Time to Code Your First DAG

A real ETL step by step

Your First Project

- The DAG:
  - The purpose of this DAG is to show you a typical ETL process which can be automated using Apache Airflow and Twitter. For simplicity, I'm not going to fetch data from the twitter API but if you want it, let me know.
  - It is composed of 4 tasks:
    - Fetching tweets
    - Cleaning tweets
    - Uploading tweets into HDFS
    - Loading data into HIVE
- For those who have never coded before, no worry, I will explain everything and the files which are not directly related to Apache Airflow will be provided.

The Project

- cd ~
- (from section 2, you should have the airflow_files directory)
- source .sandbox/bin/activate
  - If you don’t see the prompt changing with (.sandbox), type:
    - rm -rf .sandbox
    - python3.6 -m venv .sandbox
    - source .sandbox/bin/activate

The Project

- After having untar the archive you should have the following arborescence:

Don’t worry if you don’t have the .pyc files

```
(.sandbox) [airflow@localhost first_dag]$ tree
.
├── cleaning_tweet.py
├── data.csv
├── fetching_tweet.py
├── __pycache__
│   ├── cleaning_tweet.cpython-36.pyc
│   └── fetching_tweet.cpython-36.pyc
└── twitter.py

1 directory, 6 files
```

Let’s Start Coding Our DAG

- cd ~
- export PYTHONPATH=/home/airflow/airflow_files
  - Don’t forget to execute this command in each session (terminal) you use (where you run the web server, the scheduler etc).
- vim airflow/dags/twitter.py

The Code

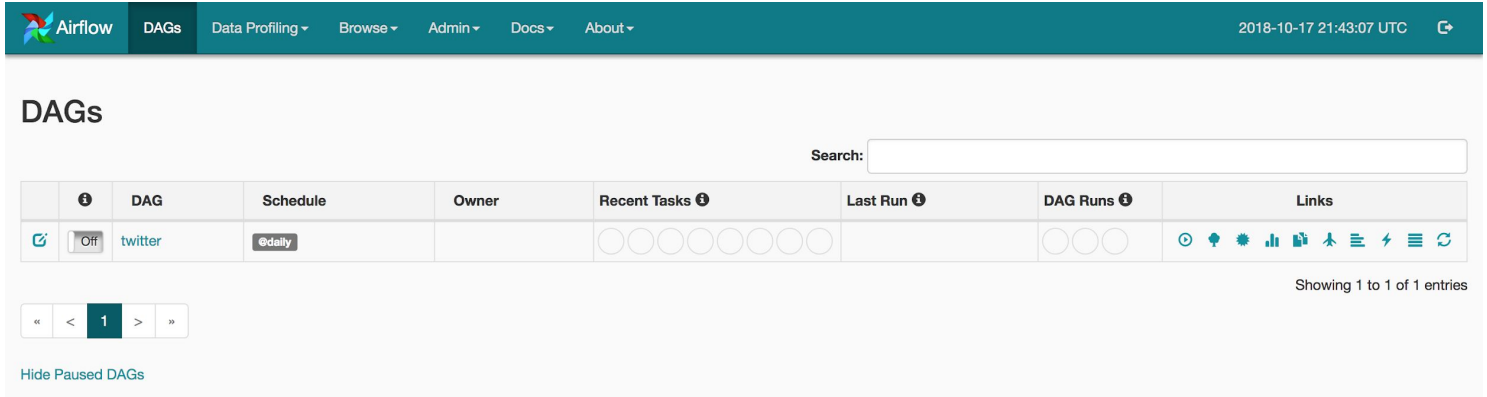
You should take look at the code of twitter.py contained into the first_dag folder.

Running Airflow

- Once you DAG is created and placed into ~/airflow/dags don't forget to run the web server as well as the scheduler with the following commands:
- `In a new terminal: airflow webserver`
- `In a new terminal: airflow scheduler`

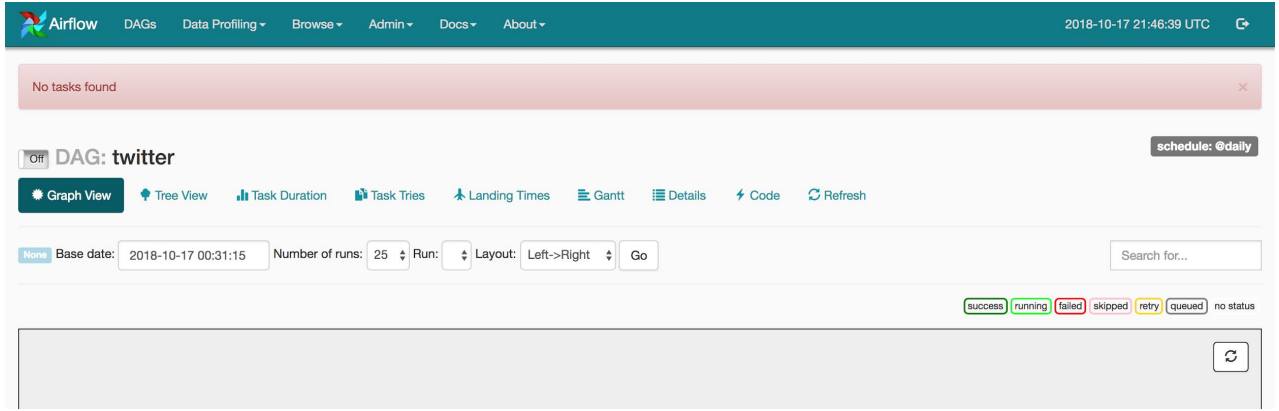
Airflow UI

- Once your Airflow UI is running you should see your dag as follow:



Airflow UI

- If you click on your dag and go to “Graph View” you should see this:



Final Note

- We basically have initialised our first DAG which is about getting tweets, clearing them and loading them into our HDFS.
- But, we still need to create our Tasks to inform Apache Airflow of how it should run our scripts and load data into HDFS.
- Let’s do this in the next lesson...