io/providers)

***\*\* idempotent*** means for a give input, output will remain same irrespective of the number of executions of the tasks.

***\*\* Providers:*** Providers can be considered as source connectors which is used to configure operators.

1. ***Task/ Task Instance***: Task is a specific instance of an Operator. When an Operator is assigned to a DAG, it becomes a Task.

Tasks are the actual unit of work that gets executed when DAG runs.

1. ***Sensors*** are specialized tasks designed to wait for a particular condition to be met before allowing the DAG to proceed.

***Sensor "sense" external events or states***, such as checking if a file exists in a specific location, confirming that data is available, or verifying that a process in an external system is complete.

Sensors are useful when your workflow depends on resources or events outside of Airflow's immediate control.

1. ***Workflow:*** A Workflow is the entire process defined by DAG, including all tasks and its dependencies, as can be seen in below snippet.

A screenshot of a computer

Description automatically generated

## Types of Airflow Architecture:

1. ***Single Node Architecture:***

In Single Node Architecture, all components of Airflow run on a single node.

As can be seen below snippet of Single Node Architecture, all components are running on a single machine.

It is used during development or running small scale workflows.

A diagram of a computer

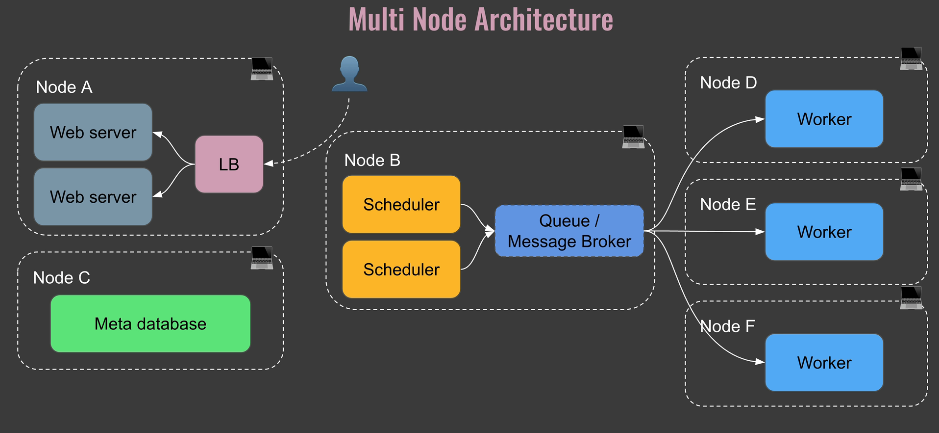
Description automatically generated

***Scheduler and Queue is generally part of Executor.***

1. ***Multi Node Architecture:***

In Multi Node Architecture, components of Airflow run on different nodes.

***Details of multi-node architecture:***



One node will have one or multiple Web Servers with Load Balancer.

* Dedicated node for Scheduler and Queue, Executor is generally part of Scheduler and for Queue Message Brokers are used like RabbitMQ.
* Each worker will have its own node.
* Meta Database will be hosted on separate node.

***Pros of using multi node architecture:***

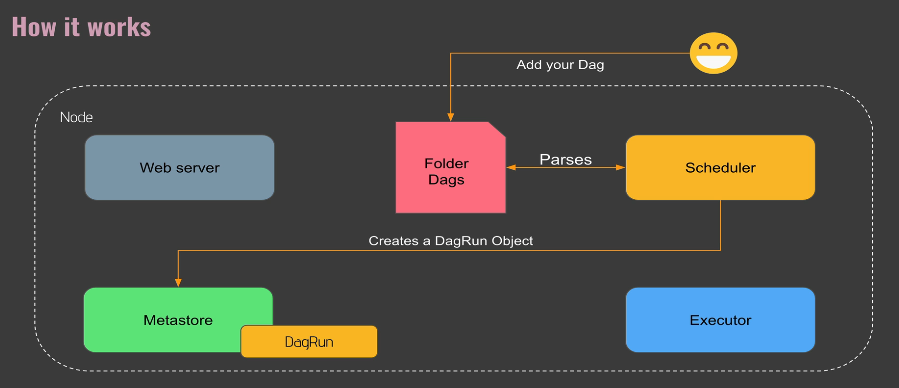
* It is easy to scale.
* Provides fault tolerance (Reliability) i.e. if one of the worker nodes goes down then its tasks will be handled by other worker nodes.
* This architecture also improves the performance by distributing the tasks to multiple workers.

