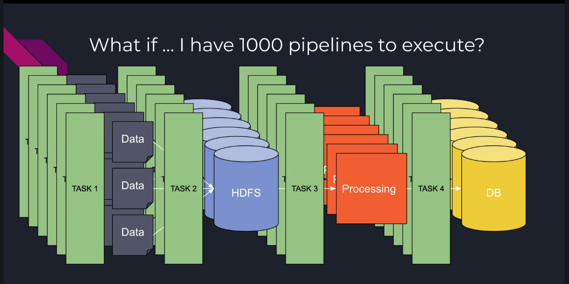
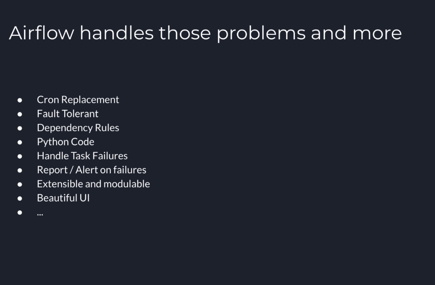
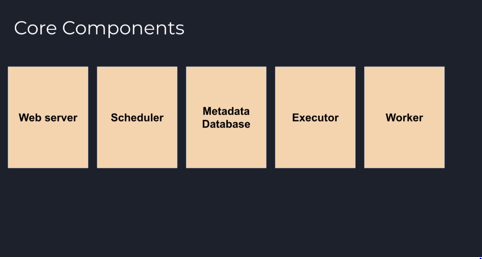
Apache Airflow:

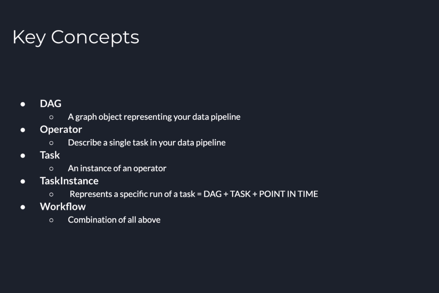
Airflow is used to process 1000’s of data pipelines in HDFS system using Apache Spark.

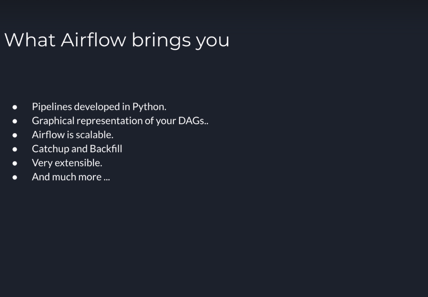
Apache Airflow is a way to programaticaly author,schedule and monitor data pipelines.



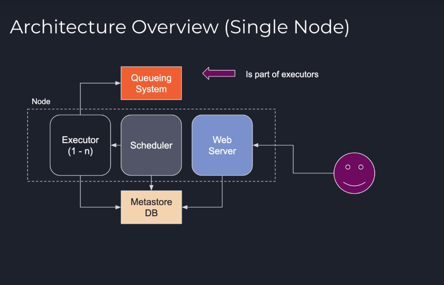




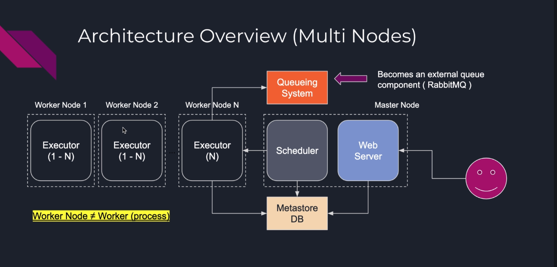




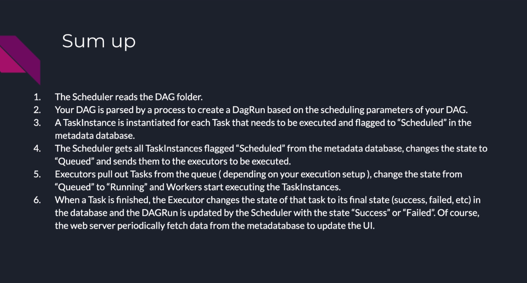
**Airflow Architecture in Singleton mode:**



