National College of Ireland

Configuration Manual

MSc Research Project
MSc Cloud Computing

Sumit Kumar Sahoo
Student ID: 21154589

School of Computing
National College of Ireland

Supervisor: Sean Heeney

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sumit Kumar Sahoo
Student ID:	21164967
Programme:	Msc in Cloud Computing **Year:** 2022-2023
Module:	Msc Research Project
Supervisor:	Sean Heeney
Submission Due Date:	01/01/2023
Project Title:	Open-source ETL Framework using Big Data tools & Orchestration on AWS Cloud Platform
Word Count:	1091 Lines **Page Count: 18**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: 

Date: 01/01/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Sumit Kumar Sahoo
Student ID: 21164967

Open-source ETL Framework using Big Data tools and Orchestration on AWS Cloud Platform

1 Introduction

This is a configuration manual for Setting out Tools , Software and AWS services used for Developing a Big Data ETL framework using Python, Pyspark and Apache airflow for orchestration. We have used Terraform for Infrastructure as Code. We have used S3 bucket for using it as Source and destination and Redshift For Data warehouse Data Analytics. We have used CloudTrail for Auditing and AWS price calculator for calculating Total Cost of Operation.

2 Tools and Softwares Used

1. Visual Studio Code IDE
2. GitBash
3. Python – 3.8
4. Terraform => 0.12
5. PySpark
6. Apache Airflow – 2.2
7. Aamazon Webservices

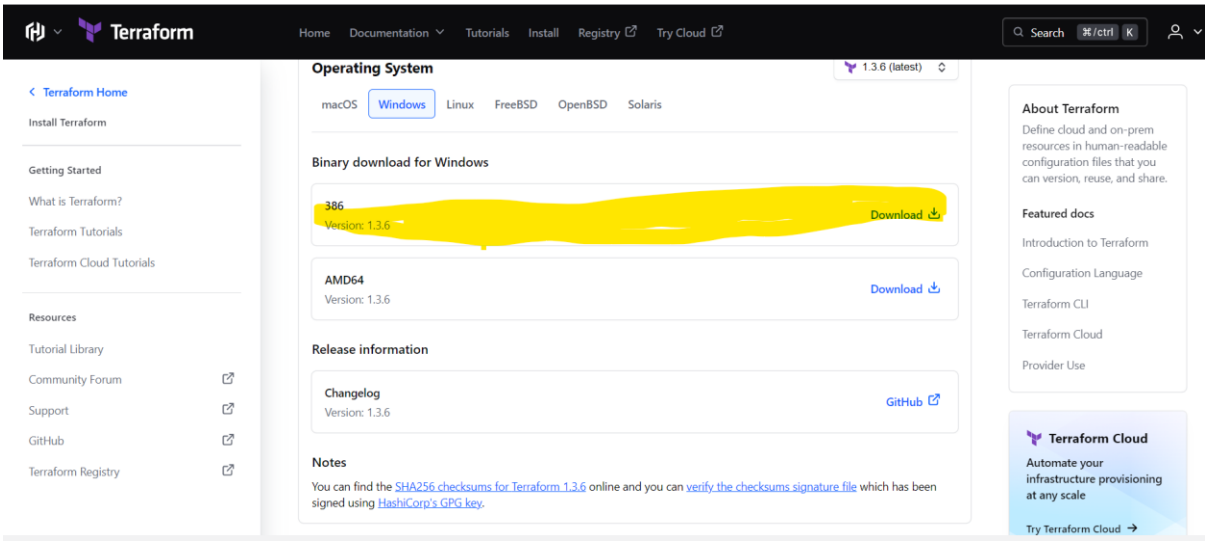
3 Software Installation

3.1 Visual Studio Code –

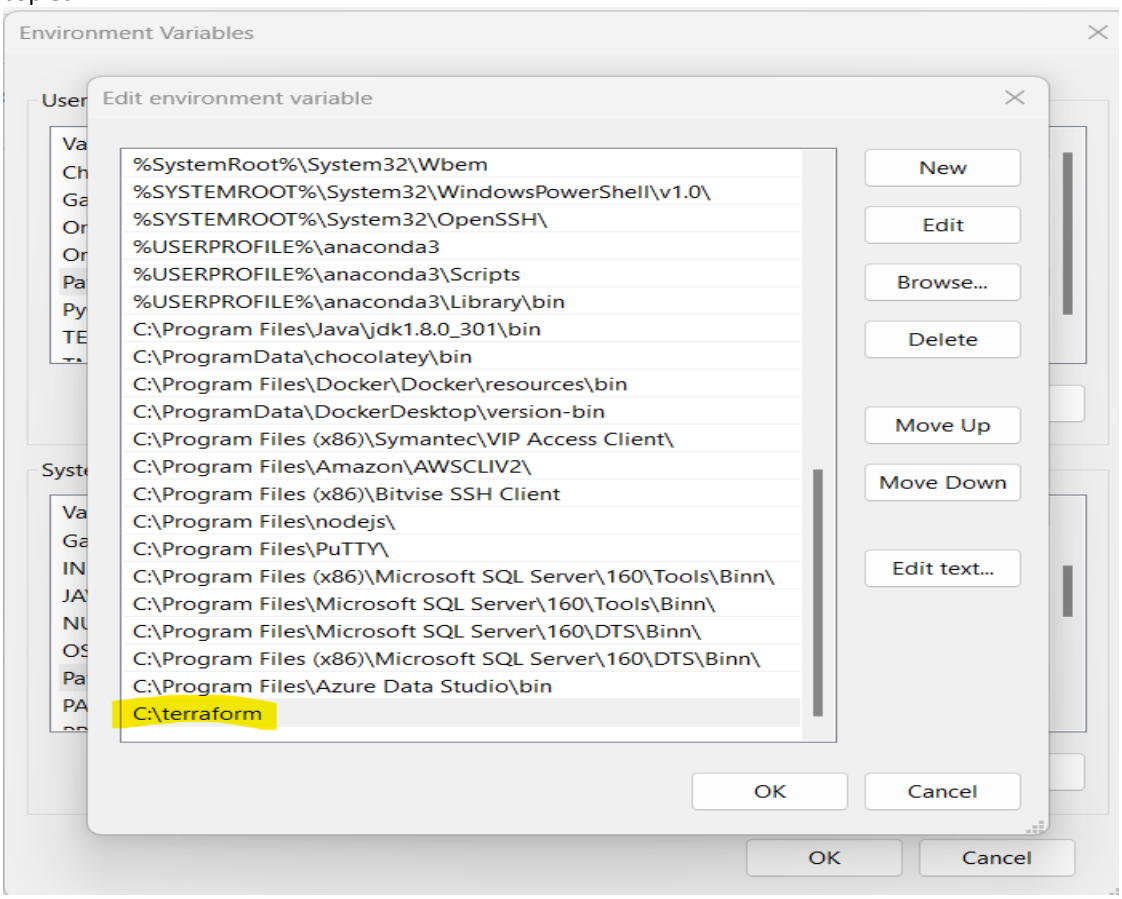
1. Download Visual Studio Code for Windows using https://learn.microsoft.com/en-us/visualstudio/install/install-visual-studio?view=vs-2022

3.2 Setup Terraform

1. Download Terraform for Windows using - https://developer.hashicorp.com/terraform/downloads

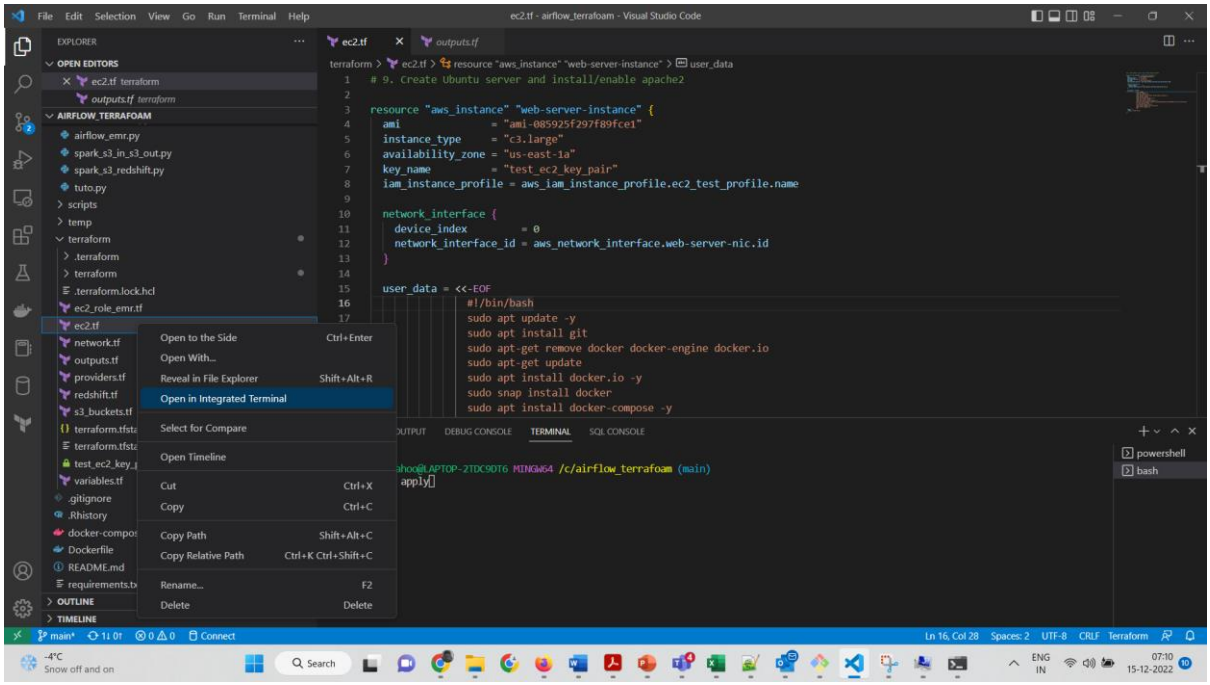


2. Extract the .zip file and store the folder in C:\ Drive
3. Setup environment and user variables by Editing the Path and pasting the Path where the above folder is
copied.