## How does Airflow work

### Execution Steps

**Step 1:**

User submits the python DAG file in the DAG Folder, which will be further parsed by the scheduler and creates a DAGrun object which will be stored in the Metastore and mark it as scheduled.



**Step 2:**

Once DAGrun is scheduled, now scheduler will create TaskInstance object and store it in Metastore and mark it as scheduled .

A diagram of a diagram

Description automatically generated

**Step 3:**

Once the task instance is scheduled now Scheduler will submit the TaskInstance Object to Executor, which will further decide how to execute the Tasks and it will assign a worker (not shown in the instance) to execute the same.

A diagram of a computer

Description automatically generated

**Step 4:**

Once the task is successfully executed the Worker, Executor will update the status of the tasks on Metastore.

A diagram of a computer

Description automatically generated

**Step 4:**

Once the all the tasks associated to a DAG are successfully executed then DAGRun object will be marked as Executed.

A diagram of a workflow

Description automatically generated

**Step 4:**

Web Server will continuously take to Metastore and get the latest update of the DAGs i.e. Scheduled DAGs, Executing DAGs/ Tasks, Failed Tasks etc..

A diagram of a dog

Description automatically generated

### Important Points to Remember

Whenever a DAG is executed, a DAGRun object is created, which has timestamp information ***data\_interval\_start*** and ***data\_interval\_end.***

***data\_interval\_start:***

* It marks the beginning of the data interval, or the range of data that a task is responsible for processing.
* It essentially marking the start of the ***logical interval*** *also known as* ***logical data***.
* ***execution date*** was the older term for ***logical date*** - meaning they all represent the start of the data interval, not the actual time the DAG is executed.

***data\_interval\_end:***

* It marks the end of the data interval, or the end of the data collection period for a DAG run.

A diagram of a diagram

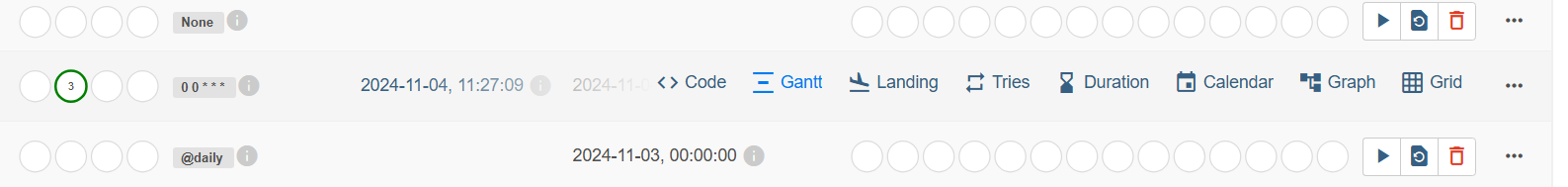
Description automatically generated

In the above example shown in the snippet, if the DAG is deployed on 2022-01-01 and scheduled to execute at 2022-01-02 midnight then –

1. During 1st scheduled execution - ***data\_interval\_start*** will be ***2022-01-01*** and ***data\_interval\_end*** will be ***2022-01-02.***
2. During 2nd scheduled execution - ***data\_interval\_start*** will be ***2022-01-02*** and ***data\_interval\_end*** will be ***2022-01-03.***

# Different Views to Monitor DAG

There are multiple views provided by Airflow Web UI to monitor DAGs.



***Below is the list of views:***

1. Grid View

It provides the details of the DAG and its tasks like Operator used in the tasks, execution time etc.

A screenshot of a computer

Description automatically generated

1. Graph View

Graph view will help to understand the dependencies among tasks.

A screenshot of a computer

Description automatically generated

1. Gantt View

Gantt View will help to understand the start and end time taken by the tasks to execute; this view is helpful to identify which tasks are consuming most time for the execution.

