Kaushik Dey

Kaushikdey1984@yahoo.com

**Twitter Data Pipeline using Airflow**

1. Up & Running Ubuntu server via EC2 Instance and install the Airflow, pandas, s3fs.

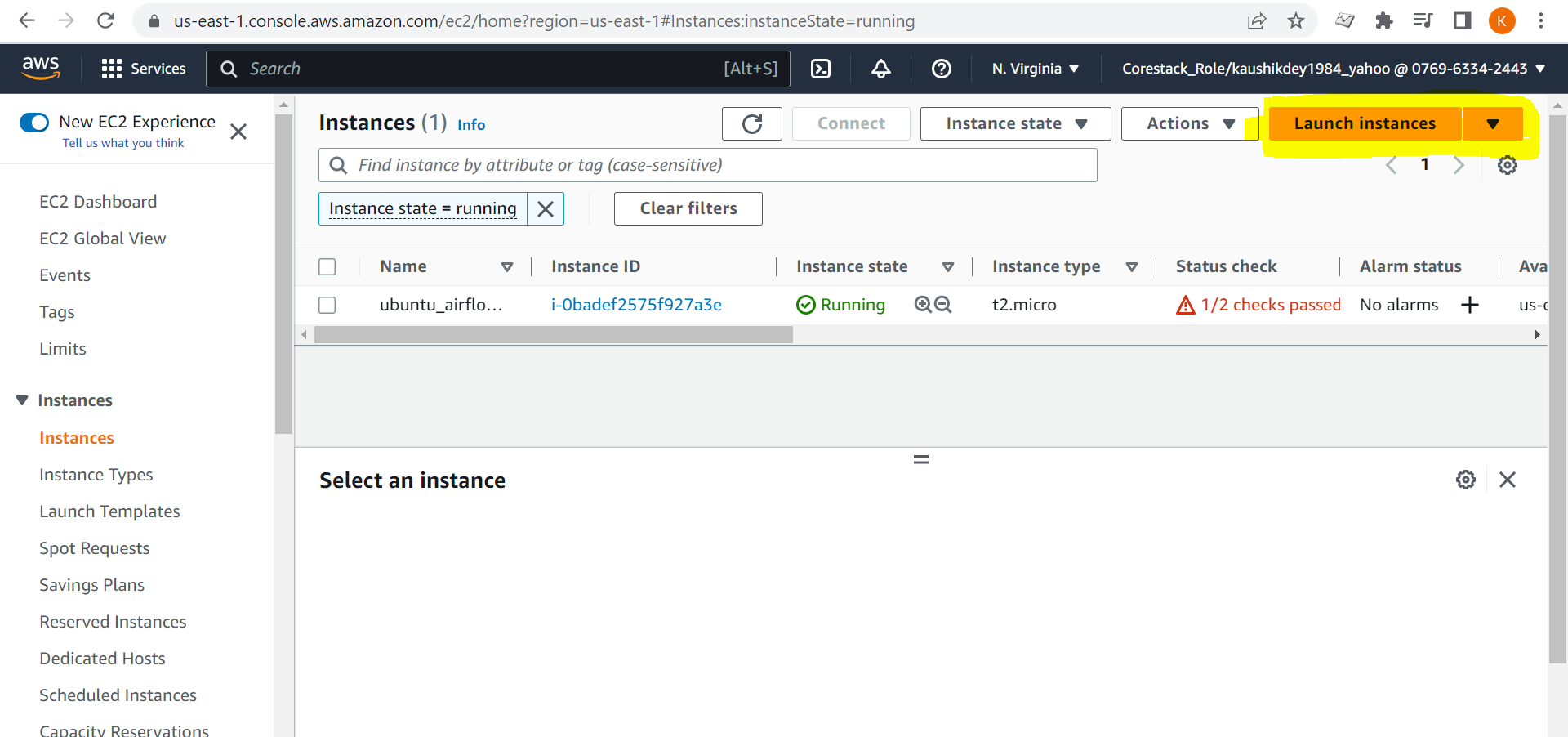
Step 1

First, we must create one ubuntu instance. So, we must take ec2 service from Aws. So, we must follow some steps.