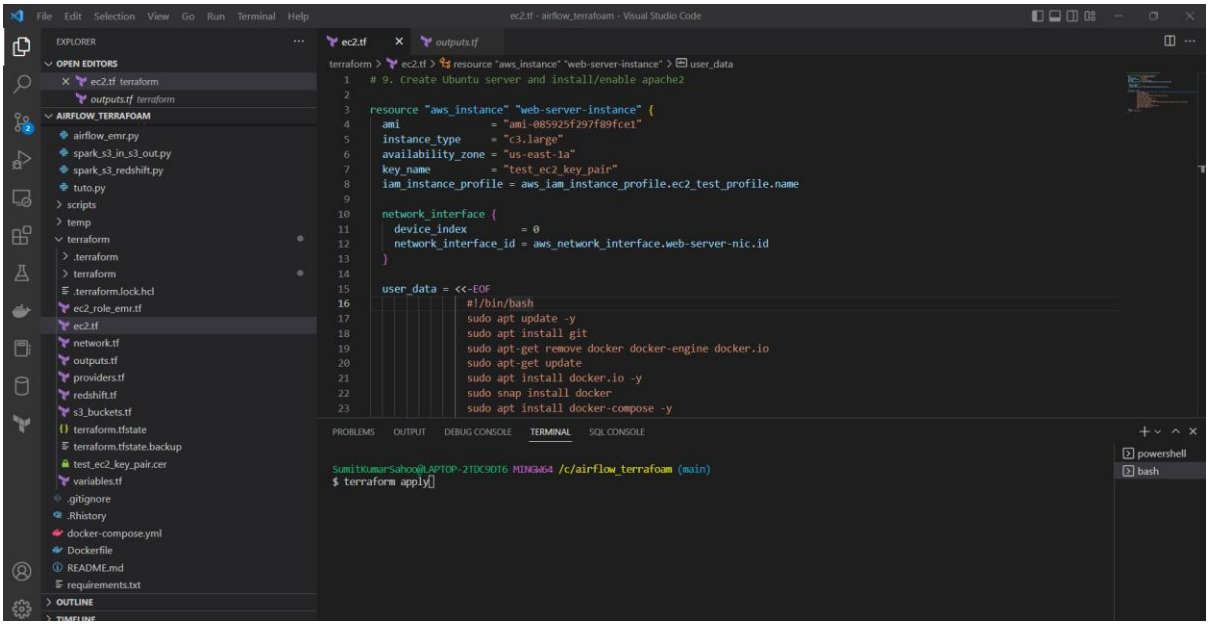


4. Goto the Visual Studio and open the Project and right click on ec.tf file -> Click on Open in integrated terminal -> cmd and run terraform plan

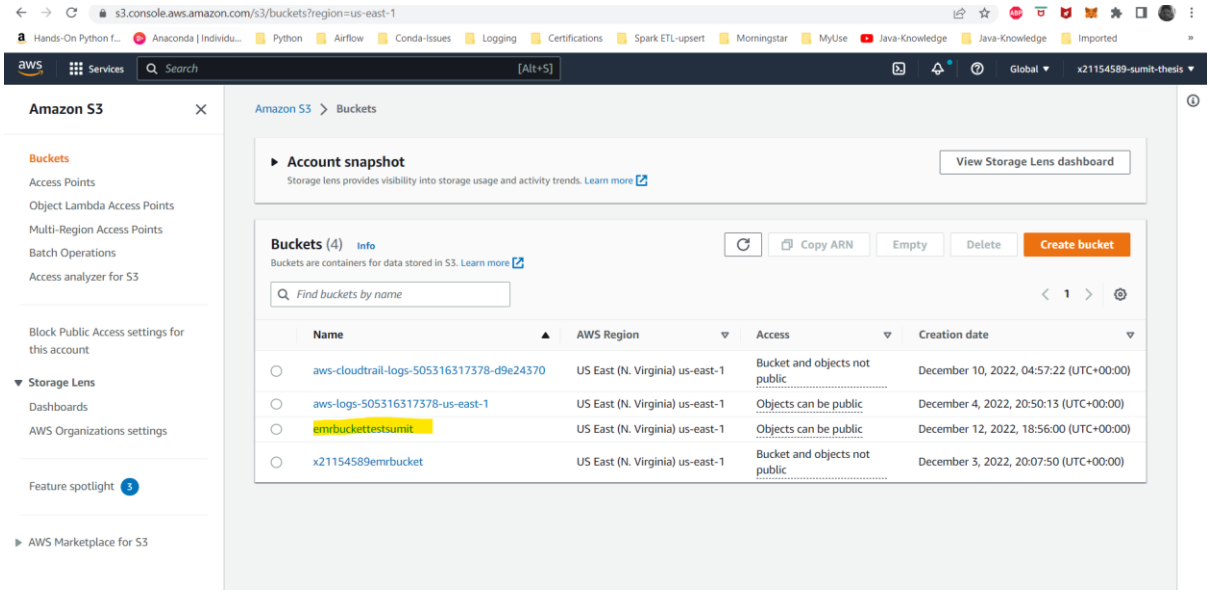
4 Deployment of EC2 which hosts Apache Airflow and S3 in us-east-1



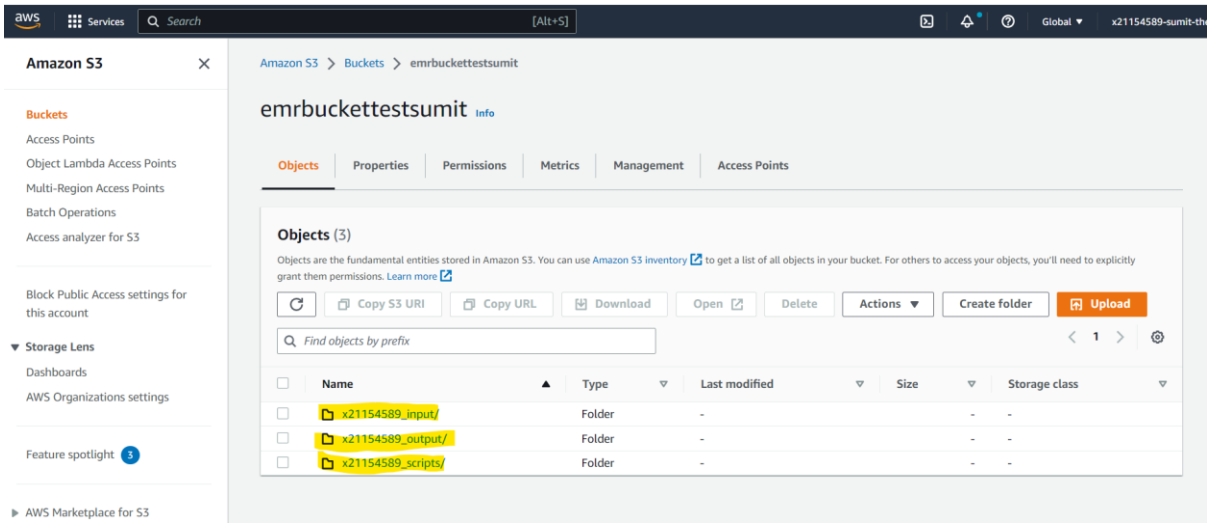
1. Run command -> terraform plan



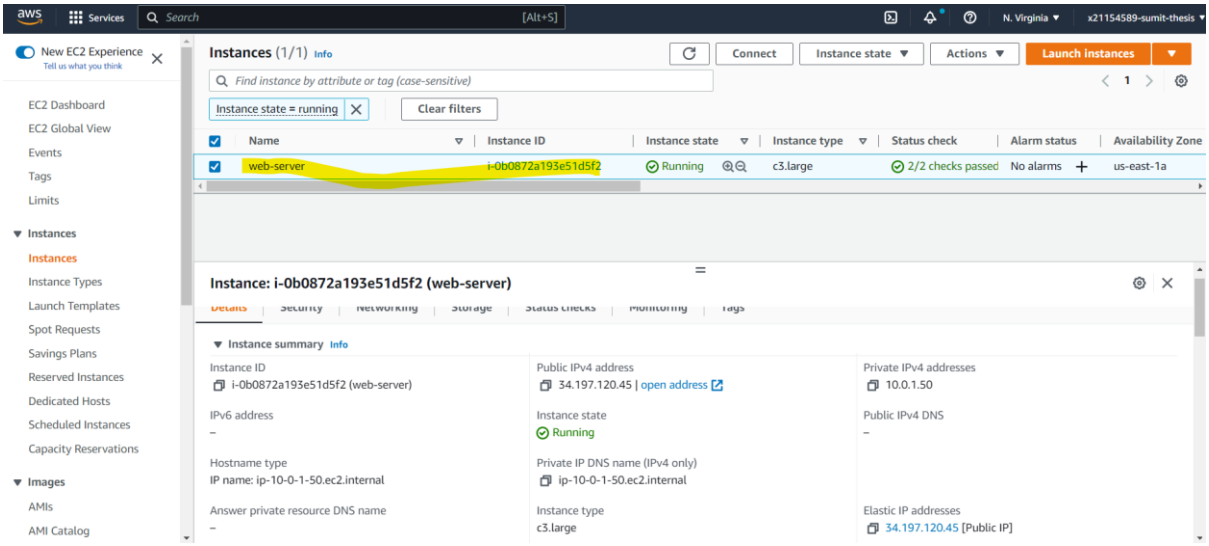
2. This will start automated deployment of :
    a. S3 Bucket – emrbuckettestsumit



    b. Create files as shown below

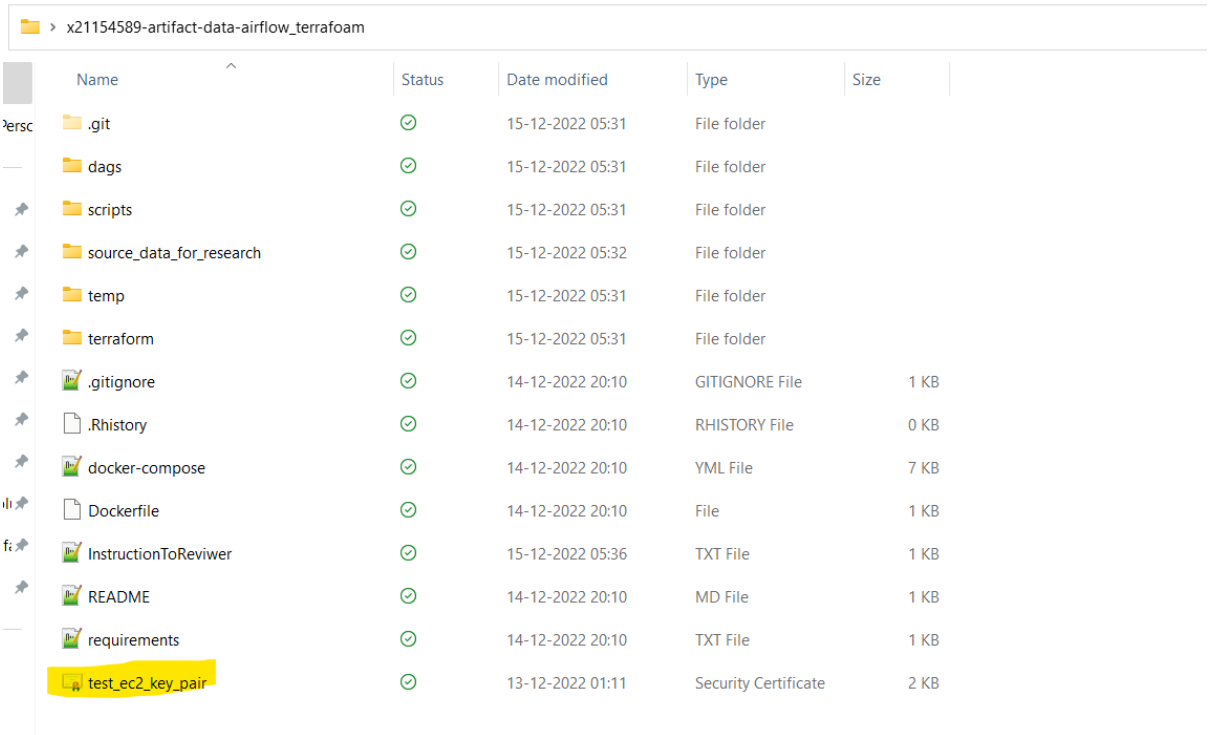


    c. Create EC2 c3.large in us-east-1



d. The terraform script will pull the code from github repository - https://github.com/sumitkumarsahoo-x21154589/airflow_terrafoam.git and clone the data inside the EC2 webserver spin up above