Cluster or Multi-Node mode:



Airflow component workflow:



How to install airflow:

pip install “apache-airflow[celery,crypto,postgres,rabbitmq,redis]==1.10.6”

pip install –upgrade Flask

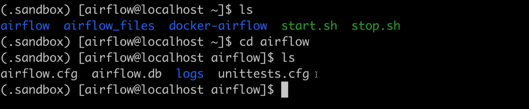
airflow initdb

Go to aiflow folder to see all config files

To edit airflow config db , after you edit airflow.cfg file you need to run

Airflow initdb to instantiate the changes

airflow initdb



Create dags folder in order to put in all your dags

Activate sandbox:

Source .sandbox/bin/activate

>airflow scheduler

>airflow webserver

Both will instantiate and run the scheduler and webserver

Go to : localhost:8080 to see the airflow dashbaord

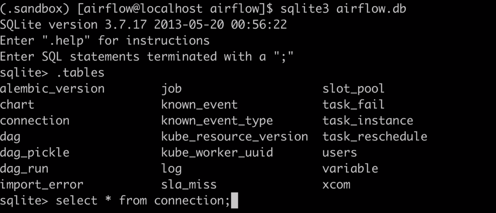
Now we can query the airflowdb by going into airflow cli( command line interface)

You can use the sqlite database to query the various tables inside the db

$ sqlite3 airflow.db

Sqlite> .tables

Sqlite> select \* from connection;



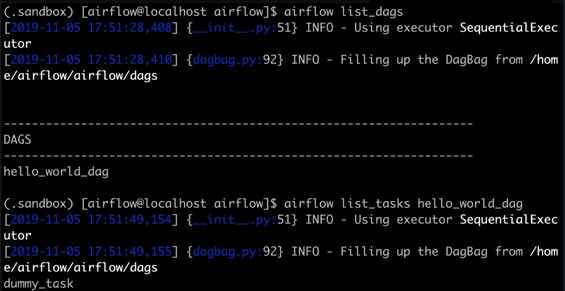
>airflow list\_dags 🡪 This will list all the dags

>airflow list\_tasks hello\_world\_dag 🡪 This will list all the details about a dag

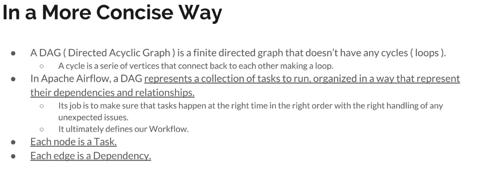
>airflow list\_tasks hello\_world\_dag –tree 🡪 This will list the order of execution of a dag

> airflow

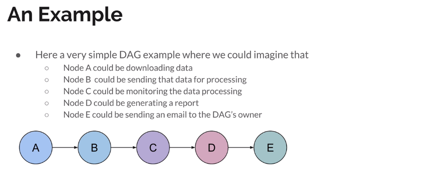
> airflow –h ( this will list all the help commands)

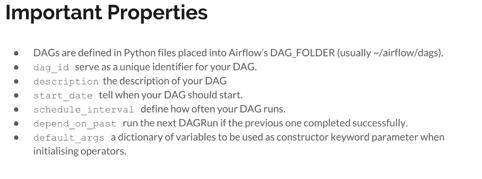


What is a DAG exactly?



Main parameters of a dag.cfg file.

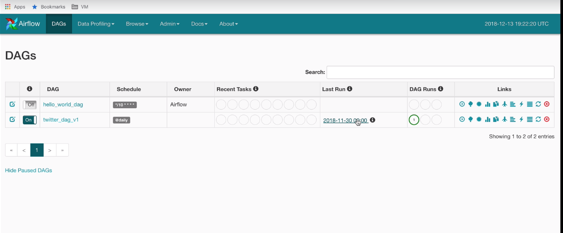




How does a DAG look like?

