Name: Abishek Vellineni

Email Id: av739@njit.edu

Module 07 Assignment 01: Programming Assignment 2

Wine Quality Prediction AWS Spark Application

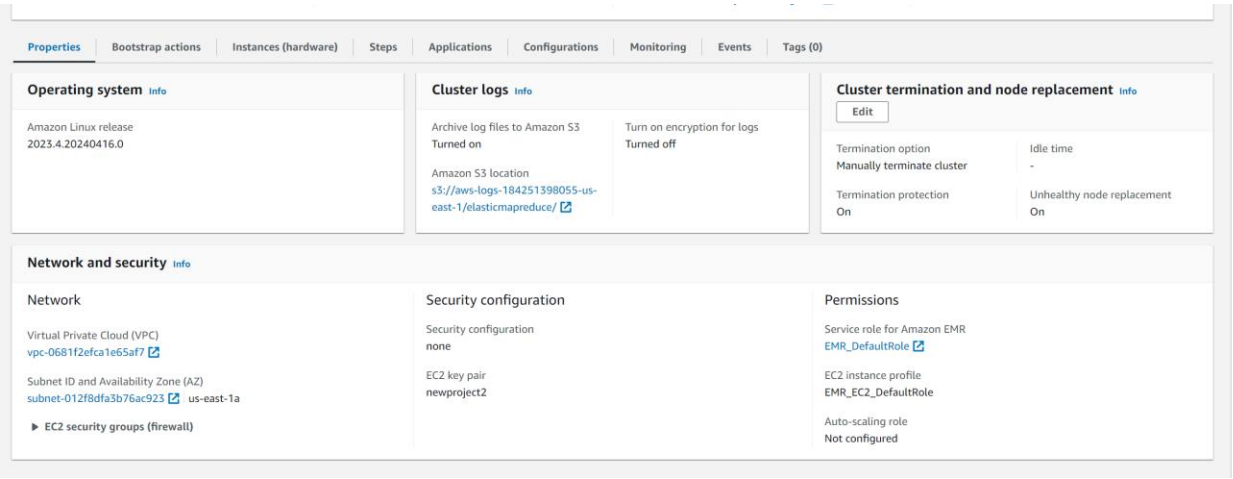
The application is deployed on an Amazon Web Services (AWS) Elastic MapReduce (EMR) cluster. The primary objective is to parallelly train a machine learning model on EC2 instances for predicting wine quality using publicly available data. Subsequently, the trained model is employed to predict the quality 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/Abishek183/wine_prediction

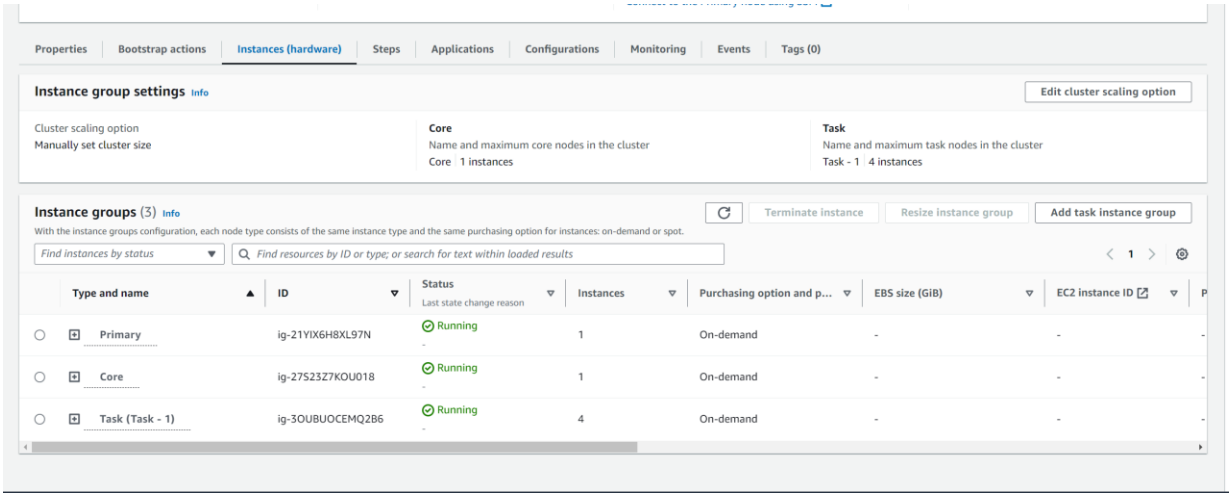
Link for Docker - https://hub.docker.com/repository/docker/abishek183/wine_predict/tags