A screenshot of a computer

Description automatically generated

1. Code View

Code view helps to view the python code used to create the DAG and the associated tasks. This is helpful to check if the changes done to the Python code is applied to the DAG.

A screenshot of a computer

Description automatically generated

1. Landing View

Landing View is used to visualize the overall time taken by the DAG and its Tasks to execute over the period of time.

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Calendar View

Calendar View helps to check the status of the DAGs.

A screenshot of a computer

Description automatically generated

# Code Data Pipeline

In this exercise, A DAG will be created along with following tasks –

1. A task to create a table in Postgres.
2. A task to check the availability of API.
3. Extract data from API.
4. Push data to Postgres Table.

Dependencies of the tasks will be defined after the creation of tasks.

## Setup Environment to write Python code

Create a python script in *./airflow-working-folder/dags* folder with name “***user-processing.py***”

Use the above python file for adding DAG and Operators logics, as this folder is mounted with Airflow Docker’s DAG folder structure. Airflow’s scheduler will pick the python file from that Docker’s DAG folder location to execute.

Details of the mount point can be found in docker-compose.yaml file.

## Define a DAG

### DAG

While defining DAG, following parameters need to be passed –

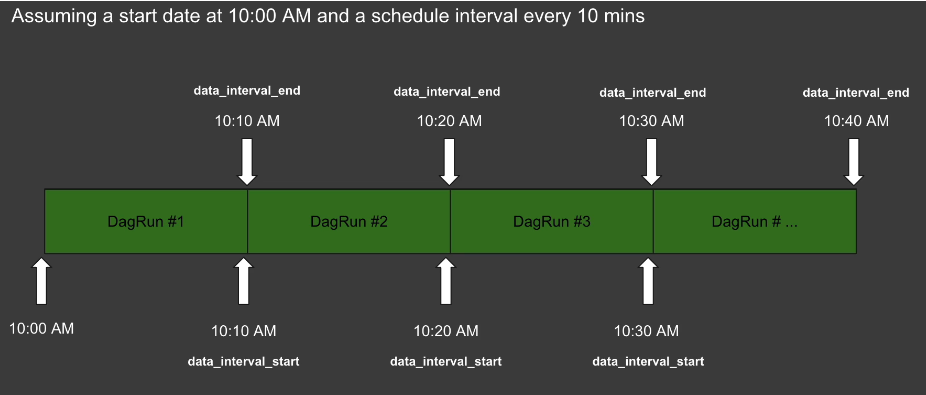
* ***dag\_id***: Provide unique name to the DAG.
* ***start\_date:*** Assign execution Start Date of the DAG.
* ***schedule\_interval:*** Assign Schedule Interval to execute the DAG at a set frequency – It accepts CRON Format.

There is another parameter called ***timetable*** which can be used to schedule the DAG using Airflow’s *timetable* *class*.

Follow below link to go through Airflow’s scheduler

<https://airflow.apache.org/docs/apache-airflow/1.10.10/scheduler.html#dag-runs>

A DAG is triggered after the ***start\_date/last\_run+ schedule\_interval***



In the above example, let’s say DAG *start\_date is 2022-01-01* and is scheduled to execute every 10 mins.

Initially Scheduler will wait for 10 mins before the actual execution so, *data\_interval\_start = 2022-01-01* 10 AM and after 10 mins i.e. at 10:10 AM DAG is executed, so *data\_interval\_end = 10:10 AM*

Now during 2nd execution schedule, *data\_interval\_start =* 10:10 AM and after 10 mins i.e. at 10:20 AM DAG is executed again, so *data\_interval\_end = 10:20 AM*

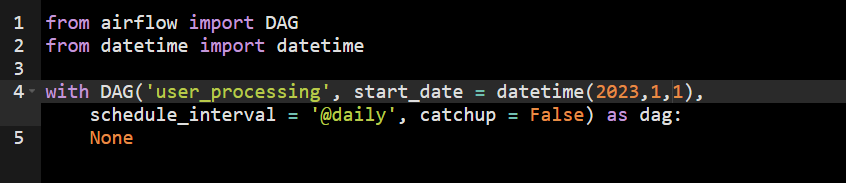
* ***catchup:*** This is used to backfill the data. For Example, If the DAG's ***start\_date*** is in the past, setting ***catchup=True*** will cause the DAG to run for each interval between the start date and today.

