**Wine Quality Prediction AWS Spark Application:**

**Pa2Winepred:** This project requires the creation of a Python application that uses the PySpark interface.

The application is running on an Amazon Web Services (AWS) Elastic MapReduce (EMR) cluster. The primary goal is to simultaneously train a machine learning model on EC2 instances to predict wine quality using publicly available data. The trained model is then used to predict the wine's quality. Docker is used to create a container image for the trained machine learning model, which simplifies the deployment process.

**Link for GitHub:**

https://github.com/Bharathwaj02/programming-Assignment-2

**Link for Docker:**

https://hub.docker.com/repository/docker/bg383/qulwinepred

**Steps for the Execution of Wine Quality Prediction AWS Spark Application:**

1. Create a Key-pair for the EMR Cluster: Go to EC2/Network/Key-pairs

Use the format of .pem and download the keypair

Created key pair as CS643key862.pem

2. Create an S3 bucket

We must create an S3 bucket in aws: cs643winequlpred2

3. Next, create an EMR cluster through the EMR console.

4. To create the spark in the AWS instance, use the EMR console:

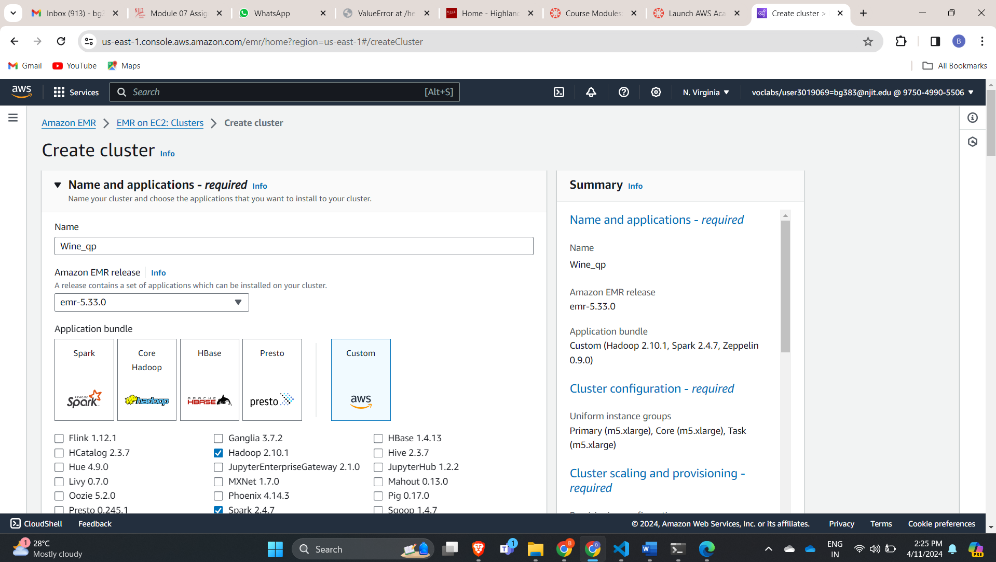
Create the spark cluster by using the EMR console, and create the 4 instances:

Name and application:

Name: Wine\_qp

Amazon EMR release: EMR-5.33.0

Application bundle: Hadoop 2.10.1, Spark 2.4.7, Zippeline 0.9.0, and Yarn



Note: It says Clone "wine\_qp\_clone" because I cloned the previous configuration instead of starting from scratch to save time.

Cluster Configuration:

A screenshot of a computer

Description automatically generated

Cluster Scaling and Provisioning:

A screenshot of a computer

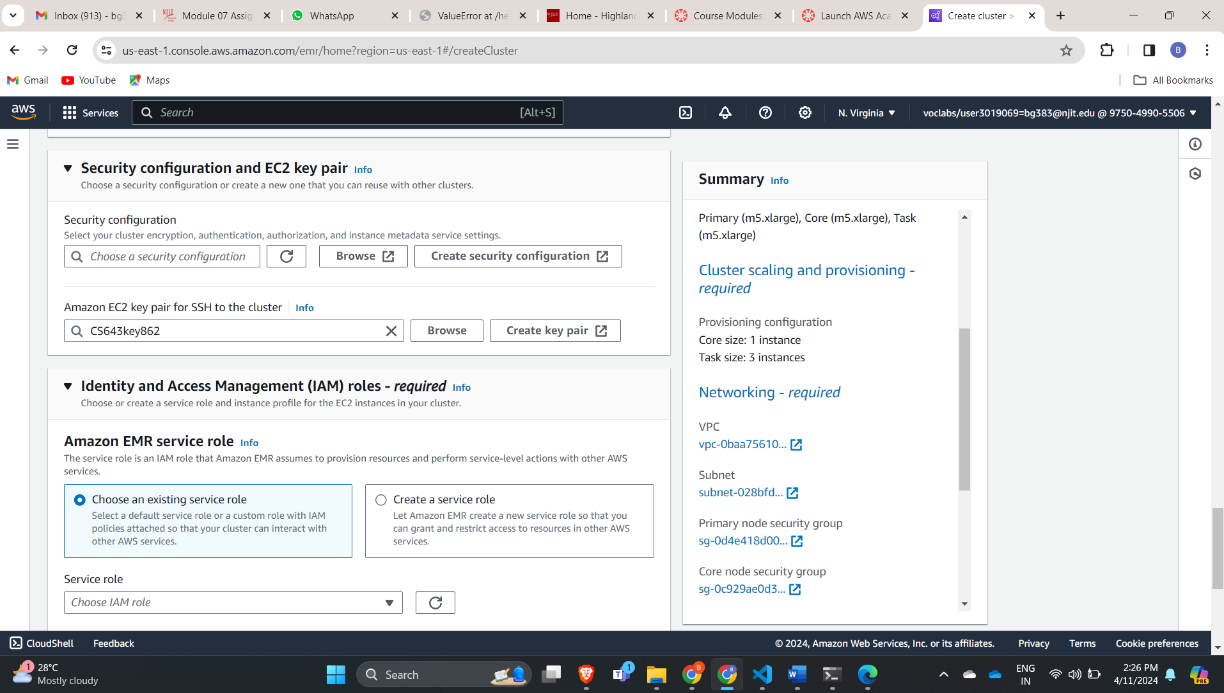
Description automatically generated

Networking & Cluster Termination:

A screenshot of a computer

Description automatically generated

Security Configuration and EC2 Key pair & Identity and access management(IAM) roles:



We can follow the above steps to create EMR cluster for the instancesA screenshot of a computer

Description automatically generated

5. Now we are training the ML model into spark cluster with ec2 instances in parallel:

1. Now the cluster will accept the tasks to run the ML model

Need to connect the Master instance in the Terminal:

ssh -i "CS643key862.pem" ec2-user@ec2-54-211-78-180.compute-1.amazonaws.com

and it is successfully logged in.

2. After the login of the Master instance then change the root by using

Sudo su



3. Submit the task by the command:

spark-submit s3://cs643winequlpred2/winequilityprediction.py

A screenshot of a computer program

Description automatically generated

A computer screen shot of a computer error

Description automatically generated

4. The trace status for the above tasks is then displayed. If the status is a success, a test .model is in the S3 bucket. s3://cs643winequlpred2

A screenshot of a computer

Description automatically generated

6. Now we are running ML model using the Docker:

1. Create a docker account and sign up.

2. After the successful login then download and set up the docker in your local system

3. Install the docker

4. Login to the docker in the power shell by the command

docker login

Pwd

5. After login you need to build the image:

docker build -t winequlpred .

A black screen with white lines

Description automatically generated

6. The push and pull into the docker hub repository:

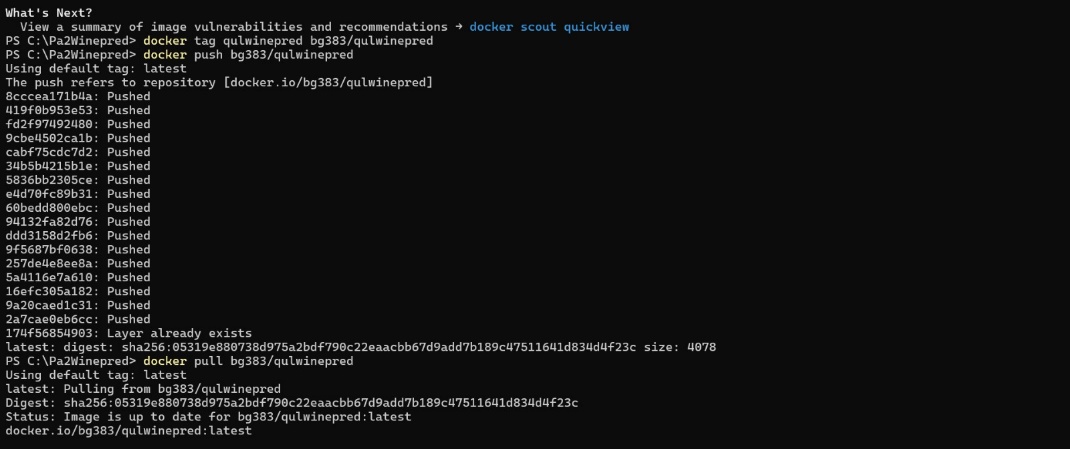
PUSH:

docker tag qulwinepred bg383/qulwinepred

docker push bg383/qulwinepred

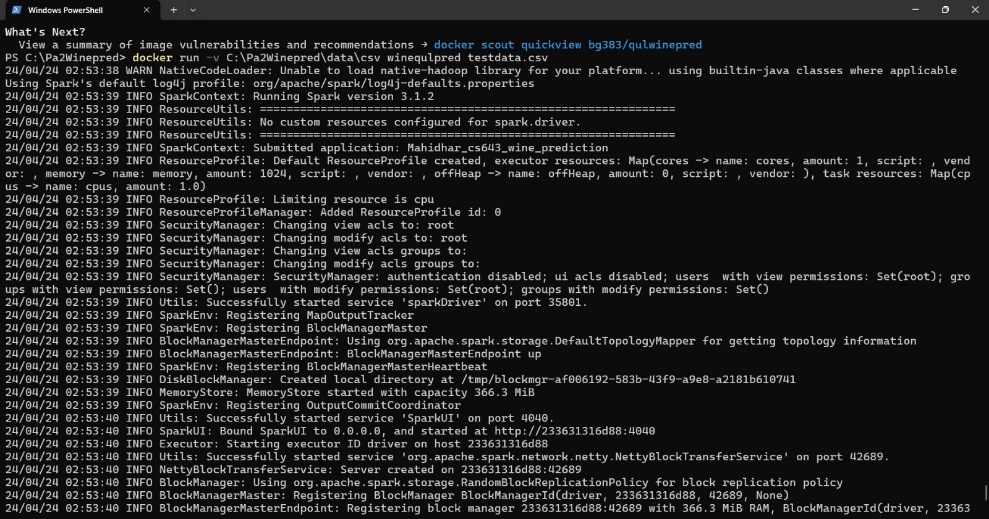
PULL:

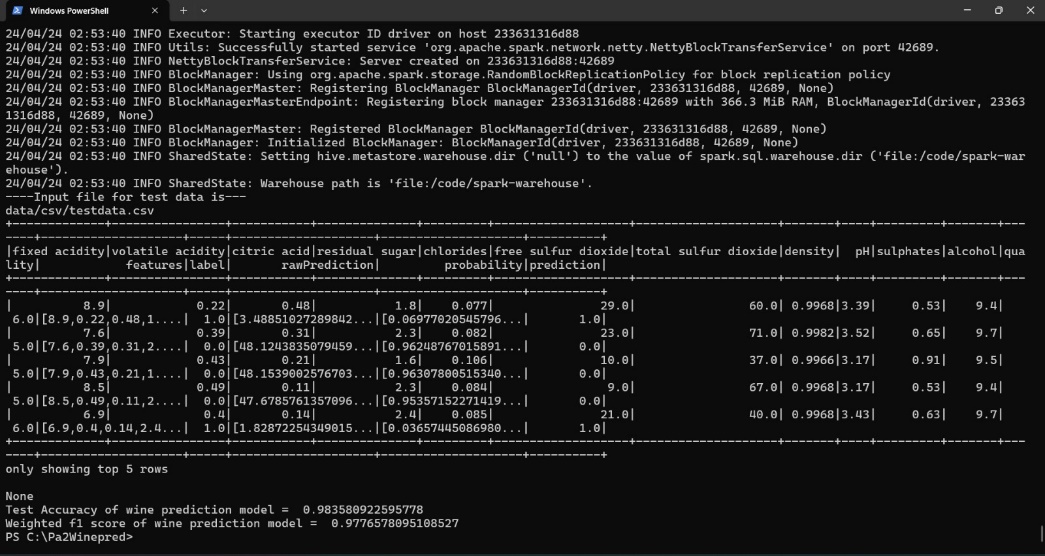
docker pull bg383/qulwinepred



7. Place your test data file in a designated folder known as "dir." Mount this directory with the Docker container, then run the container with the following command.

docker run -v C:\Pa2\data\csv winequlpred testdata.csv





**Conclusion**: As shown in the above image, got an accuracy of ~98% while predicting the wine quality.