Programming Assignment-2

AWS Spark Quality Prediction:

This project aims to create a Python application using the PySpark framework. The application runs on an Amazon Web Services (AWS) Elastic MapReduce (EMR) cluster. Its main goal is to train a machine learning model in parallel on EC2 instances to predict wine quality using publicly accessible data. After training, the model is used to make predictions about wine quality. To simplify deployment, we use Docker to build a container image for the trained model.

Link for GitHub: https://github.com/hg332Hru/PA2-Assignment

Link for Docker: <https://hub.docker.com/repository/docker/hg332/predwine>

Execution for AWS Spark Quality Prediction :

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: hrudaykey1.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

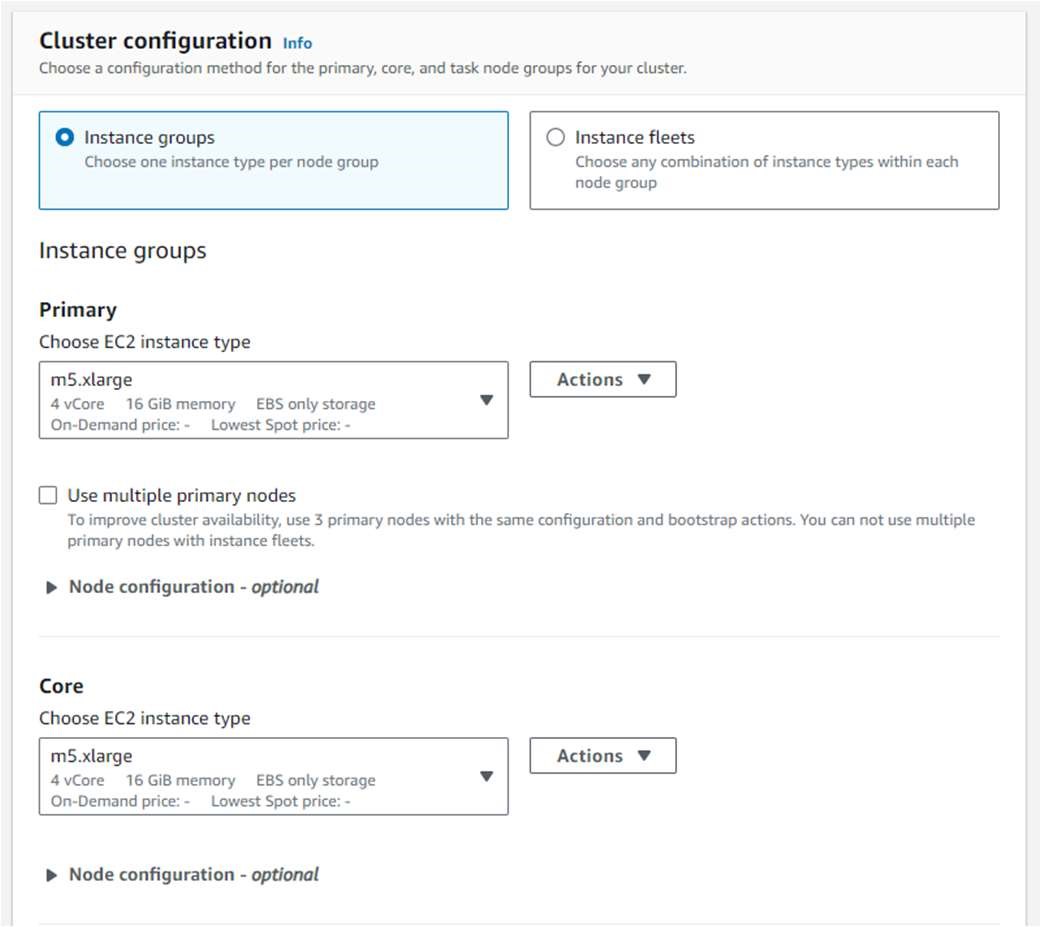
Amazon EMR release: emr-5.33.0

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

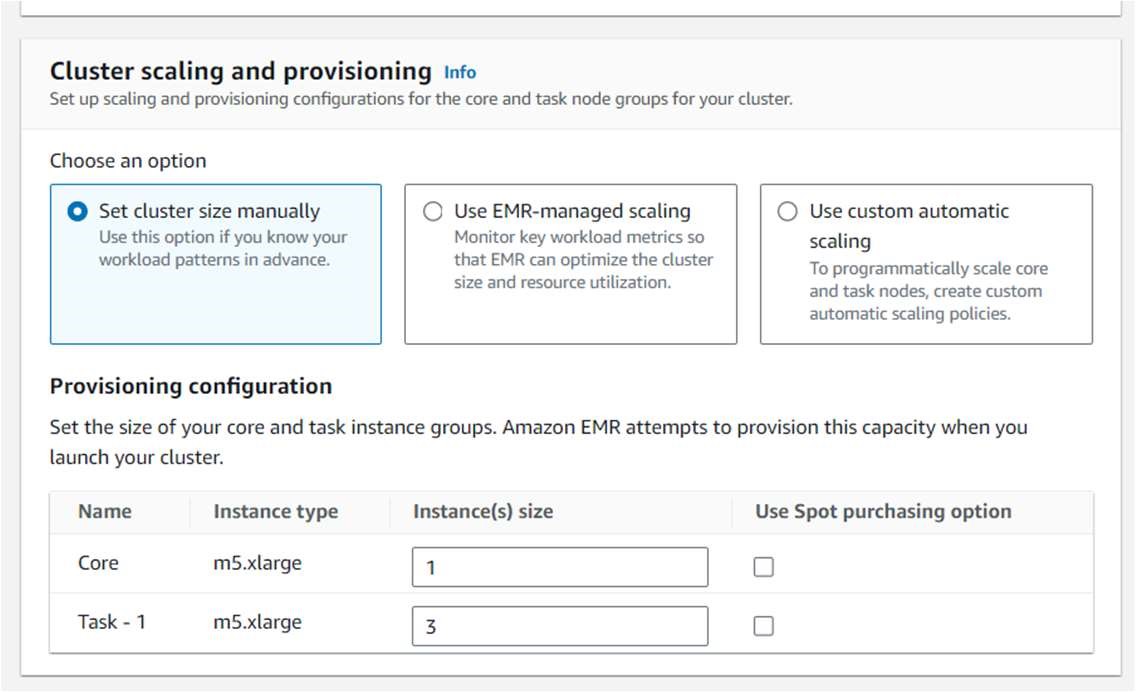
A screenshot of a computer

Description automatically generated

Cluster Configuration:



Cluster Scaling and provisioning:



Networking & Cluster Termination:

A screenshot of a computer error message

Description automatically generated

Security Configuration and EC2 Key pair:

A screenshot of a computer program

Description automatically generated

Identity and access management(IAM) roles:

A screenshot of a computer

Description automatically generated

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

Starting:

A screenshot of a computer

Description automatically generated

Waiting:

A screenshot of a computer

Description automatically generated

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 "hrudaykey1.pem" ec2-user@ec2-18-206-219-210.compute-1.amazonaws.com

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

Sudo su

A screenshot of a computer

Description automatically generated

3. Submit the task by the command:

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

A screenshot of a computer code

Description automatically generated

A screenshot of a computer screen

Description automatically generated

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://hrudaybucket.