To avoid this backlog of runs, set ***catchup=False***. This ensures the DAG only runs from the current date forward, preventing unnecessary historical runs.

A screenshot of a video game

Description automatically generated

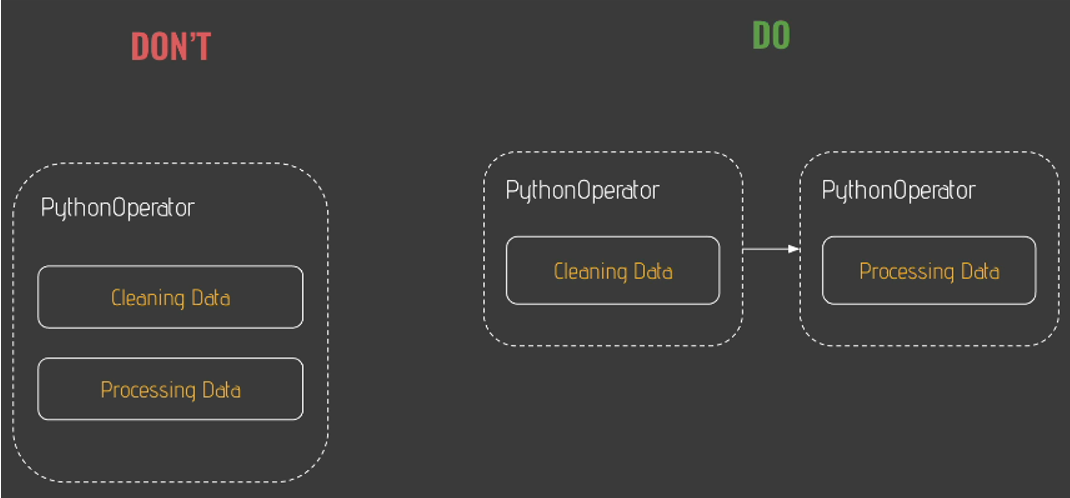
### DAG Python Script Details



## Define Operator

### Operator

While creating operator don’t combine multiple steps in 1 operator as shown in below snippet. Break multiple processing steps into multiple Operators.



***There are 3 types of Operators:***

1. ***Action Operator***: Used to execute an action like set of code.
2. ***Transfer Operator***: Used to transfer data from point A to B.
3. ***Sensor Operator***: Wait for a condition to be fulfilled.

Providers

To enable Apache Airflow to interact with various data sources, specific providers for each source needs to be installed.

***\*\**** *Providers can be considered as source connectors which is used to configure operators.*

***Install core Airflow***: pip install apache-airflow

***Install Snowflake Provider***: pip install apache-airflow-snowflake

A screenshot of a computer

Description automatically generated

*While configuring an Operator, a connection needs to be established in Airflow’s Web UI and connection\_id used in the connection will be referenced while configuring the Operator in Python Script.*

### Operator Python Script Details

In this example, Postgres Operator is used to create a table in Postgres DB.

Operator requires certain parameters, one of such parameters is connection Id, to get the connection id, set up Postgres connection in Airflow Web UI and use the connection name provided while setting up the connection in the python script.

#### Airflow Web UI Setup Postgres Connection using below details

* *Host: postgres*
* *DB: airflow*
* *Login: airflow*
* *Password: airflow*
* *Port: 5432*

A screenshot of a computer

Description automatically generated

#### Python Code Snippet

A screen shot of a computer

Description automatically generated

To understand the parameters used in the operator use the documentation associated to that operator. Use the link below to access the documentation and navigate to Python API section.

* All the providers can be found at

<https://airflow.apache.org/docs/>

* Postgres provider’s operator documentation:

<https://airflow.apache.org/docs/apache-airflow-providers-postgres/5.8.0/_api/airflow/providers/postgres/operators/postgres/index.html>

A screenshot of a computer

Description automatically generated

### Test the Task

* Go to the folder path where docker-compose.yaml file is kept and execute below script to list all active containers.

|  |
| --- |
| ***docker-compose ps*** |

* Open the Airflow docker shell from the Host Machine, using the below script.