e. We can check it using ssh inside the EC2 server using git bash ssh by goin to the folder having test_ec2_key_pair.cer file



f. The ssh can done using following command:

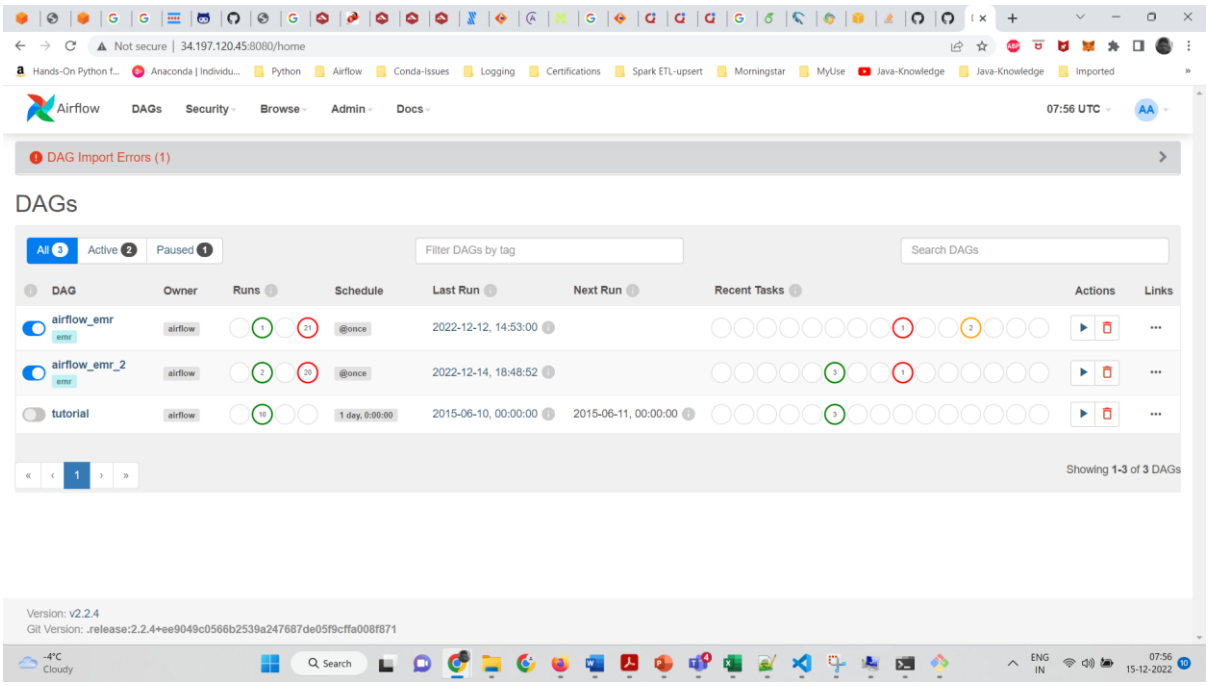
```
>> ssh -i test_ec2_key_pair.cer ubuntu@3.214.183.171
```

g. We can verify entire code is present or not in the EC2 using the command and make sure the root access is given to the folder.

>> **cd /** .

3. Now the Apache Airflow web server is up and running on the EC2 instance it can be checked using the link http://34.197.120.45:8080/home

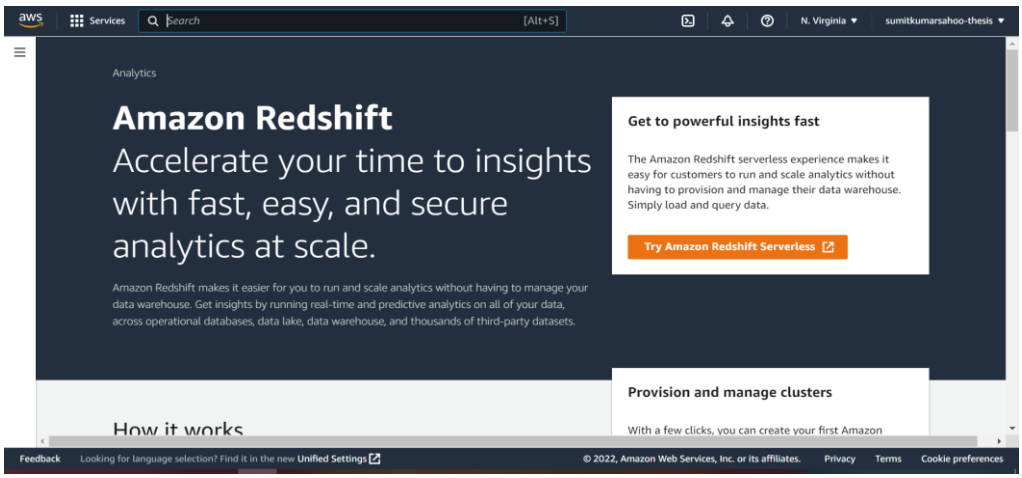
Which is http://<publicipaddressofec2>/home



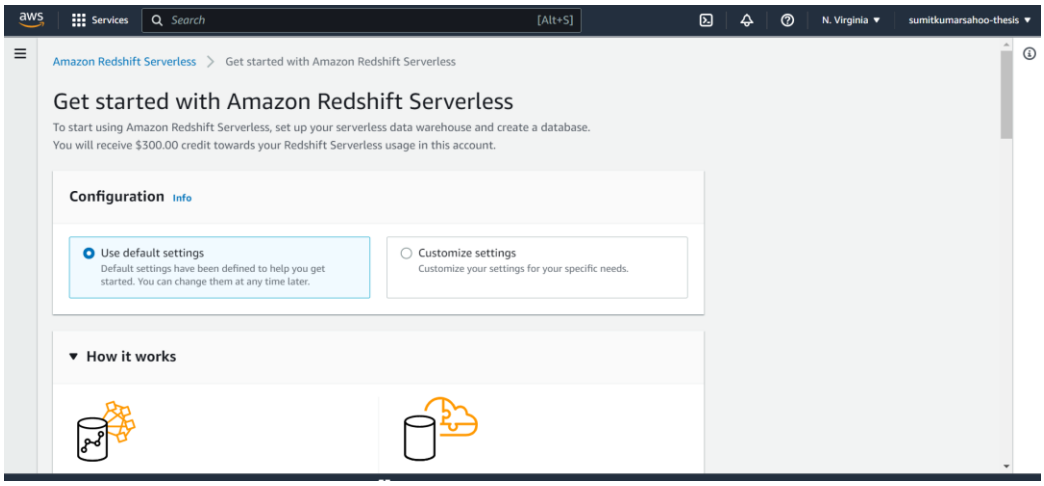
This completes the deployment steps of EC2 and S3

5 Setup AWS Services

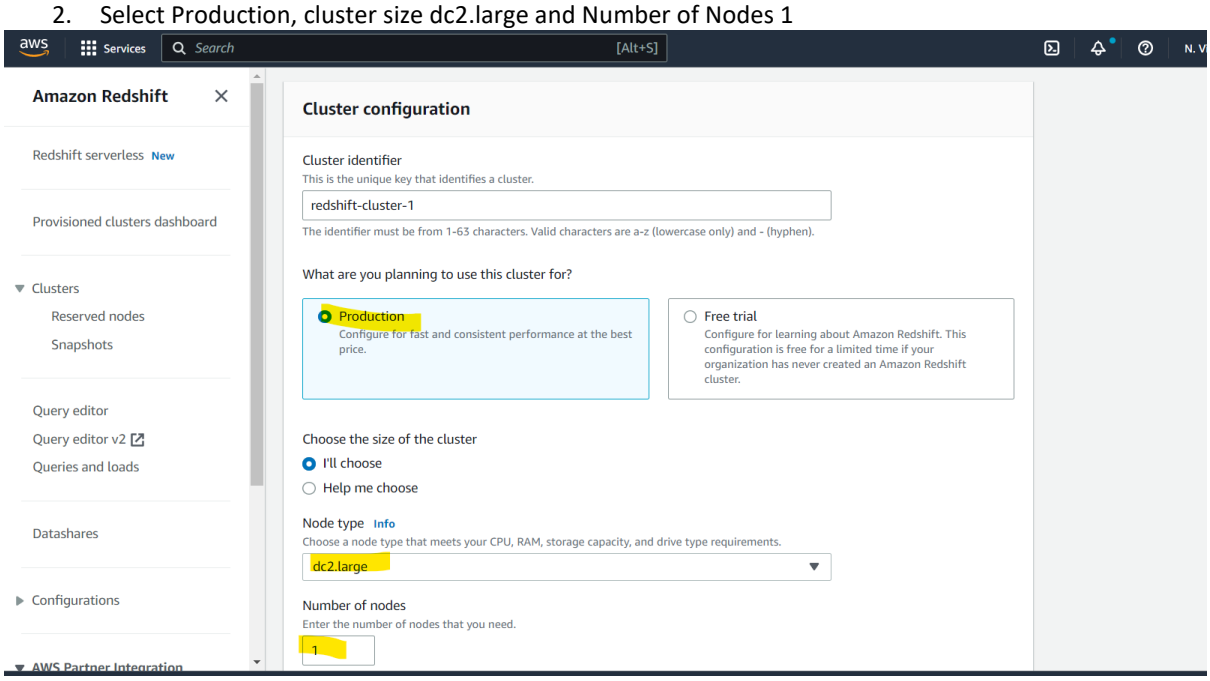
5.1 AWS redshift



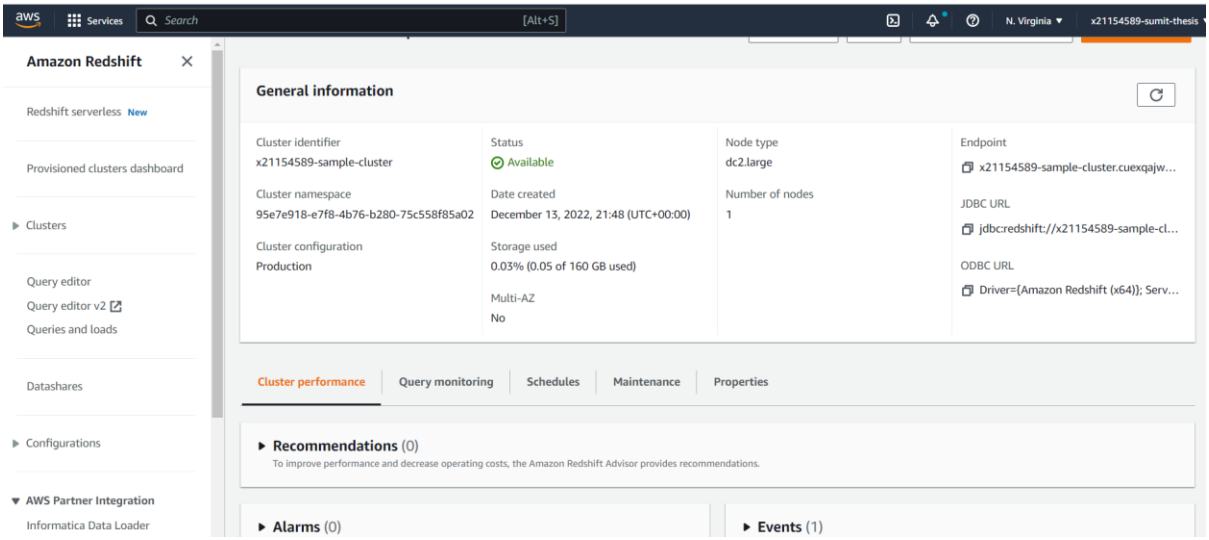
1. Search for Redshift & Create Cluster



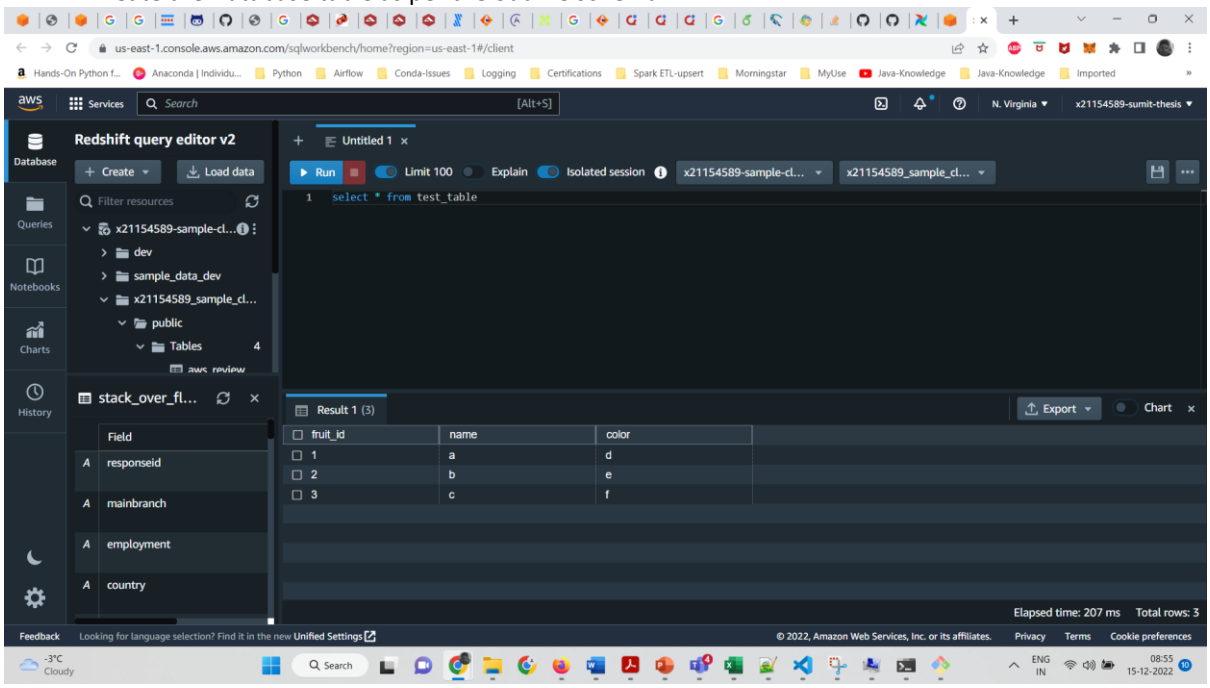
2. Select Production, cluster size dc2.large and Number of Nodes 1



3. Click on Create Cluster with other deafult config. This creates Cluster



4. Reate the Database table as per the out file schema

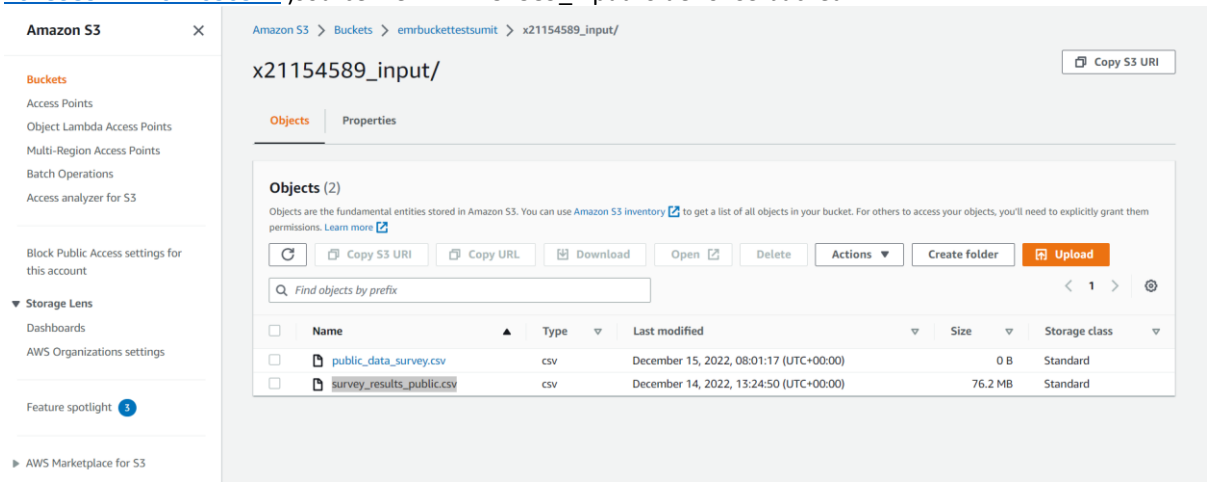


6 Running Apache Airflow DAG

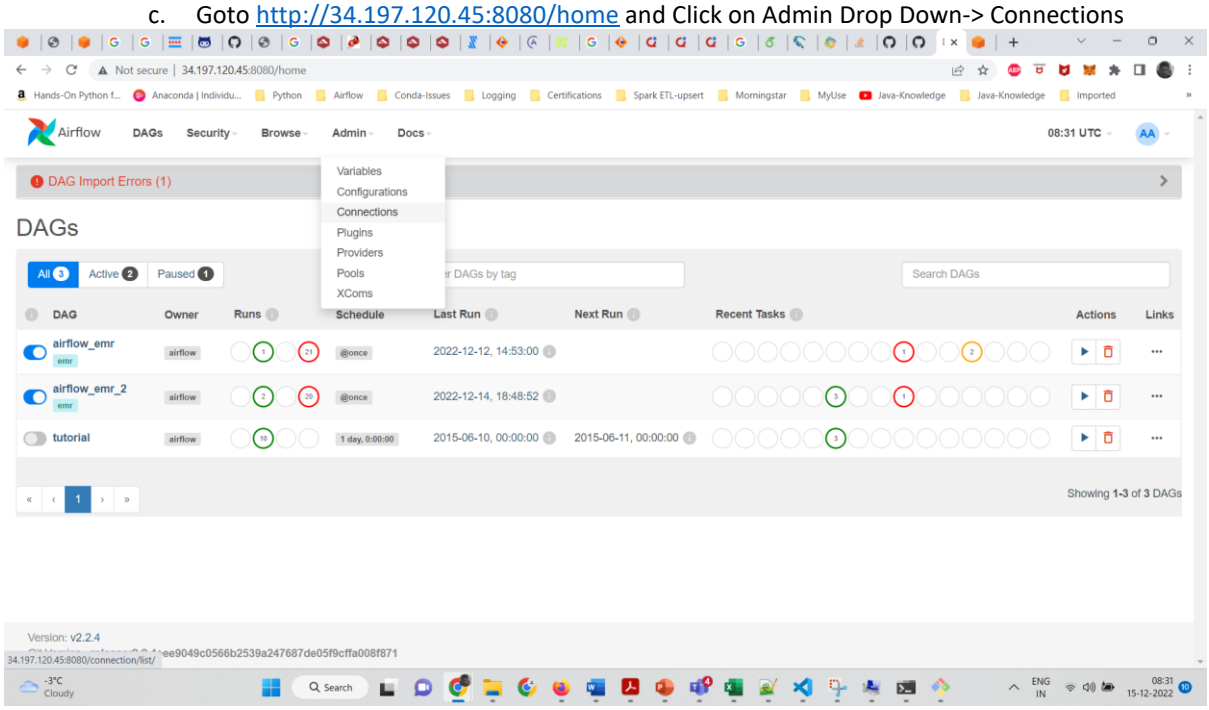
1. In Visual Code Studio, after development changes are done, they can be pushed to GitHub using the command in the terminal:
   git remote add origin https://github.com/sumitkumarsahoo-x21154589/s.git
   git branch -M main
   git push -u origin main