Graphical user interface, email, website

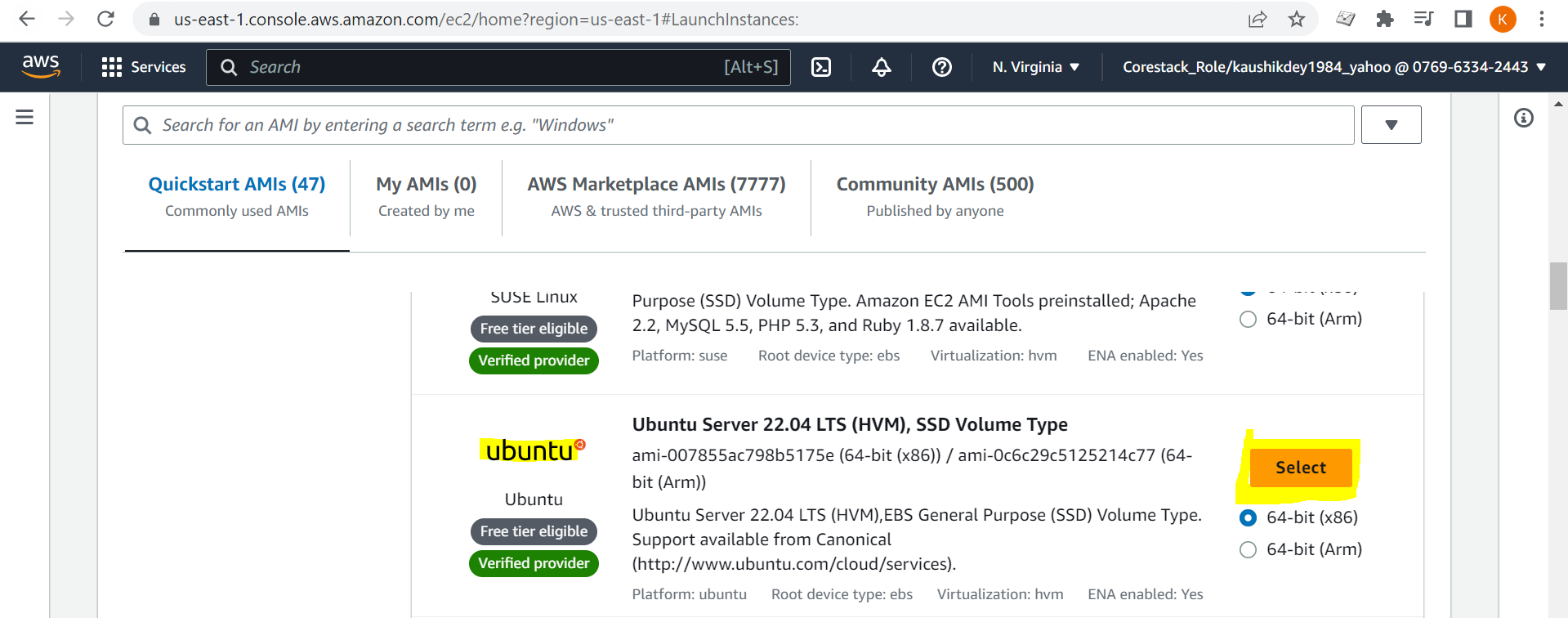
Description automatically generated

Step 2



Step 3

Now we must Browse AMI from the catalogue. The screenshot is given below. (Ubuntu Server 22.04 LTS)



Step 4

Now we must create new key pair for this instance.