Cluster Creation AWS SPARK:

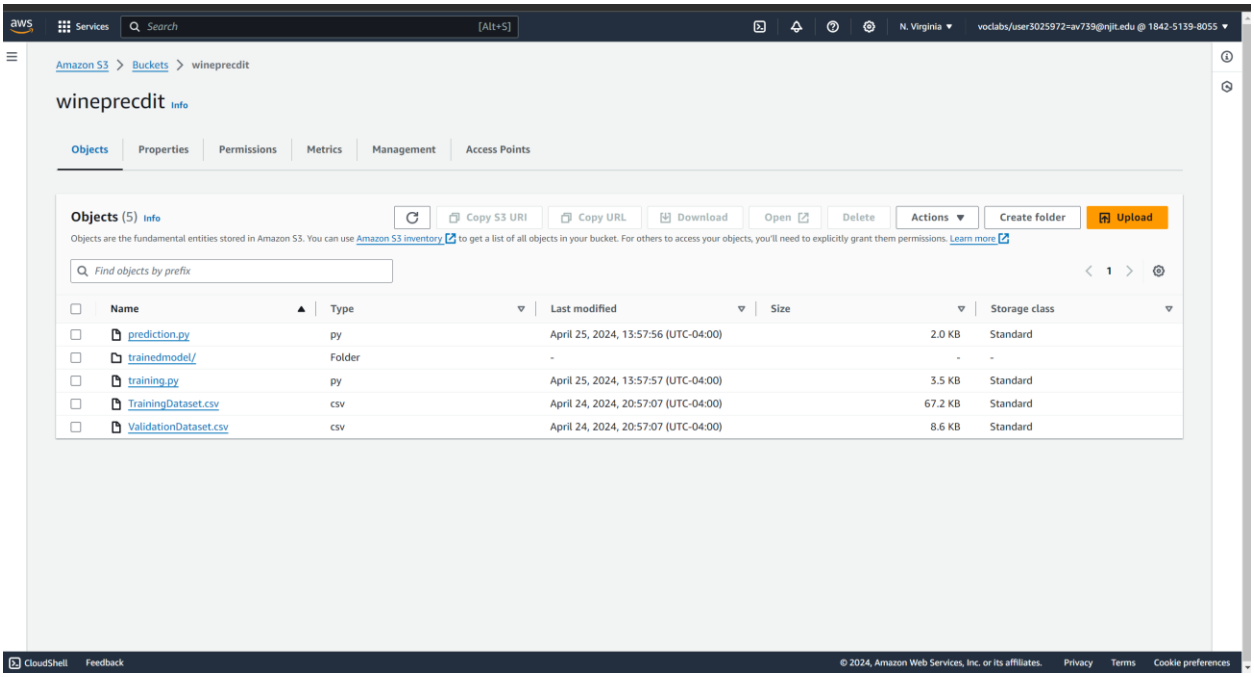
1. Navigate to the EMR console and then click on create new instance.
2. Provide the name for your cluster.
3. In cluster Termination, change to manual from automatic.
4. Provide the key_pair which is a .pem file in security configuration



5. In instance creation, provide 1 for core and 4 for tasks as we need to run on 4 EC2 instances.
6. Then select default roles for the IAM roles.