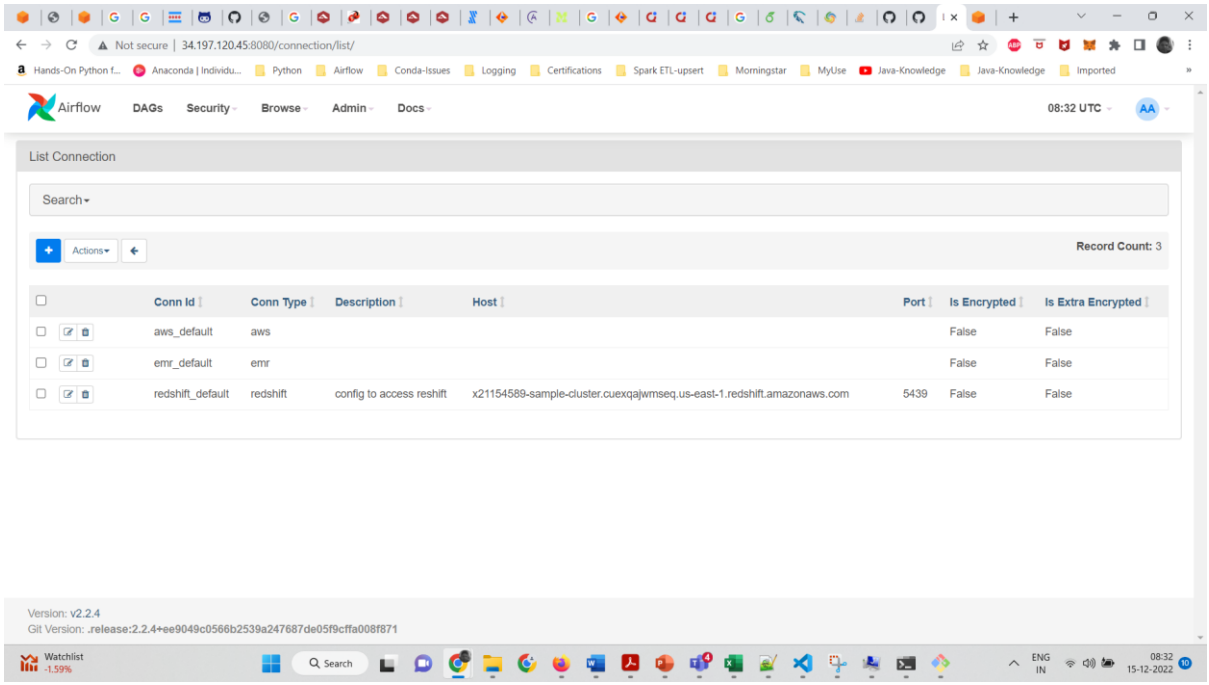
2. Prerequisite of running DAG,
   a. setting up source in S3.

b. Upload the survey_results_public.csv from Public Secondary Dataset (https://insights.stackoverflow.com/survey?_ga=2.91719770.170077947.1671050547-2078385242.1671050547)source file in x21154589_input folder of S3 bucket

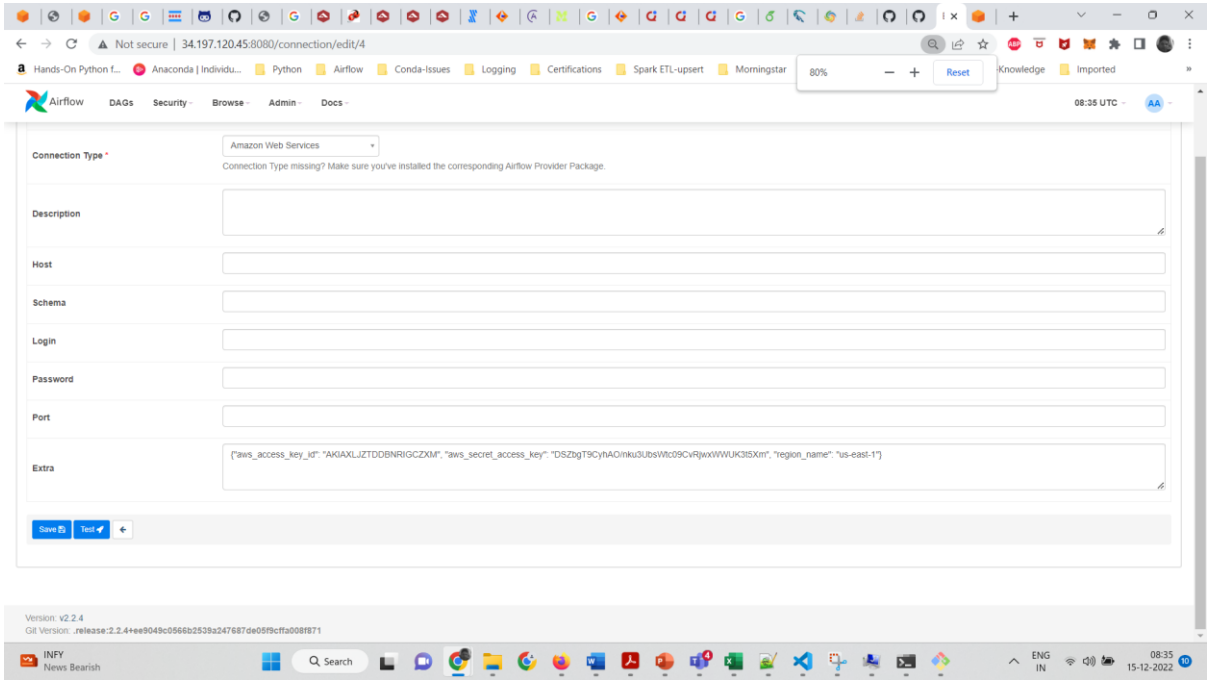


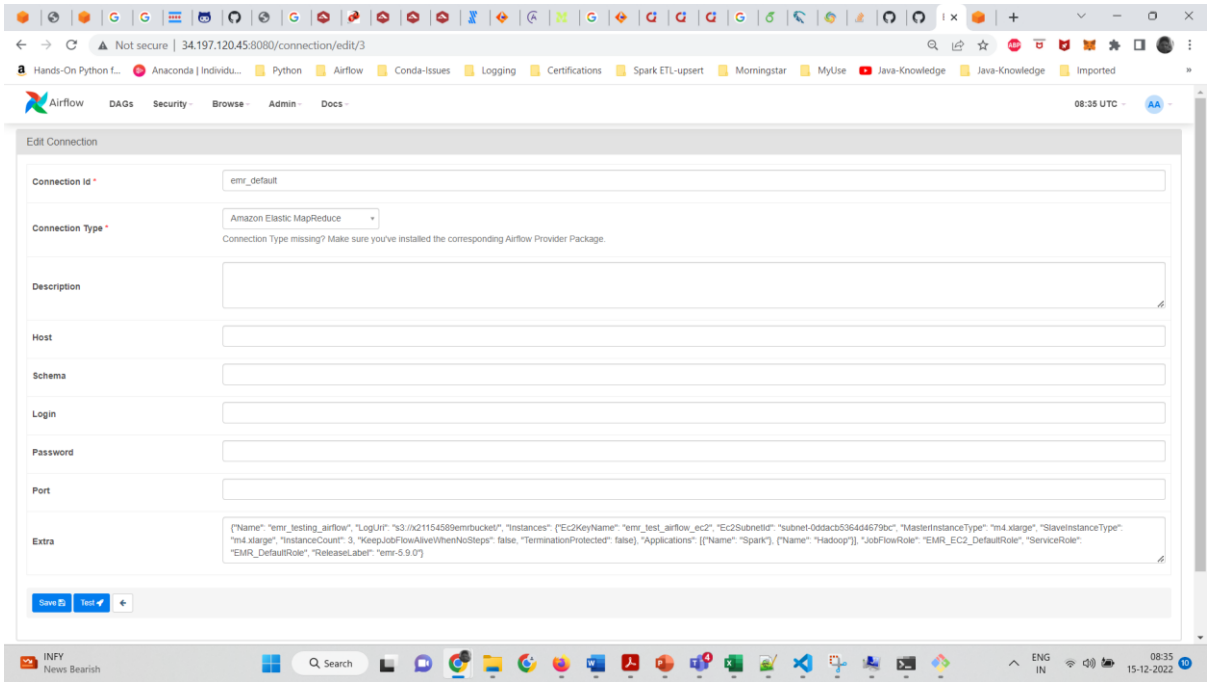
c. Goto http://34.197.120.45:8080/home and Click on Admin Drop Down-> Connections

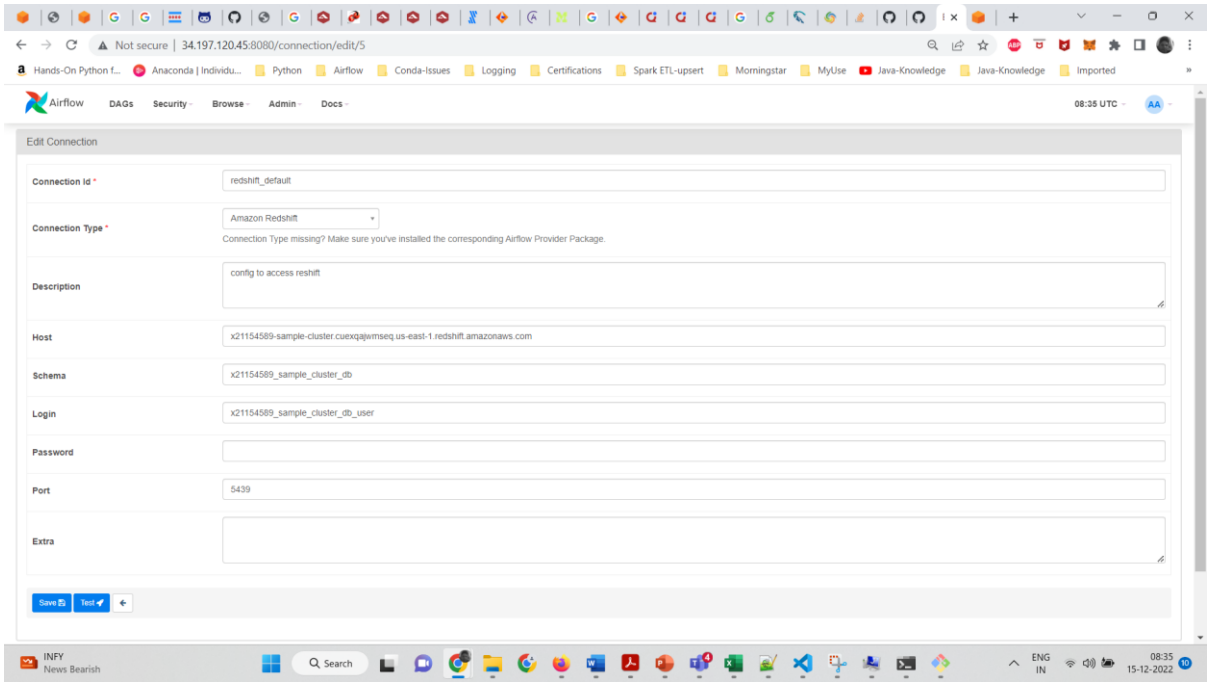




d. Setup connection string for AWS login , EMR spin up and close(for cost saving) and redshift connection to move proccesed data from S3 bucket emrbuckettestsumit/ 21154589_output/



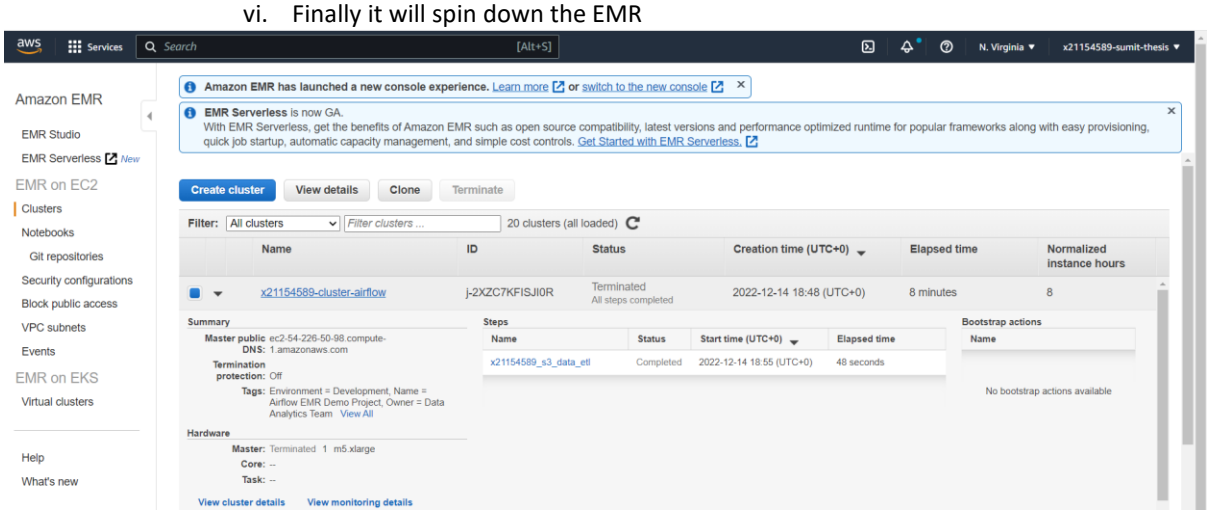




3. Now we are set to run dag

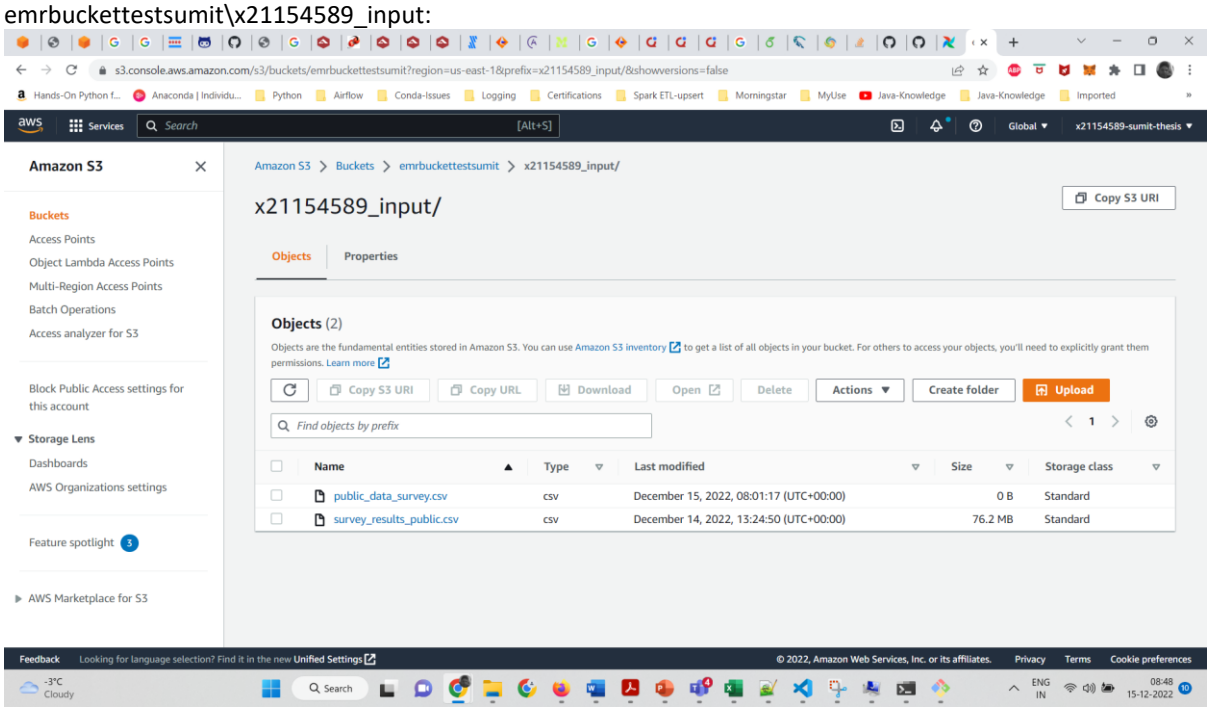
4. Running the DAG:
    a. Broef description of what the DAG will do:
        i. First it will pull data from S3 bucket emrbuckettestsumit/ 21154589_input/ survey_results_public.csv
        ii. Then it will spin up EMR cluster
        iii. It will run the PySpark code for transformation on input dataset
        iv. Then it will do checks if the first two steps are successful or not
        v. It will store the processed data in the emrbuckettestsumit/ 21154589_output/ folder in parquet format

vi. Finally it will spin down the EMR

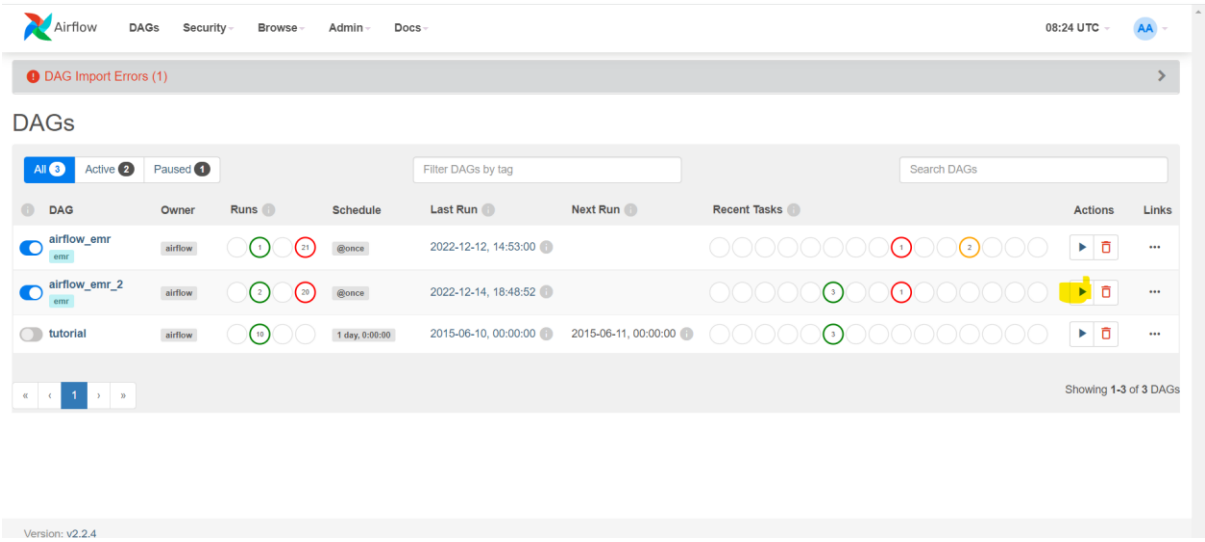


vii. The final step is the parquet file will be moved to Redshift for data analysis

b. Before running the DAG we need to move the source data into s3 bucket emrbuckettestsumit\x21154589_input:

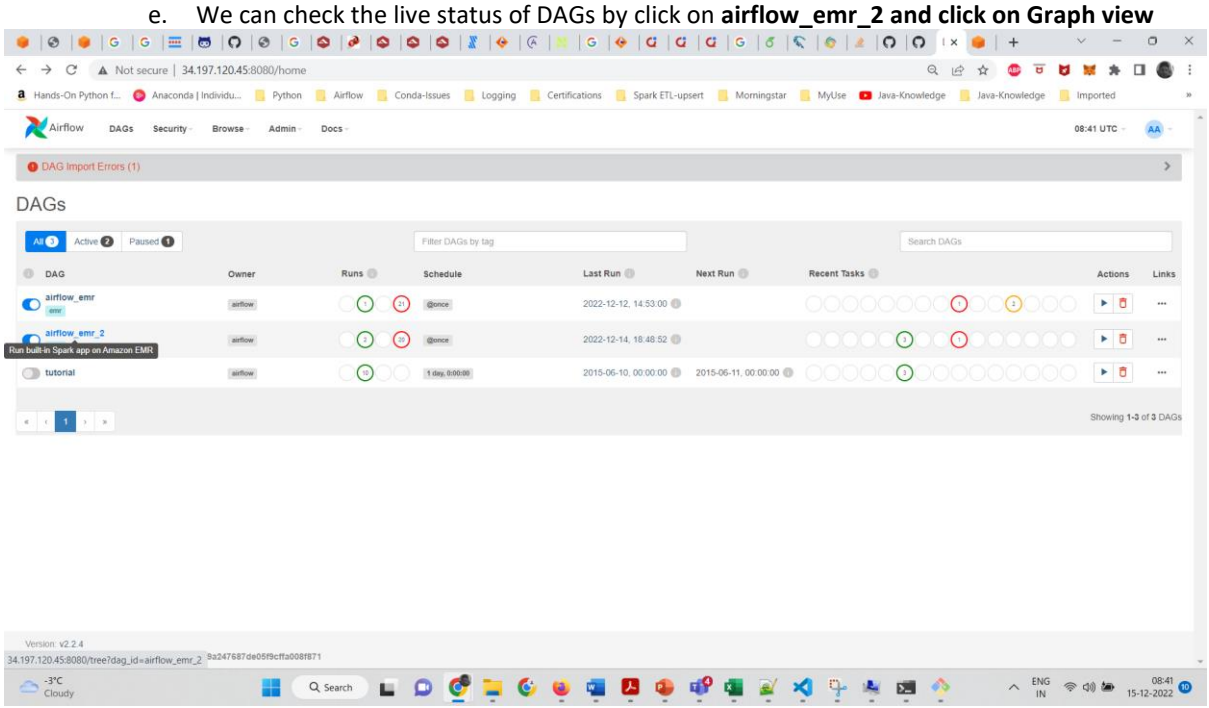


c. Go to link : http://34.197.120.45:8080/home and run the DAG as highlighted in yellow below:

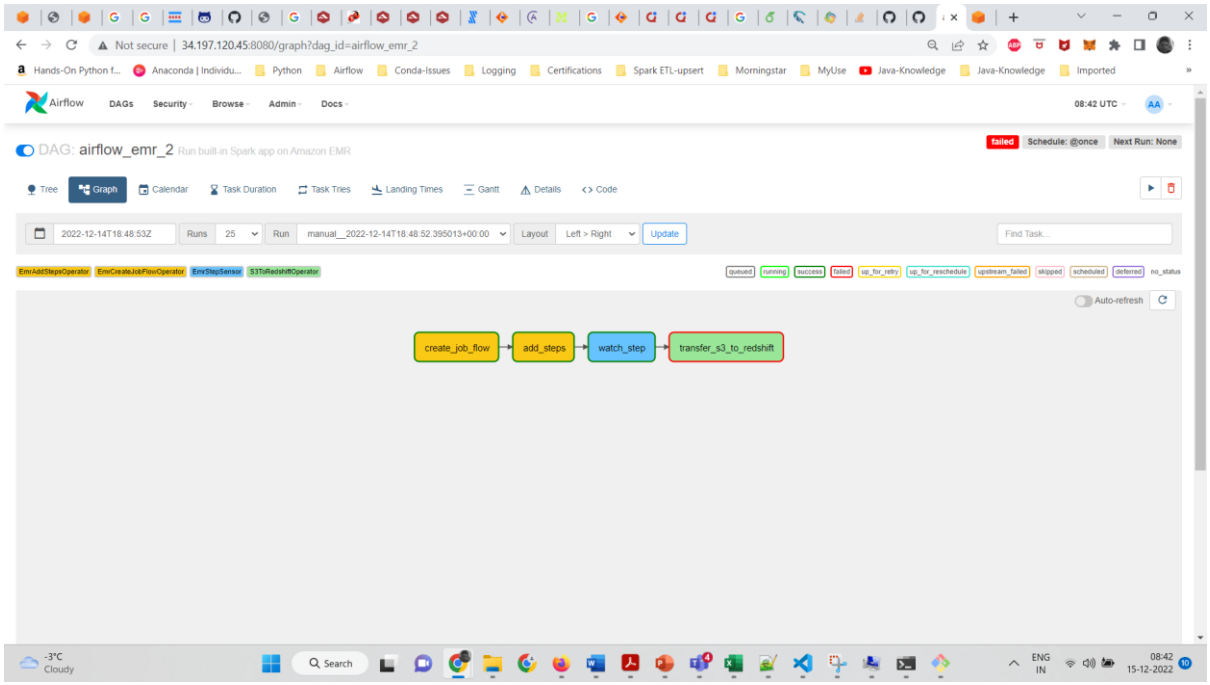


d. This would run the dag – **airflow_emr_2**

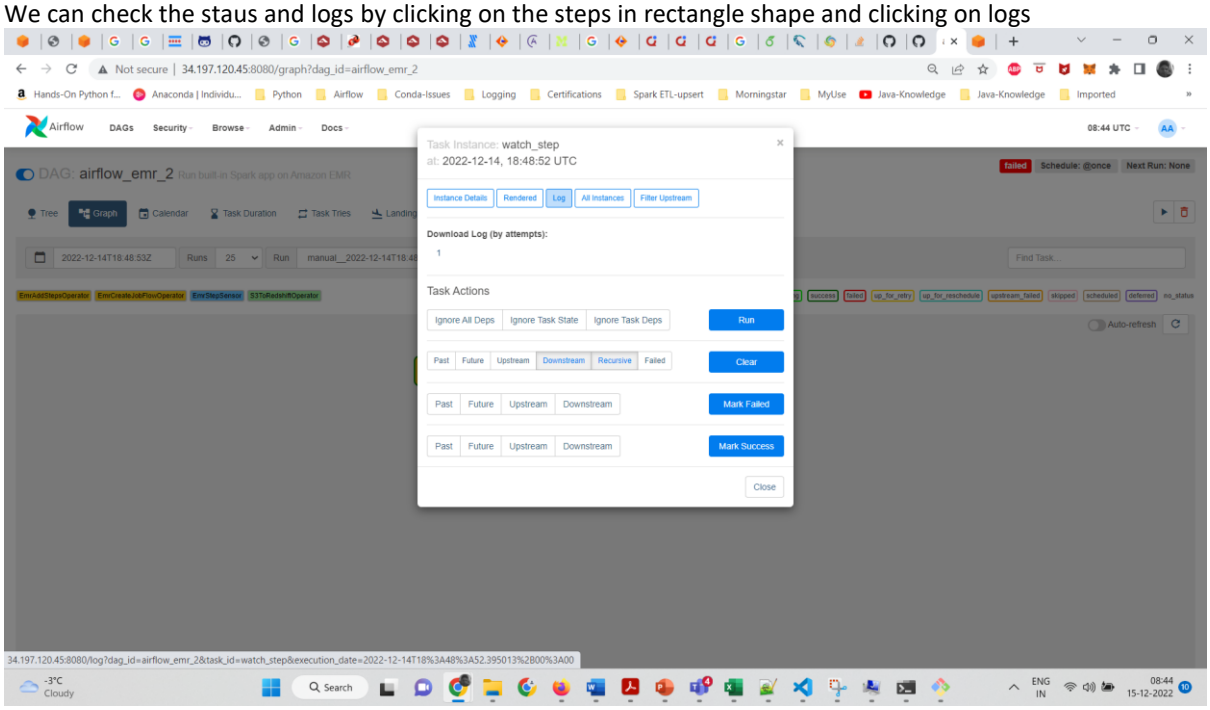
e. We can check the live status of DAGs by click on **airflow_emr_2 and click on Graph view**



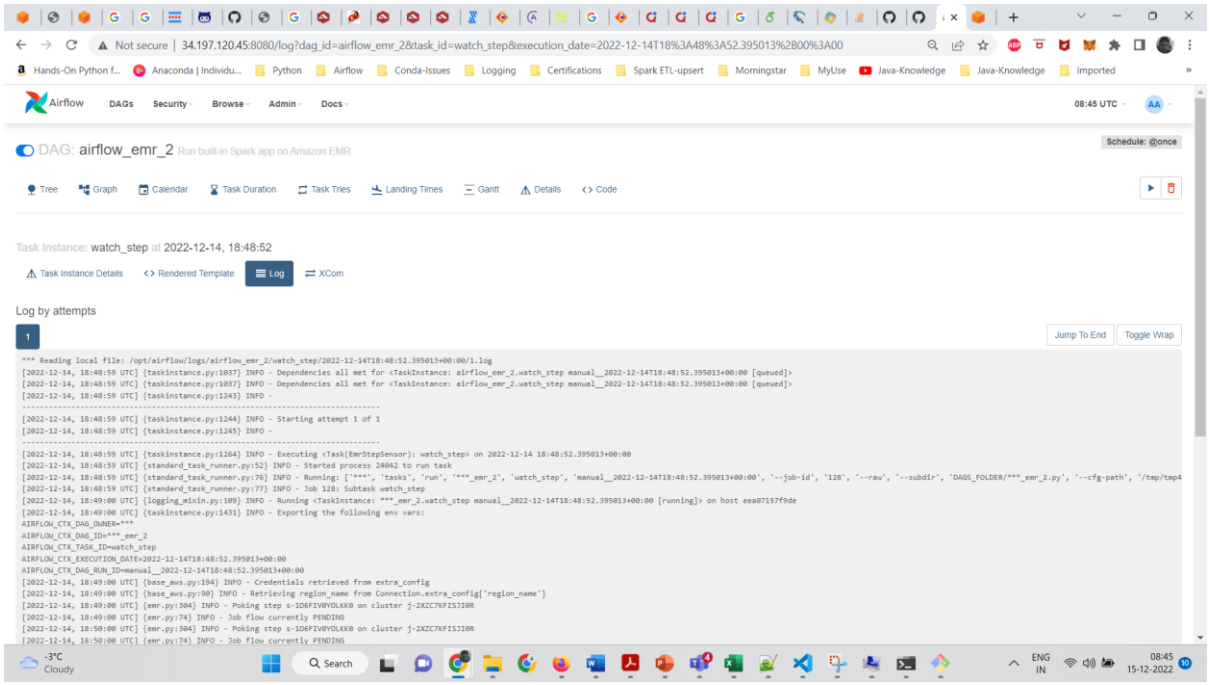
Graph view



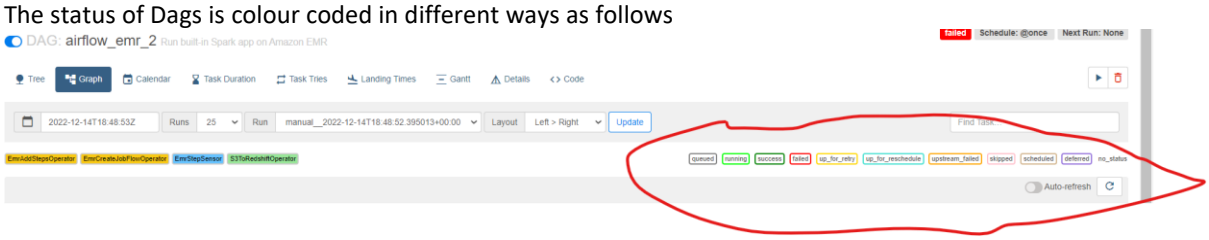
We can check the staus and logs by clicking on the steps in rectangle shape and clicking on logs