A screenshot of a computer

Description automatically generated

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

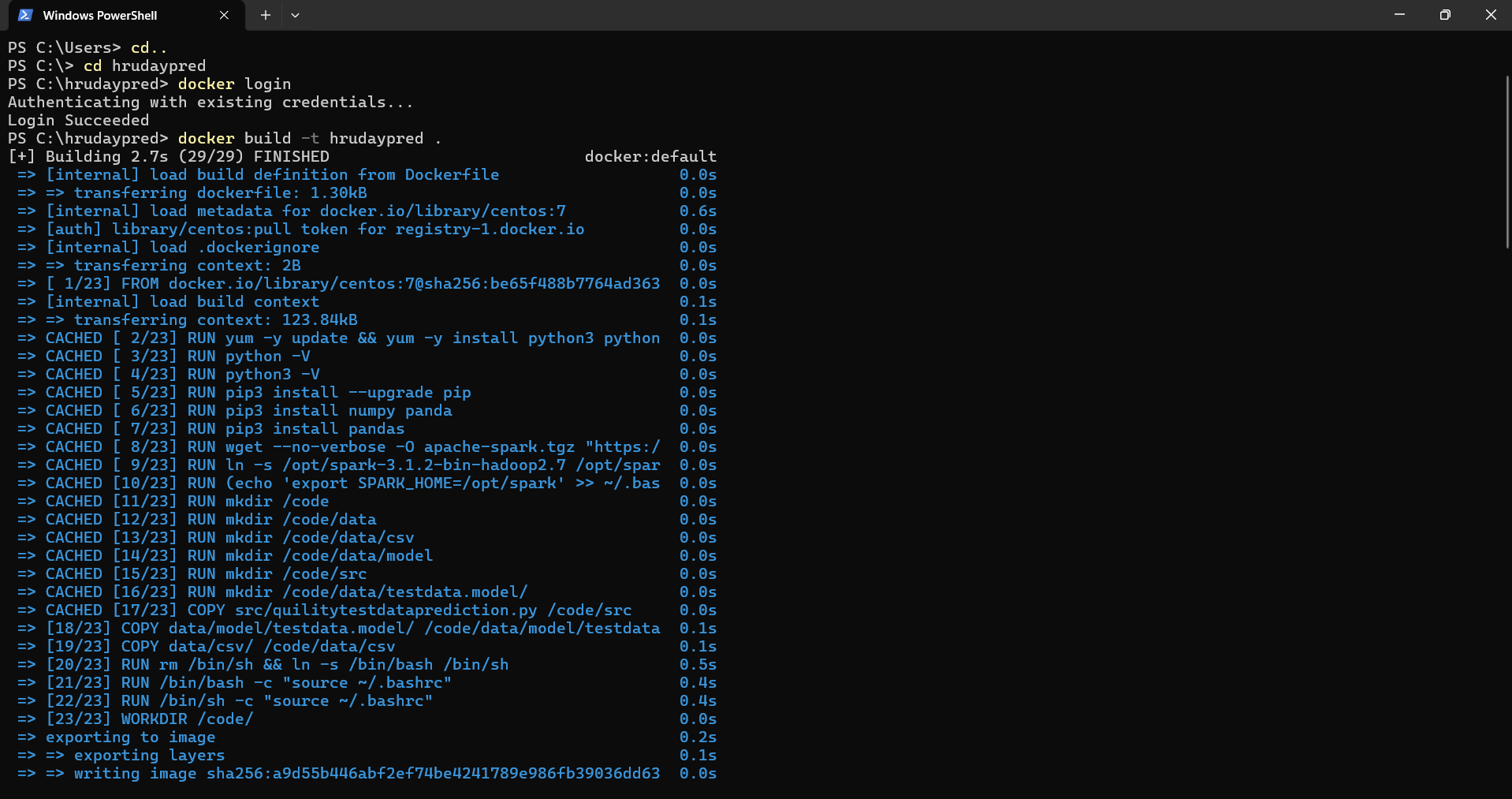
3.Install the docker

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 hrudaypred .



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

PUSH:

docker tag hrudaypred hg332/predwine

docker push hg332/predwine

PULL:

docker pull hg332/predwine

A screenshot of a computer

Description automatically generated

1. 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:\hrudaypred\data\csv hrudaypred testdata.csv

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

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