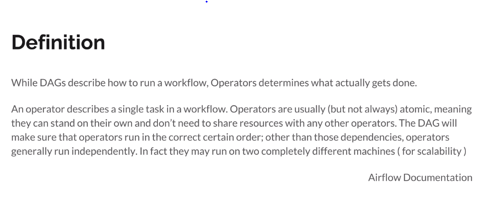


DAG view using Cloud webserver:

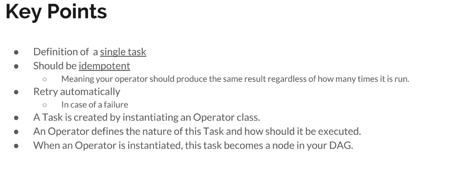


**What is an Operator:**

Operator is the actual job like mapping in Informatica. While DAG is like a scheduled workflow in Informatica.



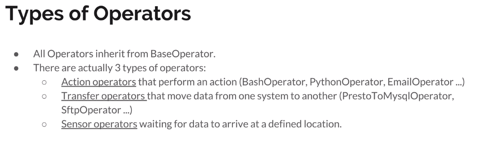
**Definition of Task**

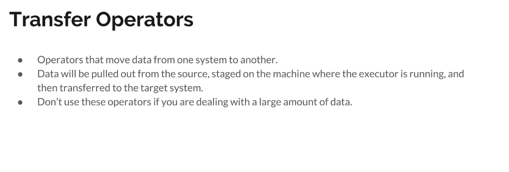


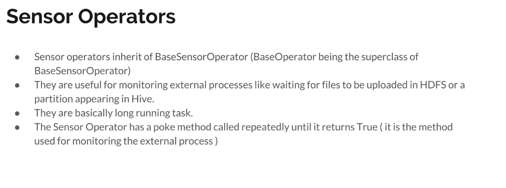
Types of Operators:

There are 4 types of operator and we can define our own operators as well:



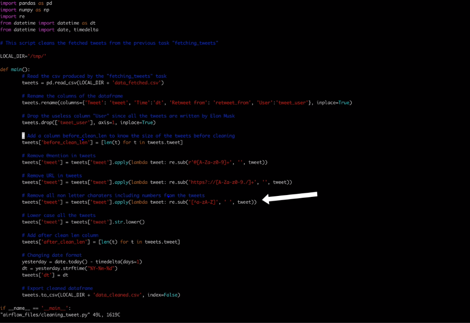


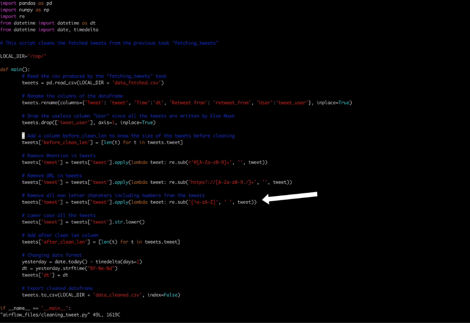


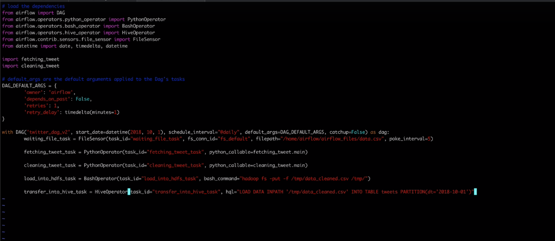


Now let’s write the DAG file for the twitter cleanup activity job:

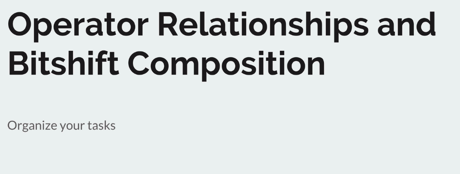
Below is the code for cleaning the tweet and scheduling the DAG