***# code snippet****: docker exec -it <container\_name> /bin/bash*

|  |
| --- |
| ***docker exec -it 2-airflow-docker-configure-airflow-scheduler-1 /bin/bash*** |

* Test the shell by using below script, this will help to check whether airflow is accessible or not.

|  |
| --- |
| ***airflow -h*** |

* Test Airflow tasks using below script:

***# code snippet****: airflow tasks test <dag\_id> <task\_id> <past-date>*

|  |
| --- |
| ***airflow tasks test user\_processing create\_table 2022-01-01*** |

* After executing above script, if Success message comes then Tasks are successfully created.

## Define Sensors

### Sensors

***Sensors*** has 2 main parameters, to control the frequency and duration of checks for a condition -

1. ***poke\_interval*** 
   * The amount of time (in seconds) between each check for the condition.
   * If *poke\_interval=60*, the sensor will check every 60 seconds to see if the target condition has been satisfied.
   * Default value is 60 seconds.
2. ***timeout*** 
   * The maximum amount of time (in seconds) the sensor will wait for the condition to be met.
   * If *timeout=600*, the sensor will wait up to 10 minutes. If the condition is not met within that time, the sensor fails, and the task is marked as failed.
   * Default value is 7 days.

*While configuring an Operator, a connection needs to be established in Airflow’s Web UI and connection\_id used in the connection will be referenced while configuring the Operator in Python Script.*

### Sensors Python Script Details

In this exercise, a sensor task is created to check whether API is available or not.

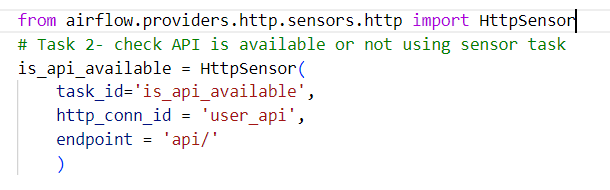
#### Airflow Web UI Setup HTTP Connection using below details:

* *Connection\_id: user\_api*
* *Connection Type: HTTP*
* *Host:* [*https://randomuser.me/*](https://randomuser.me/)

A screenshot of a computer

Description automatically generated

#### Python Code Snippet



## Additional Operators

* Create an Operator to extract users from the API

A screen shot of a computer code

Description automatically generated

* Process data, which is extracted from API, PythonOperator is used to manage it.

A screenshot of a computer code

Description automatically generated

## Define Hook

### Hook

***Hook*** is an abstract layer for interacting with external systems and services.

Hooks provide a way for Airflow tasks to connect to, authenticate, and interact with different types of data sources, such as databases, cloud storage, APIs, and message brokers.

Behind the scenes Operators talk to Data Source using Hooks, as shown in below snippet.

A screenshot of a diagram

Description automatically generated

*Hook provides access to certain functions, which is not accessible via Operators. In the below snippet copy\_expert function in PostgresHook is an example of such function which is not accessible via PostgresOperator*

### Hook Python Script Details

A screenshot of a computer

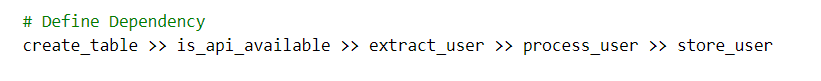
Description automatically generated

## Define Dependency

### Dependency

Dependency is defined using ***>>***, **<<** operator and ***set\_downstream*** and ***set\_upstream*** methods can also be used. This helps to define which task is dependent upon which another task.