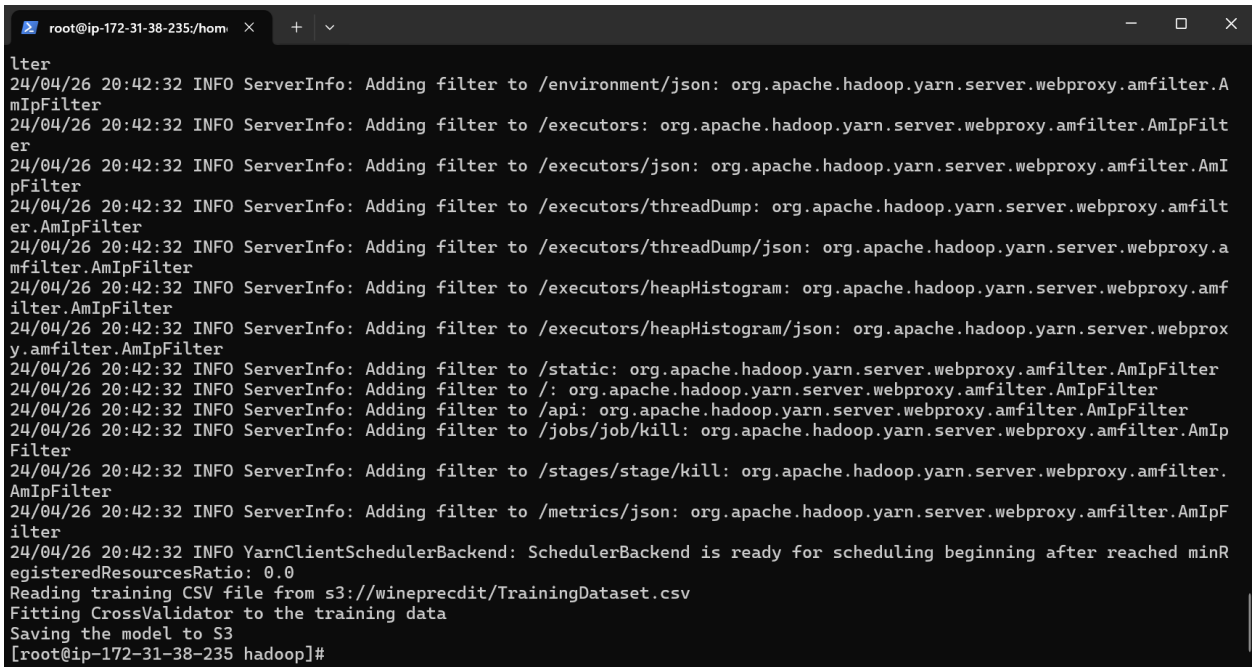
7. Create a S3 queue to upload the python and csv files.
   s3://wineprecdit



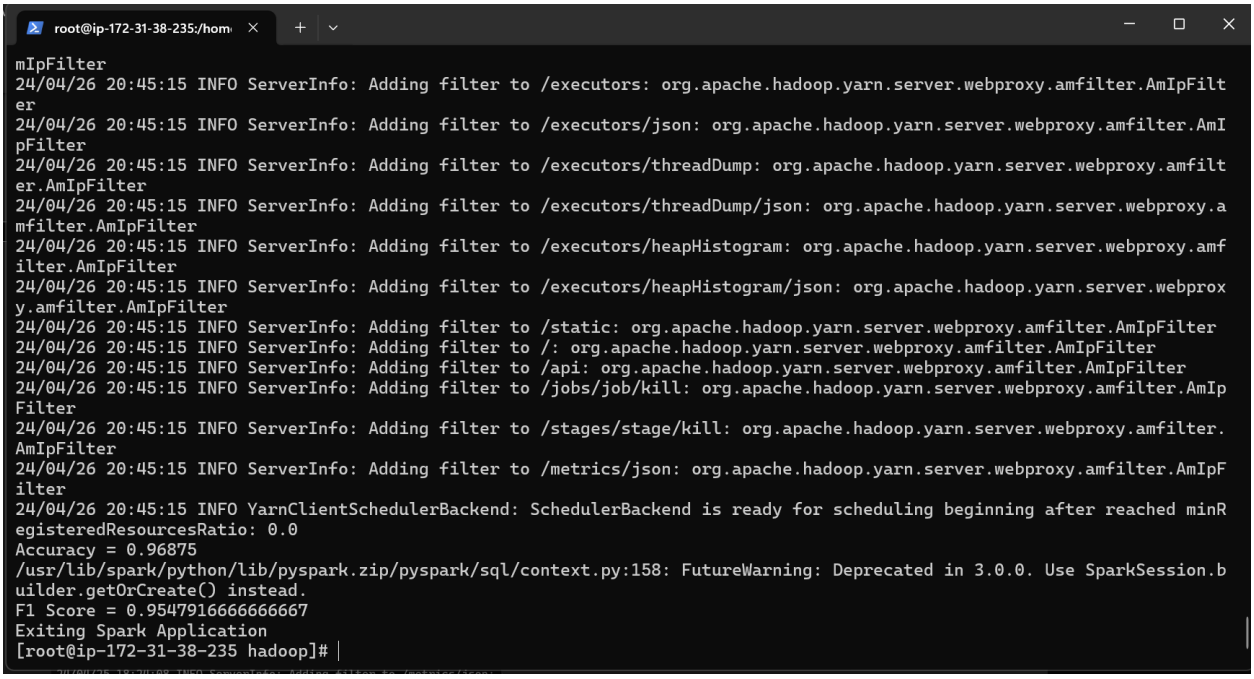
8. Open terminal and use the below command to connect to the cluster.
   ssh -i ~/newproject2.pem hadoop@ec2-52-201-250-228.compute-1.amazonaws.com