Log View

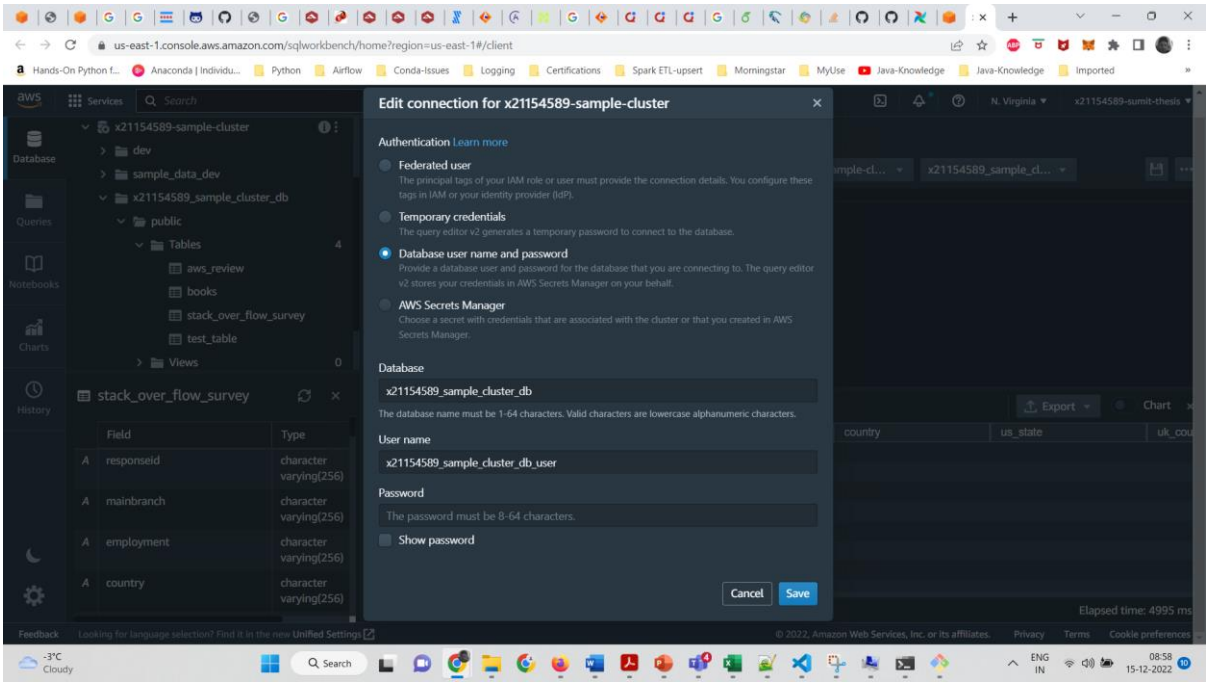


The status of Dags is colour coded in different ways as follows

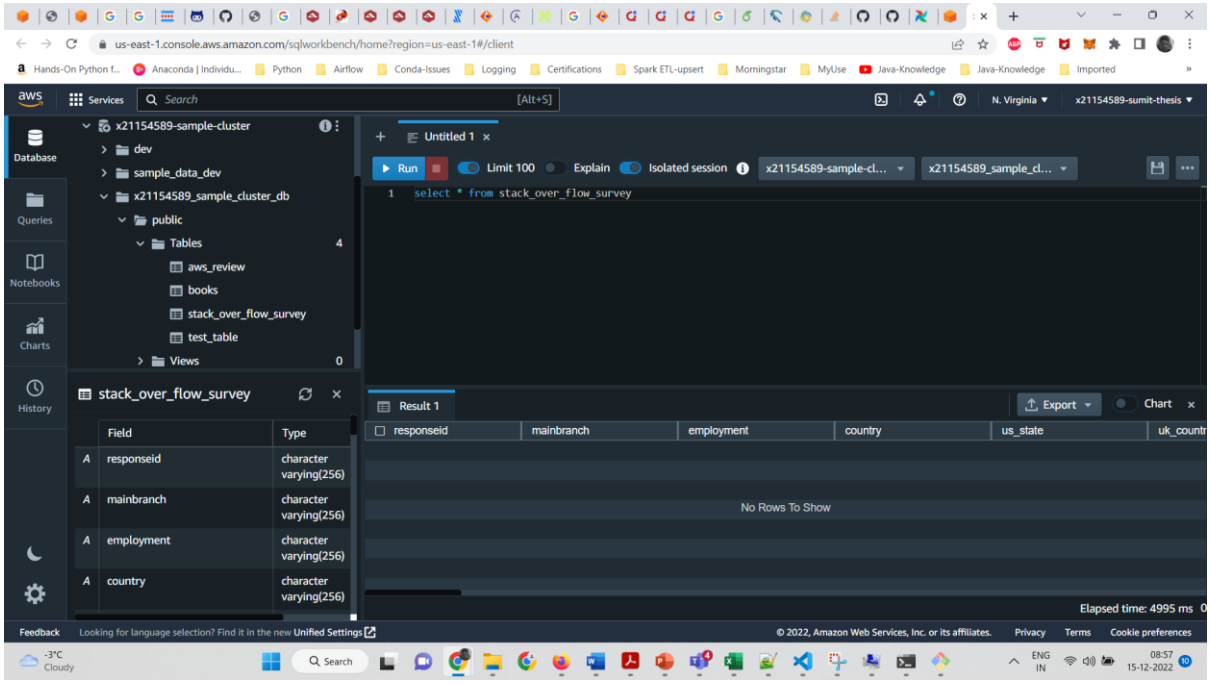


7 Data Analysis in Redshift:

1. Connect to the Database that is create using Airflow connection string

2. 

3. Data Analysis using SQL queris



8 Total Cost of ownership for the Setup infrastructure for an Year with the Hardware Configuration using AWS pricing calculator:

Successfully updated Amazon EMR estimate.

Contact your AWS representative:
https://aws.amazon.com/contact-us/

Export date: **12/14/2022** Language: **English**

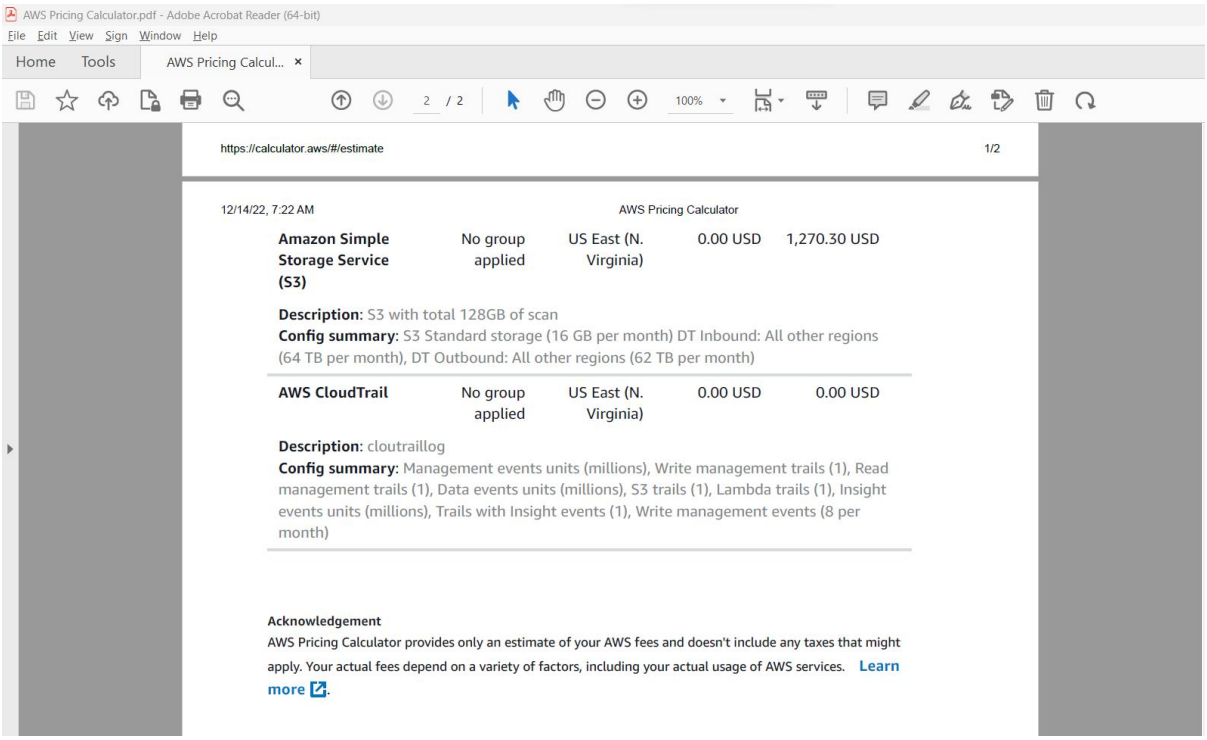
Estimate title: **My Estimate**

Estimate summary

Upfront cost	Monthly cost	Total 12 months cost
0.00 USD	**1,735.64 USD**	**20,827.72 USD** Includes upfront cost

Detailed Estimate

Name	Group	Region	Upfront cost	Monthly cost
Business support plan	No group applied	All regions	0.00 USD	157.79 USD
Description: AWS-SupportEstimate **Config summary**: Supports 24/7 phone, chat, and email access to Cloud Support Engineers for unlimited contacts, with and a response time of less than 1 hour.				
Amazon EC2	No group applied	US East (N. Virginia)	0.00 USD	77.45 USD
Description: Webserver-Airflow **Config summary**: Operating system (Linux), Quantity (1), Pricing strategy (On-Demand Instances), EBS Storage amount (8 GB)				
Amazon EMR	No group applied	US East (N. Virginia)	0.00 USD	44.10 USD
Description: EMR- startstoped by Airflow **Config summary**: Number of master EMR nodes (1), EC2 instance (m5.xlarge), Utilization (100 %Utilized/Month) Number of core EMR nodes (2), EC2 instance (m5.xlarge), Utilization (2 Hours/Day) Number of task EMR nodes (1), EC2 instance (m5.xlarge), Utilization (2 Hours/Day) Number of vCPUs per job run (2), Amount of memory per job run (GB) (8), Job runtime (120 minutes), Total ephemeral storage per job run (GB) (8)				
Amazon Redshift	No group applied	US East (N. Virginia)	0.00 USD	186.00 USD
Description: Reshift DW **Config summary**: Nodes (1), Instance type (dc2.large), Utilization (On-Demand only) (100 %Utilized/Month), Pricing strategy (OnDemand), Additional backup storage (145 GB), Data Transfer In To (8 GB)				



9 Conclusion

Using the above said tools , software and AWS services it is established that we can create an Open source ETL Framework for Big Data using AWS services and orchestrating it using Terraform with cost -effective Cloud Solution Architecture