Graphical user interface, text, application

Description automatically generated

Step 5

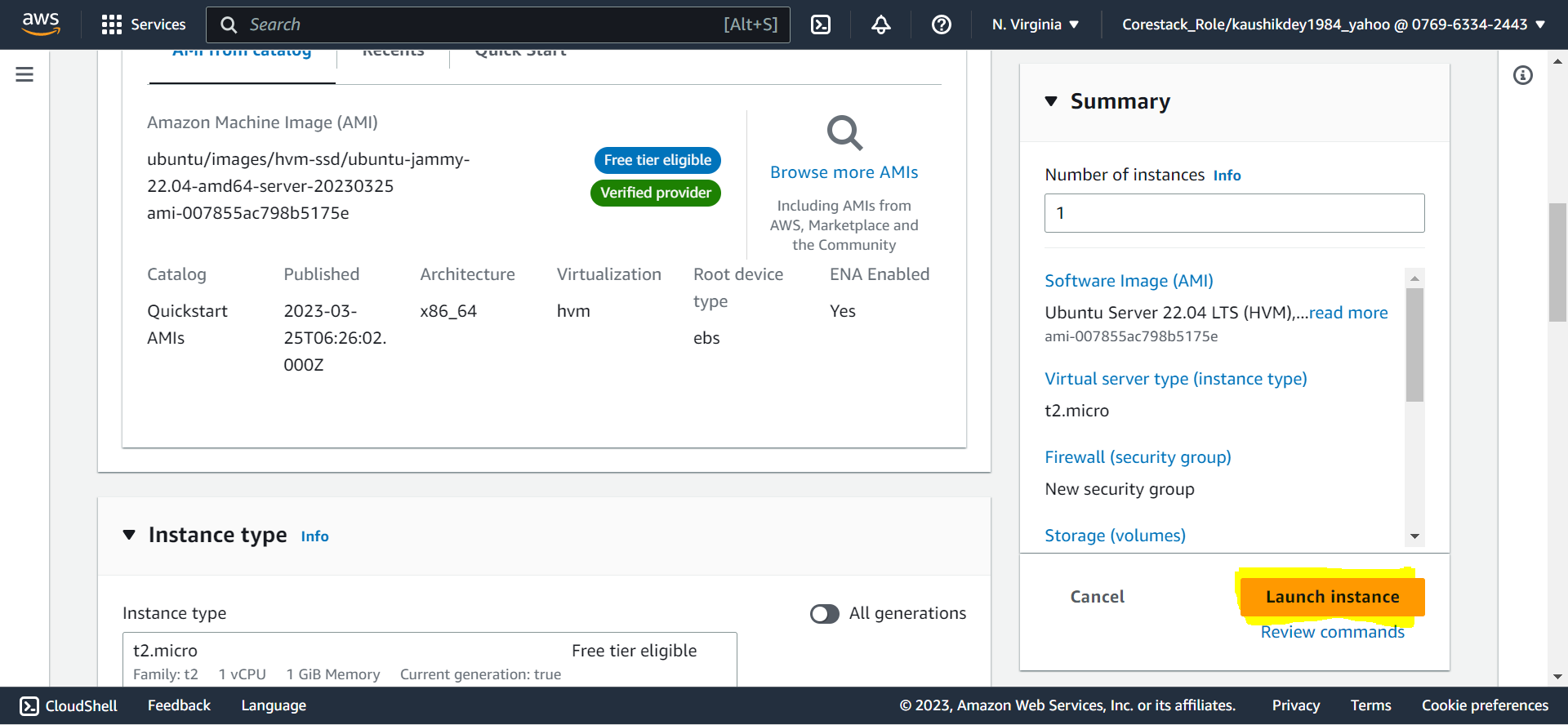
Now we can download the keypair (airflow\_server\_key\_pair.pem file). The rest of the settings will be as it is.

Graphical user interface, text, application

Description automatically generated

Step 6

Now we must launch our instance.