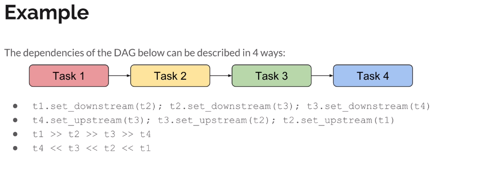




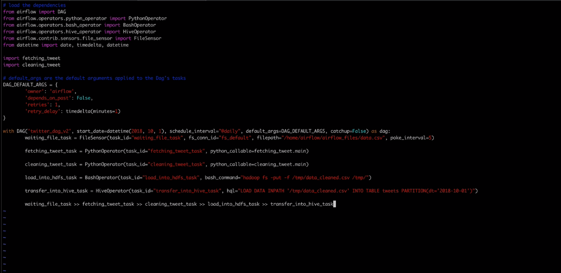
**Now we need to create dependencies:**



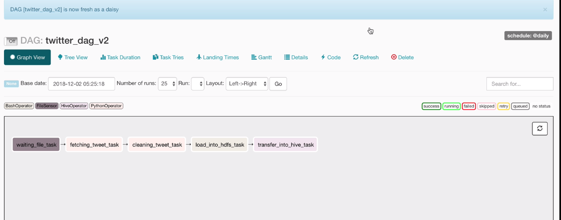
**Below is the dependency code:**



**Now let’s edit the previous DAG file with the dependencies:**

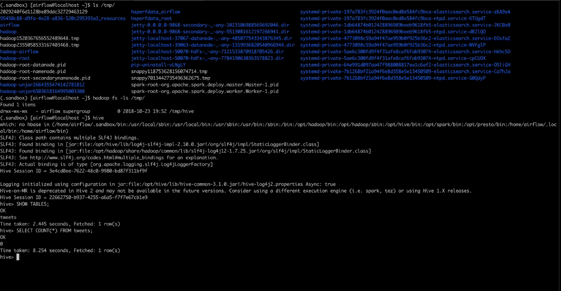


Below we can see how dependencies are made as soon as we save the dag file.

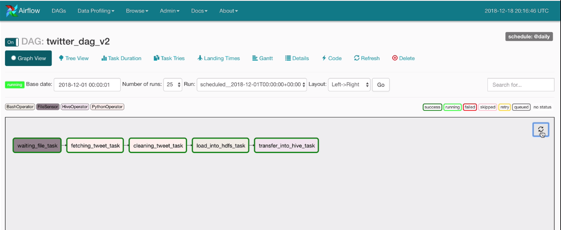


Let’s query Hadoop and Hive for the data

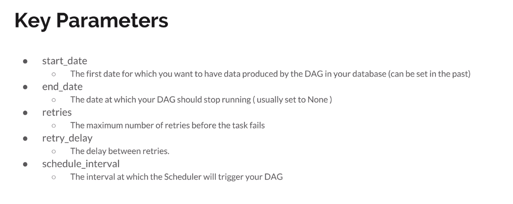
We can see that there is no data in Hive or in hadoop

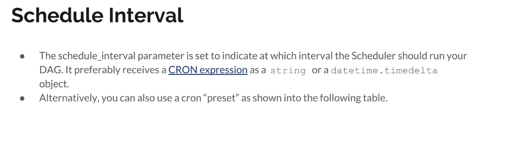


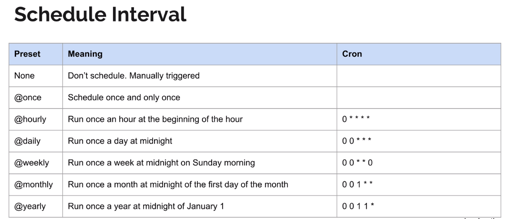
Now we can see the job completed.

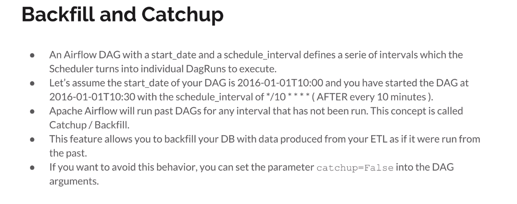


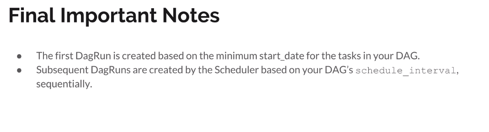
Key\_Parameters:











Below we can see how long does the scheduler take to scan for new DAG’s, we can edit this to change the time neded for the scheduler to check for new DAG’s.



What is a workflow?

It is the combination of what we saw so far:

