Wine Quality Prediction AWS Spark Application:

Pa2Winepred: This project involves the development of a Python application utilizing the PySpark interface.

The application is deployed on an Amazon Web Services (AWS) Elastic MapReduce (EMR) cluster. The primary objective is to parallelly train a machine learning model on EC2 instances for predicting wine quality using publicly available data. Subsequently, the trained model is employed to predict the quality of wine. Docker is utilized to create a container image for the trained machine learning model, streamlining the deployment process.

Link for GitHub:

https://github.com/Mahidhartanniru/Pa2Winepred/

Link for Docker:

https://hub.docker.com/r/dt37824/winequlpred

Steps for the Execution for 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: pa2assmahi.pem

2.Create an S3 bucket

Created an S3 bucket in aws: pa2winebucket1

3.Then go to EMR console and create EMR cluster

4. Creating the spark in the AWS instance by using EMR console:

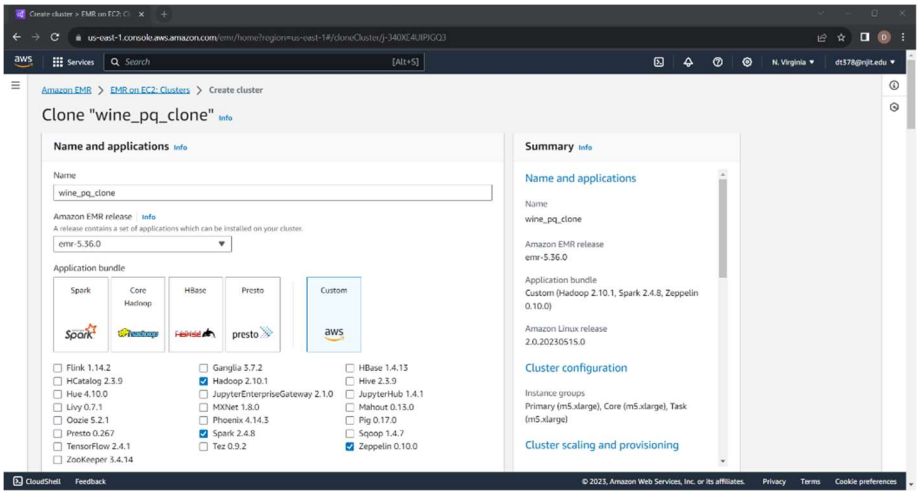
Creating the spark cluster by using the EMR console, and create the 4 instances:

Name and application:

Name:pa2winepqmahi

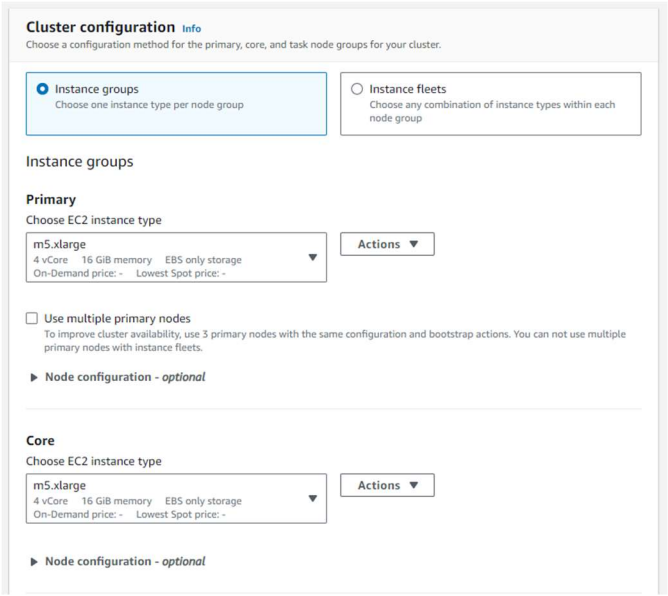
Amazon EMR release: emr-5.33.0

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

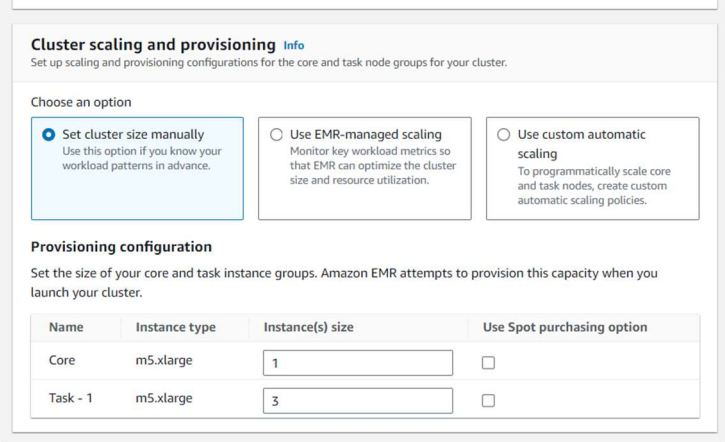


Note: Here it says Clone “wine_pq_clone” as I have cloned the previous configuration instead of creating from scratch to save time.

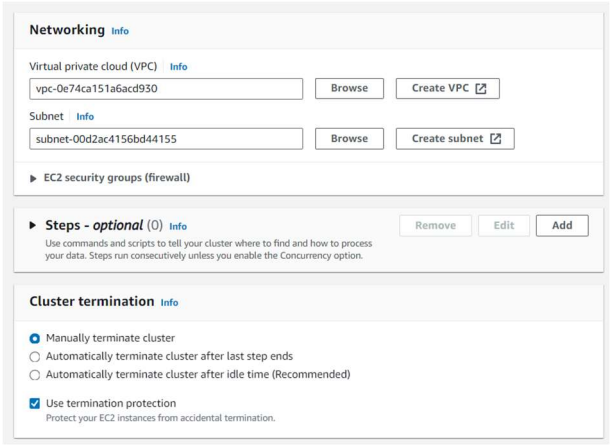
Cluster Configuration:



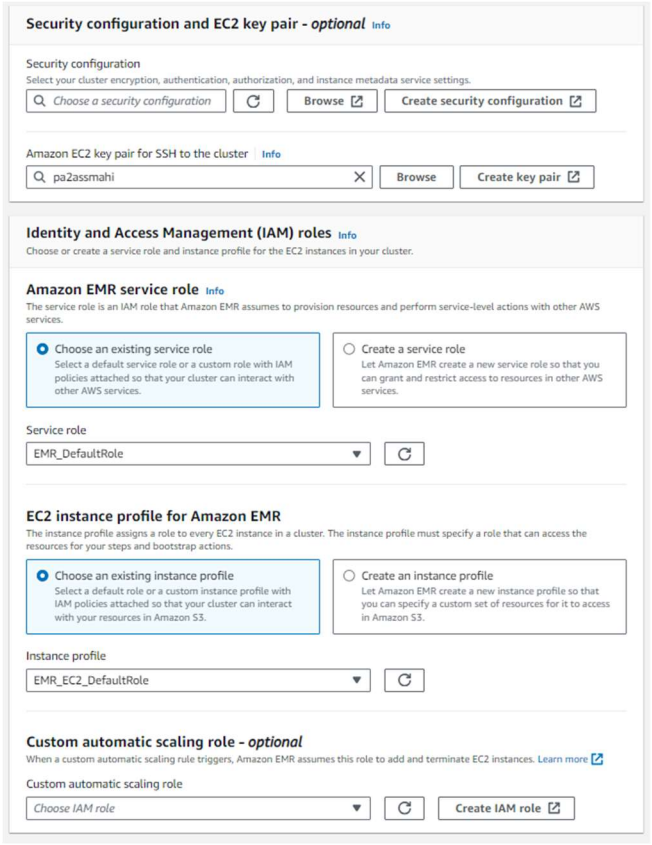
Cluster Scaling and provisioning:



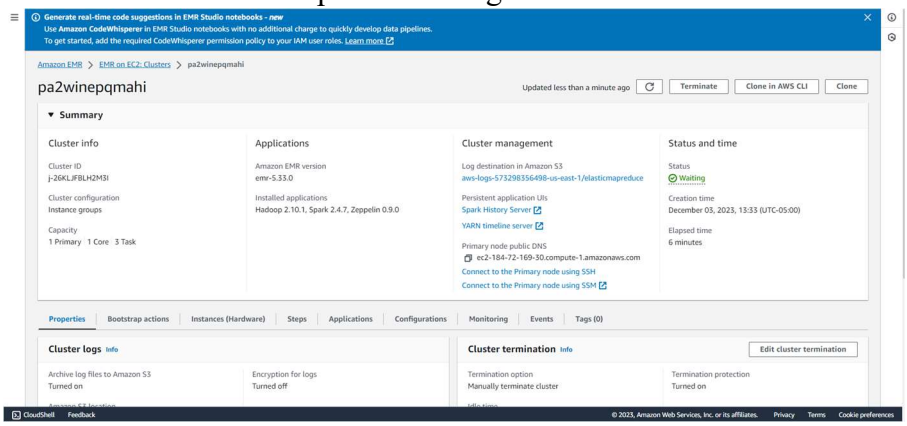
Networking & Cluster Termination:



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



We can follow above steps for creating EMR cluster for the instances



5.Now we are training 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 "pa2assmahi.pem" ec2-user@ec2-44-201-107-82.compute-1.amazonaws.com

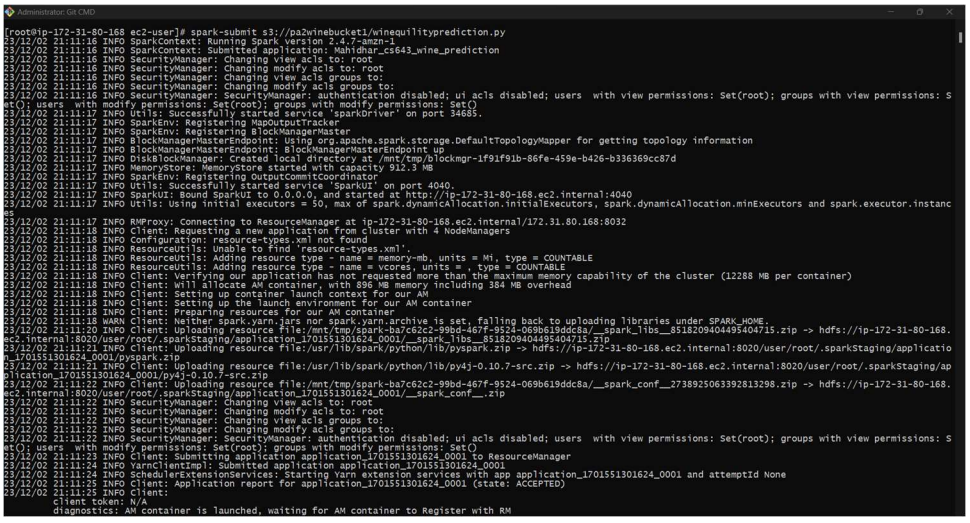
and it is successfully login.

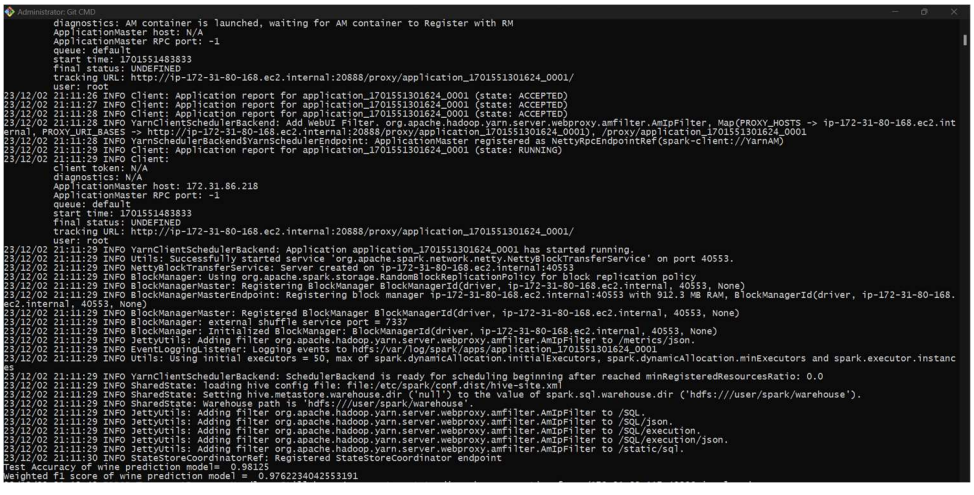
2.After the login of Master instance then change the root by using

Sudo su

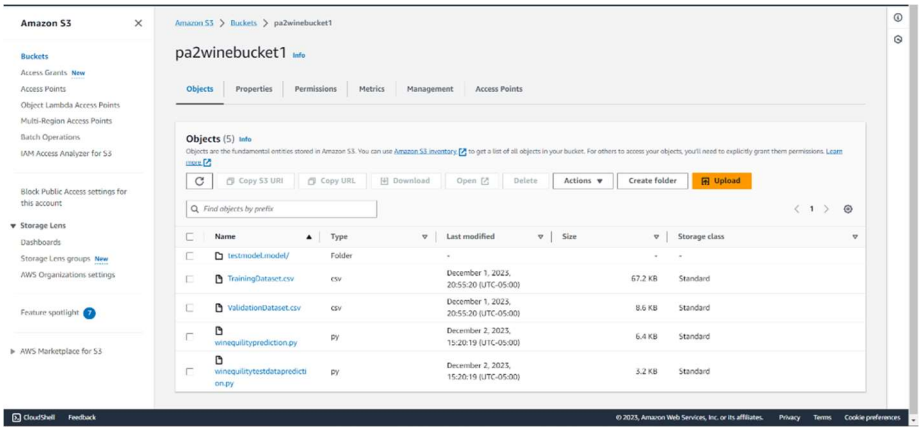
3. Submit the task by the command:

spark-submit s3://pa2winebucket1/ winequalityprediction.py





4.Then you can find the trace status for the above tasks, The status is succeed then there is a creation of test.model in the s3 bucket s3://pa2winebucket1



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

1.Create an docker account and sign up.

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

3.Install the docker

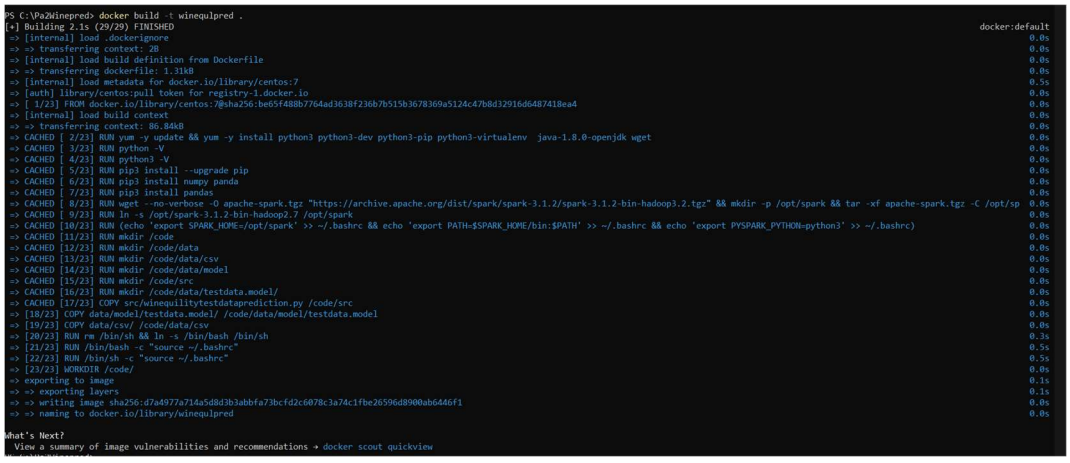
4. Login the docker in the power shell by the command

docker login

Pwd

5.After login you need to build the image:

docker build -t winequlpred .