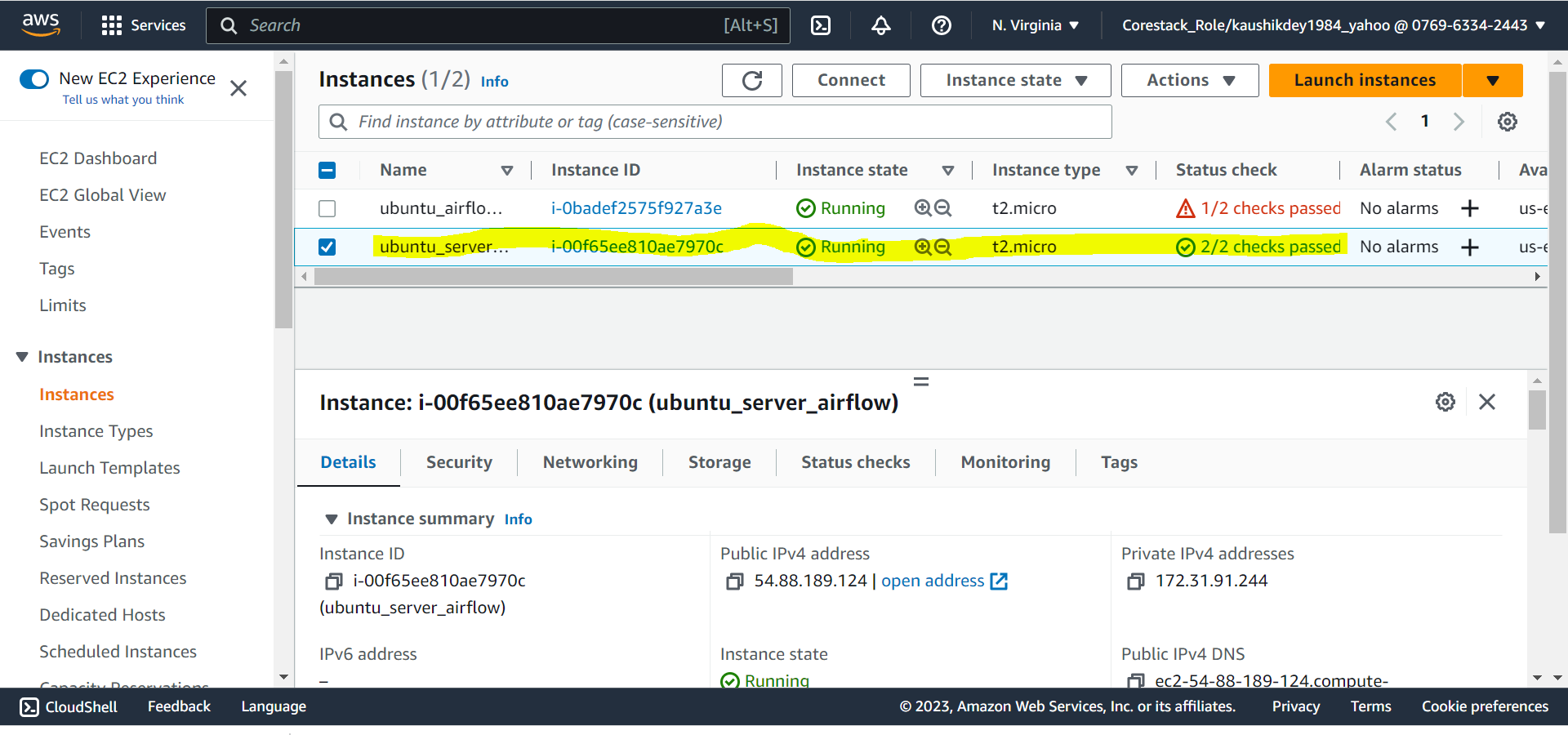


Step 7

Now we can see that our instance is created and up and running . so we can understand it from its status running. Now couple of things we have to highlighted.

Public IP Address: ec2-54-88-189-124.compute-1.amazonaws.com

Private IP Address : 172.31.91.244



Step 8

Now we have to connect our instances via browser or via ssh client.

Browser Based Access ( Not recommended)

Graphical user interface, application

Description automatically generated

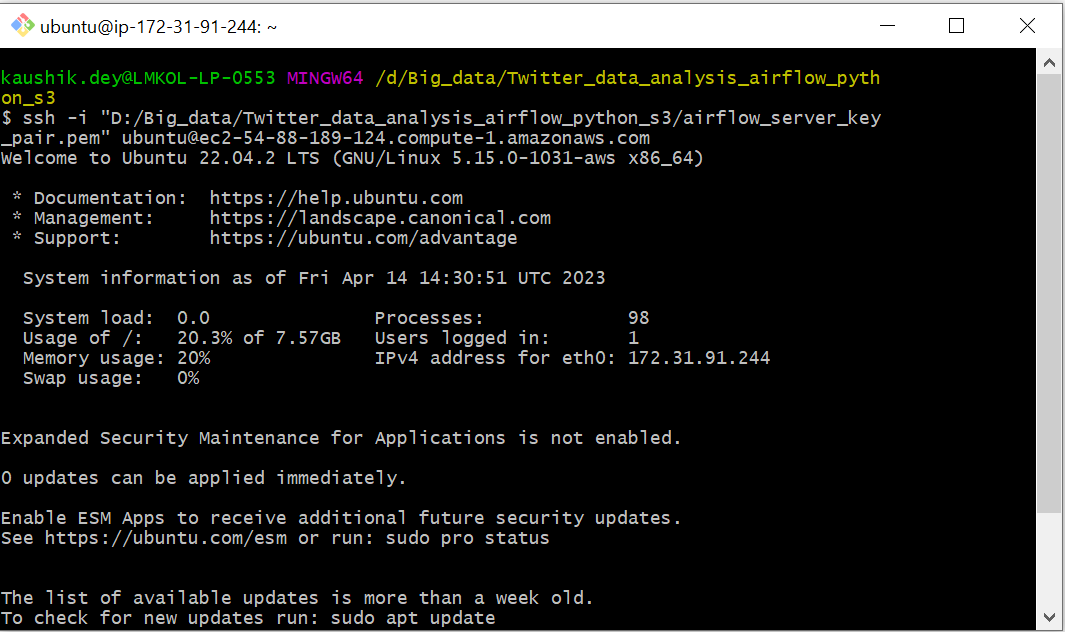
Text

Description automatically generated

SSH Client Based Access (Recommended)

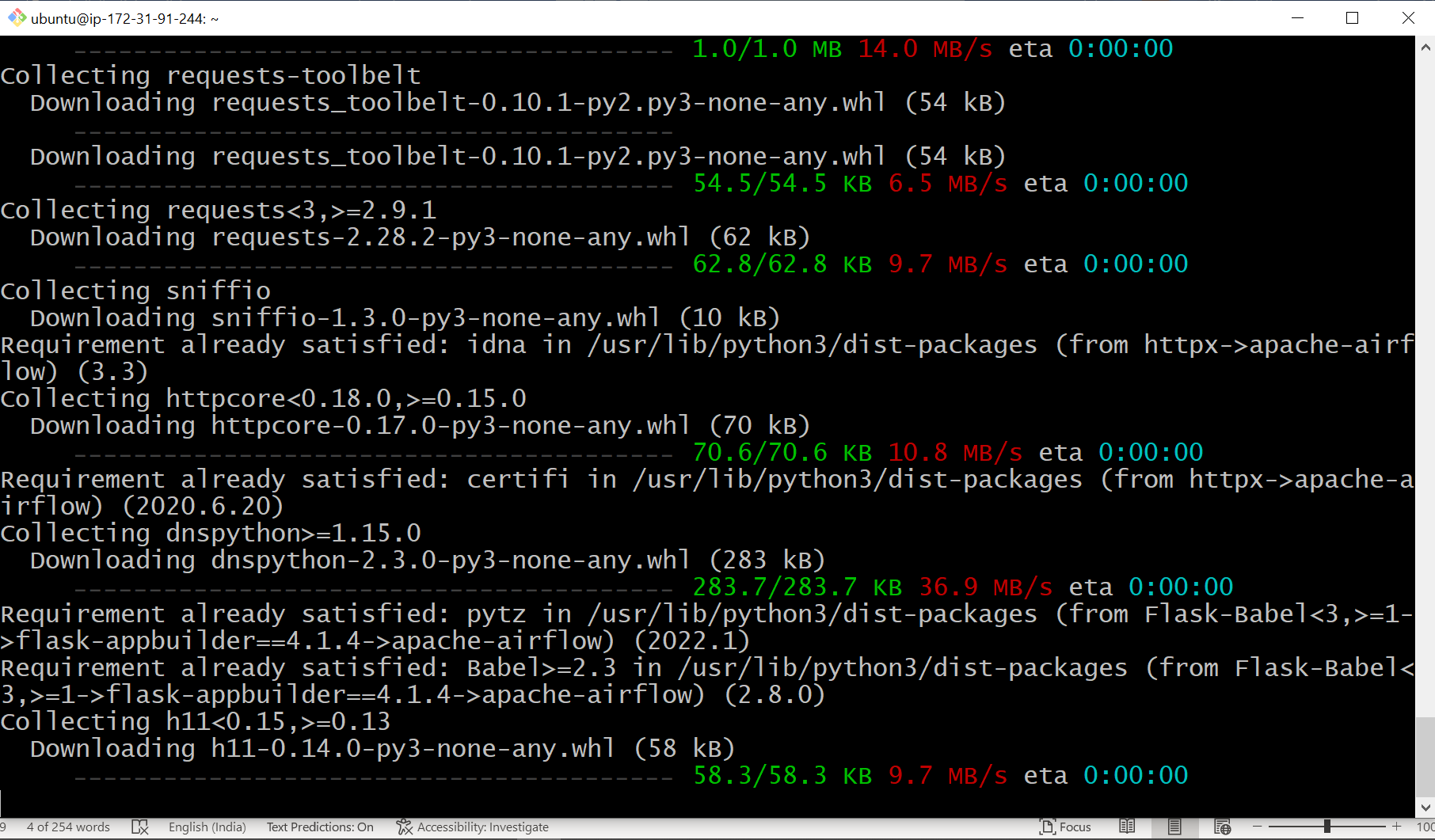
The command we must run in windows git bash shell, internally it runs Linux command. Just run this command.

ssh -i "D:/Big\_data/Twitter\_data\_analysis\_airflow\_python\_s3/airflow\_server\_key\_pair.pem" ubuntu@ec2-54-88-189-124.compute-1.amazonaws.com



Step 9

Now we must install couple of ubuntu commands to establish airflow running.

* sudo apt-get update.
* Sudo apt install python3-pip
* Sudo pip install apache-airflow
* Sudo pip install pandas
* Sudo pip install s3fs
* Sudo pip install tweepy ( twitter api )

Installation done following this way

Step 10

Now we have to up & running airflow in browser from ubuntu system. So, following two commands are there.

Airflow & airflow standalone (not recommended for production, its used for development) only two things we must know.

Login with

username: admin

password: sfkPzwUKBNsyMbG9

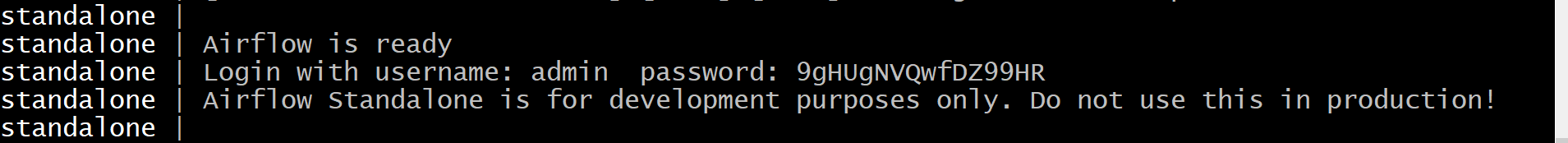
Graphical user interface, chart

Description automatically generated

Only one thing should follow with this message.