### Dependency Python Script Details

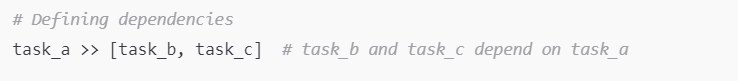


*Web UI graph view snippet to view dependency between tasks*

A screen shot of a computer

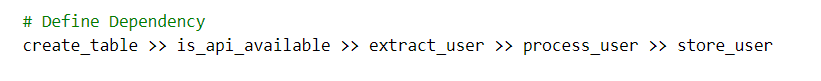
Description automatically generated

***Example 1:***



The above dependency denotes, ***task\_b and task\_c depend on task\_a***

***Example 2:***



*The above dependency can also be written as follows:*

A screenshot of a computer program

Description automatically generated

## DAG and Tasks - Python Solution

Follow below path for the python script implemented-

* ***./main/airflow-working-folder/dags/user-processing.py***

## Test DAG and Its Tasks

* Go to the folder path where docker-compose.yaml file is kept and execute below script to list all active containers.

|  |
| --- |
| ***docker-compose ps*** |

* Open the Postgres docker shell from the Host Machine, using the below script.

***# code snippet****: docker exec -it <container\_name> /bin/bash*

|  |
| --- |
| ***docker exec -it 2-airflow-docker-configure-postgres-1/bin/bash*** |

* Execute below script to open Postgres Server with User Airflow .

|  |
| --- |
| ***psql -UAirflow*** |

* Execute select statement to check data in the Table

|  |
| --- |
| ***select \* from users;*** |

* If data is populated after executing select statement, then DAG has successfully executed.

# Scheduling DAGs using Datasets

## Dataset Schedule/ Trigger

***Note: schedule\_interval*** *and* ***timetable parameters*** *will be* ***deprecated*** *and is* ***replaced*** *with parameter* ***“schedule”***

Apart from using CRON method to schedule DAGS, there’s another method called ***Dataset*** which helps to trigger a DAG after the successful execution of dependent DAG.

***Airflow dataset is a logical grouping of data***. **Upstream producer tasks can update datasets, and dataset updates contribute to scheduling downstream consumer DAGs**.

There are 2 important properties to a Dataset:

* ***URI –*** 
  + Unique Identifier of the dataset
  + Path to the dataset
  + Supports only ASCII Characters
  + Case Sensitive
  + URI schema cannot be airflow



* ***extra –*** 
  + To provide more details to the Dataset Extra is used.

A computer screen with text

Description automatically generated

## Important concepts while implementing Dataset Triggers

* ***@task*** decorator in Airflow is specifically designed for creating tasks with the ***PythonOperator*** or similar operators that use Python functions directly. This decorator is part of the **Task Flow API.**
* ***Inlets and Outlets in Apache Airflow Tasks***
  + ***Outlets*** –

A task parameter that contains the list of ***datasets*** a specific task produces updates to, as soon as it completes successfully.

All outlets of a task are shown in the DAG graph in the Airflow UI, as well as reflected in the dependency graph of the Datasets tab as soon as the DAG code is parsed, i.e. independently of whether or not any dataset events have occurred.

Note that Airflow is not yet aware of the underlying data. It is up to you to determine which tasks should be considered producer tasks for a dataset. As long as a task has an outlet dataset, Airflow considers it a producer task even if that task doesn't operate on the referenced dataset.

* + ***Inlets:***

A task parameter that contains the list of datasets a specific task has access to, typically to access ***extra*** information from related dataset events. Defining inlets for a task does not affect the schedule of the DAG containing the task and the relationship is not reflected in the Airflow UI.

## Python Script Snippet

***Producer.py***

A screenshot of a computer program

Description automatically generated

***consumer.py***

A screen shot of a computer code

Description automatically generated

## Dataset - Python Solution

Follow below path for the python script implemented-

* ***/main/airflow-working-folder/dags/producer.py***
* ***/main/airflow-working-folder/dags/consumer.py***

In this implementation, consumer DAG is dependent on producer DAG, and consumer DAG will be triggered when producer DAG writes in the dataset file.

# Executor

# Appendix

## Important Links

* Astronomer Providers

<https://registry.astronomer.io/providers>

* Airflow Providers

<https://airflow.apache.org/docs/>

* Airflow Operators

<https://airflow.apache.org/docs/apache-airflow/stable/_api/airflow/operators/>

* Airflow Schedules and Triggers

<https://airflow.apache.org/docs/apache-airflow/1.10.10/scheduler.html#dag-runs>