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

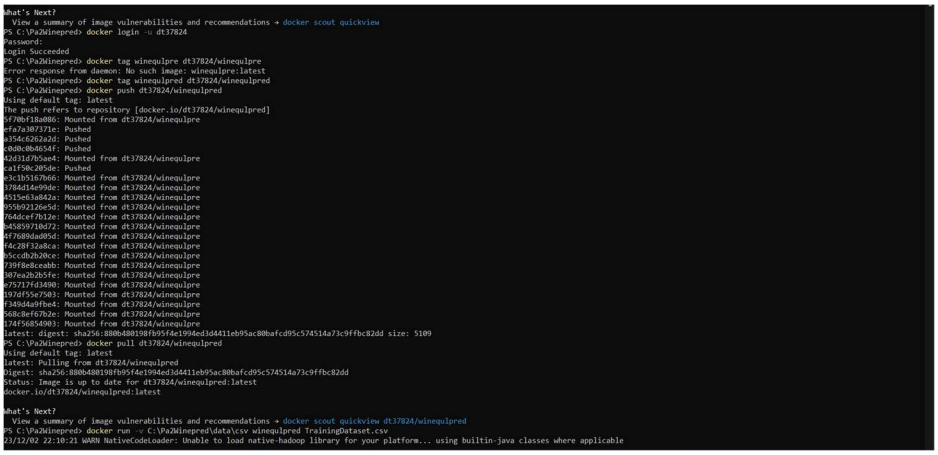
PUSH:

docker tag winequlpred dt37824/winequlpred

docker push dt37824/winequlpred

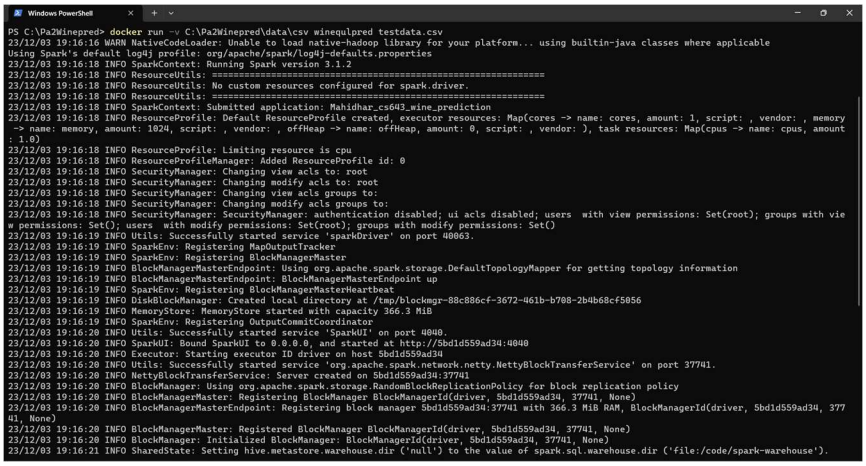
PULL:

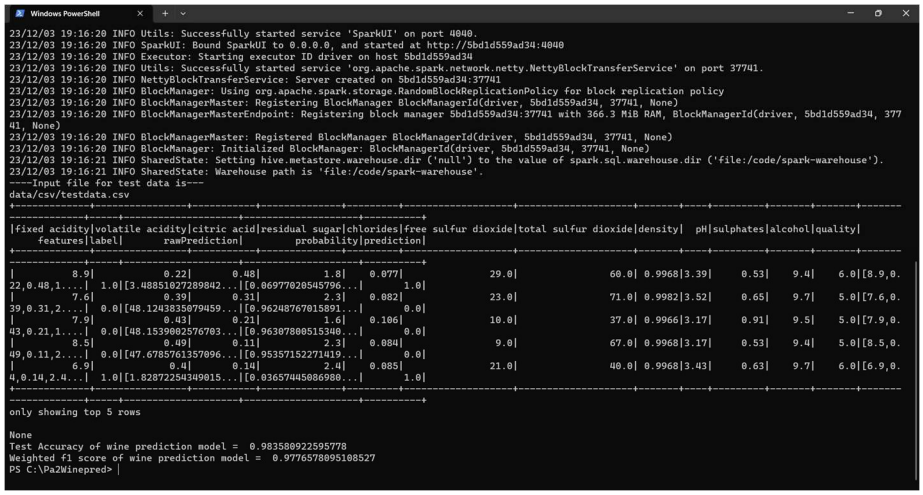
docker pull dt37824/winequlpred



7. Store your test data file in a designated folder, referred to as "dir." Mount this directory with the Docker container, and execute the container using the following command.

docker run -v C:\Pa2Winepred\data\csv winequlpred testdata.csv





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