STEP 1 :- CREATED AWS USER AND IAM USERS

STEP 2:- What is Airflow?

<https://www.analyticsvidhya.com/blog/2020/11/getting-started-with-apache-airflow/>

Basically it is for workflow management. It has DAGs directed acrylic graph.

Installation step:- need python installed and pip installed. Then set a directory for airflow

* set AIRFLOW\_HOME=~/airflow
* pip3 install apache-airflow
* airflow db init

STEP 3:- SETTING UP TWITTER API

It will give you keys – store them safely

STEP 4: Install visual studio code, tweepy, pandas and s3fs. (s3fs is used to store write data into AWS S3 buckets)

MAJOR ISSUE RESOLVED:- *453 - You currently have Essential access which includes access to Twitter API v2 endpoints only. If you need access to this endpoint, you’ll need to apply for Eleva*

*ted access via the Developer Portal.*

So to resolve the above error I requested for the free elevated access to Twitter for developer account APIs.

STEP 5:- I converted the retrieved Tweets data into a structured JSON with only the relevant info and stored it into a csv using Pandas DataFrame

STEP 6:- Create an AMAZON EC2 instance to get deploy our AirFlow instance onto.

STEP 7: Then we have to connect to the EC2 instance and deploy our Airflow on it. Here I faced a big error. Airflow scheduler was not getting started so I followed steps mentioned in this link:-

<https://eliasbenaddouidrissi.com/how-to-setup-apache-airflow-on-ec2-free-tier/>

(if it still has errors, try to close every process and start the scheduler first and then the webserver)

STEP 8: Then create DAG in Python

STEP 9: Create an S3 bucket in AWS to save our data frame in S3