Step 11

Now our airflow is ready with that following screen-shots.



Step 12

Now the url of this airflow ui is as following.

The port 8080 is for airflow default port.

our ec2 public Ip is 54-88-189-124.

Public ipv4 DNS is http://ec2-54-88-189-124.compute-1.amazonaws.com

Airflow url is <http://ec2-54-88-189-124.compute-1.amazonaws.com:8080/>

Step 13

Now we have to open the port from security group. We have to all traffic , ssh, http and https , the following screenshot is given below.

Graphical user interface

Description automatically generated

Step 13

Airflow login

URL : <http://ec2-54-237-236-91.compute-1.amazonaws.com:8080/login/?next=http%3A%2F%2Fec2-54-237-236-91.compute-1.amazonaws.com%3A8080%2Fhome>

Graphical user interface, application

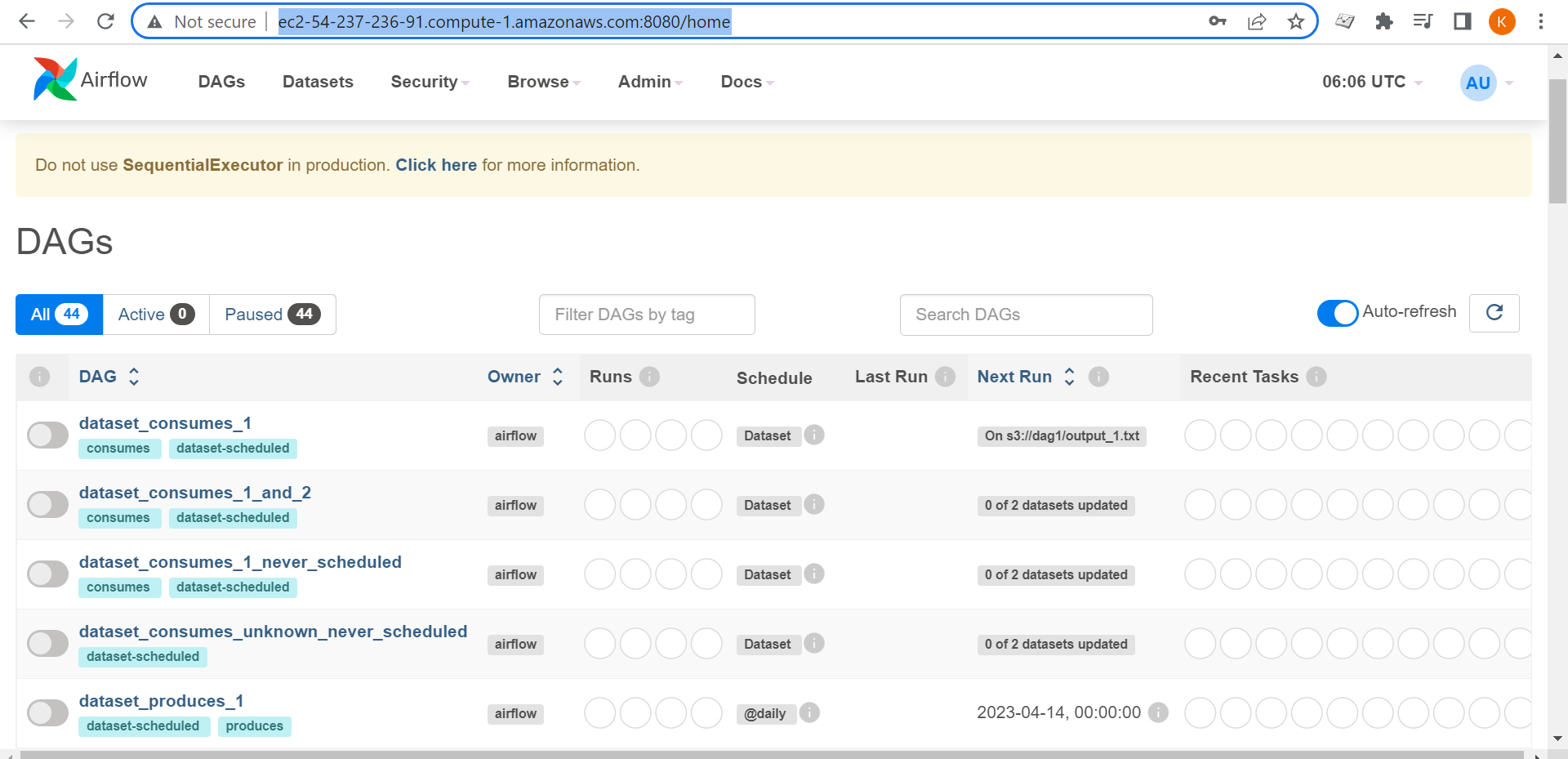
Description automatically generated

Username: admin

Password: 98gKZFDKxSAaHnC7

Step 14

DAG LINK : <http://ec2-54-237-236-91.compute-1.amazonaws.com:8080/home>



Step 15

S3 bucket creation

Graphical user interface, text, application, email

Description automatically generated

Step 16

Now we have to login in twitter account.

URL <https://developer.twitter.com/en/portal/projects/1647044527285477377/apps/26931143/keys>

you must generate access key & secret key (

access key, access secret, consumer\_key && consumer\_secret )

Graphical user interface, application

Description automatically generated

Step 17

We have to create twitter\_etl.py and twitter\_dag.py files and upload inside the airflow. So some settings need to be changes .

1. First need to check airflow folder structure.

Text

Description automatically generated

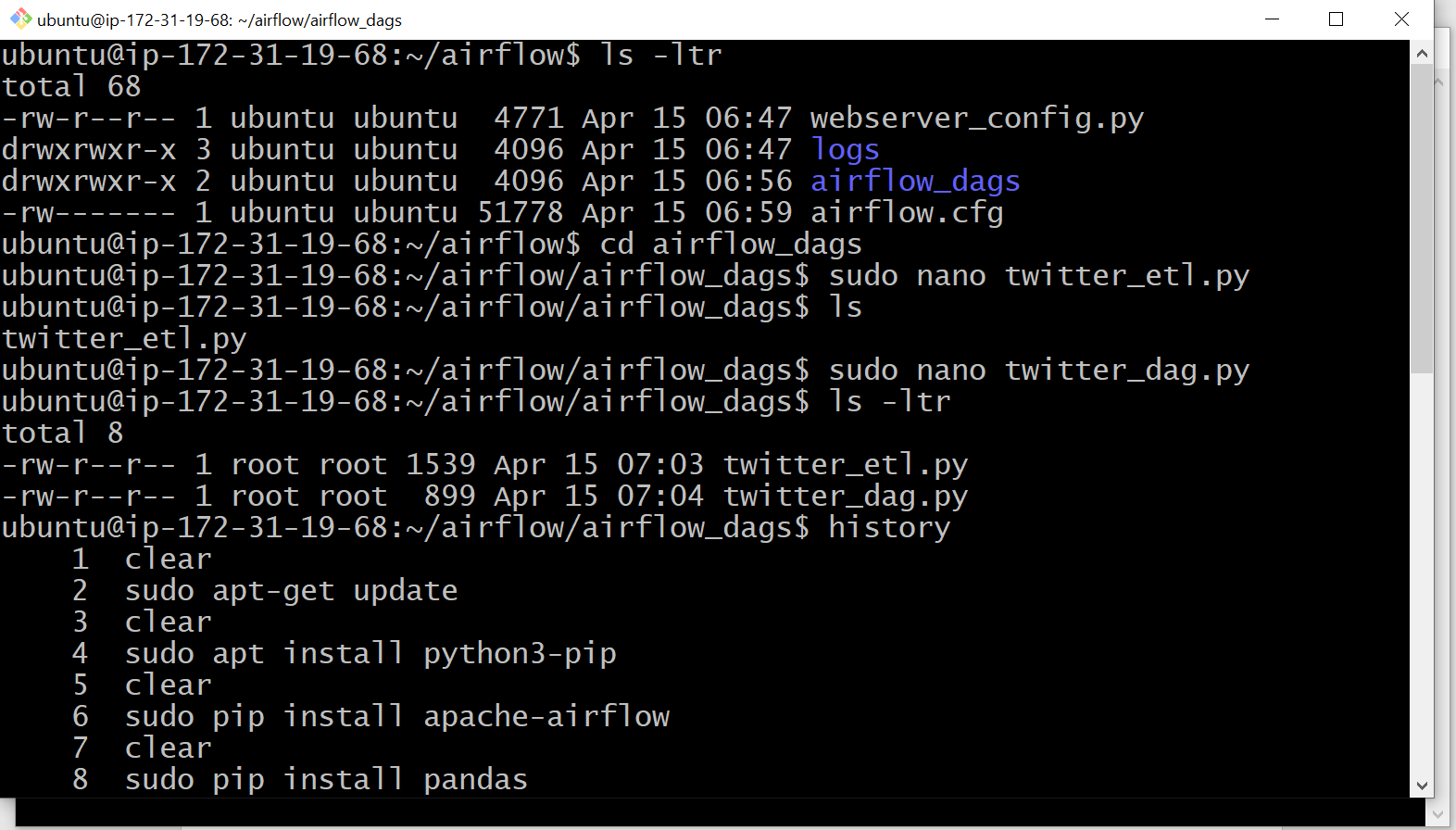
1. We must edit the airflow .cfg file in order to add our own airflow dag folder and create the ETL. We must change the dag folder name.

Text

Description automatically generated

Need to add airflow\_dags folder inside airflow.cfg file.

dags\_folder = /home/ubuntu/airflow/airflow\_dags .

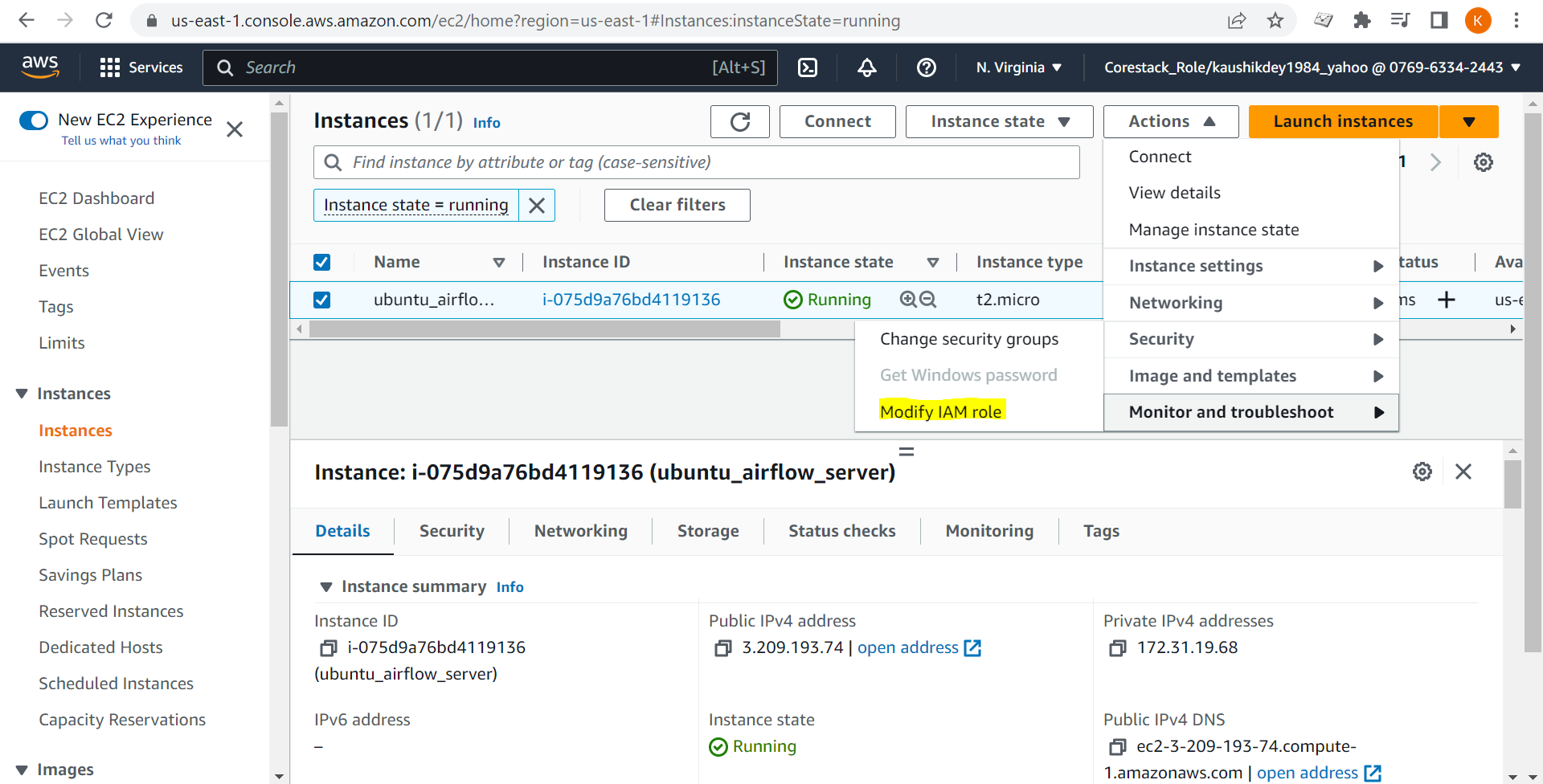


Step 18

Now restart the airflow server to take effect the new dag changes. The following screenshot is given below.

Step 19

To access the s3 bucket inside EC2 instance we have to add modify IAM role. The screenshot is given below.



Graphical user interface

Description automatically generated with medium confidence

We must provide two access. (AmazonS3FullAccess && AmazonEc2FullAccess)

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Now we have to create one role , the screenshot is given below.

Graphical user interface, text, application, email

Description automatically generated