Execution without Docker

- Run “sudo su” command to change user.
- Install numpy by using ‘pip install numpy --user’
- Then run “spark-submit s3://wineprecdit/training.py”. It runs the file from S3 bucket and creates a ML model



- Then run “spark-submit s3://wineprecdit/prediction.py s3://wineprecdit/ValidationDataset.csv”. It uses the model created and validates the data from the csv file and provides the result.
- We can infer that it provides an accuracy of 95.479% from the below image.

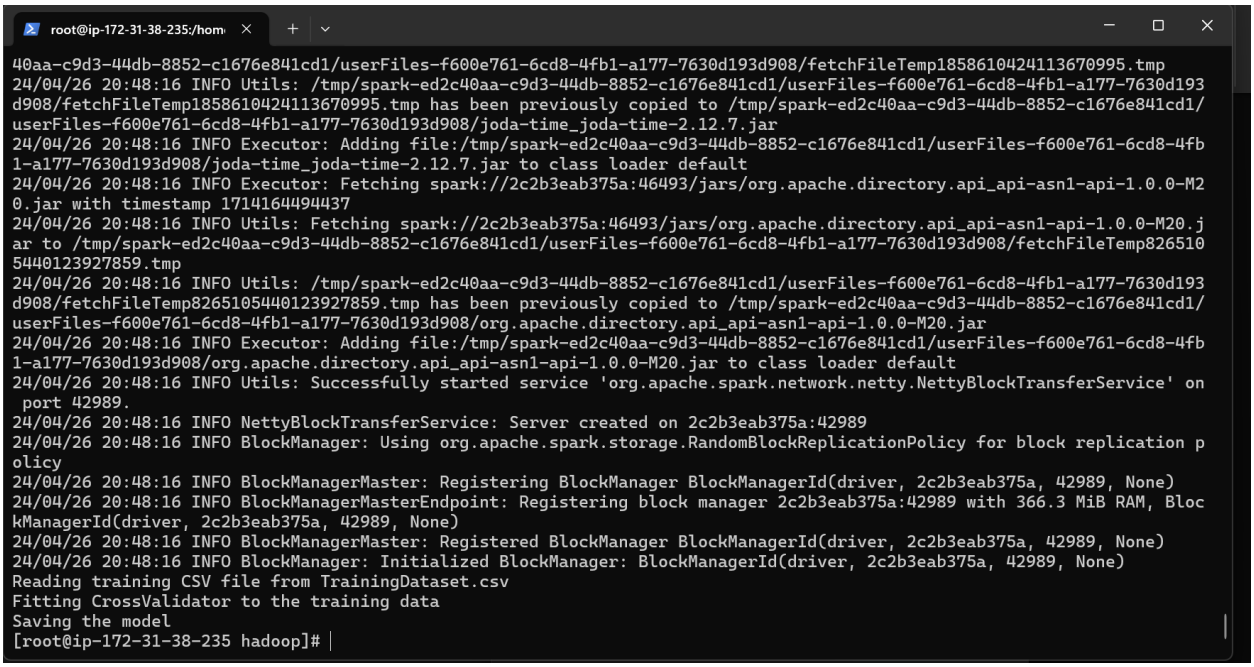


Execution with Docker

- Run the below commands to start the docker in the EC2.
  - sudo systemctl start docker
  - sudo systemctl enable docker

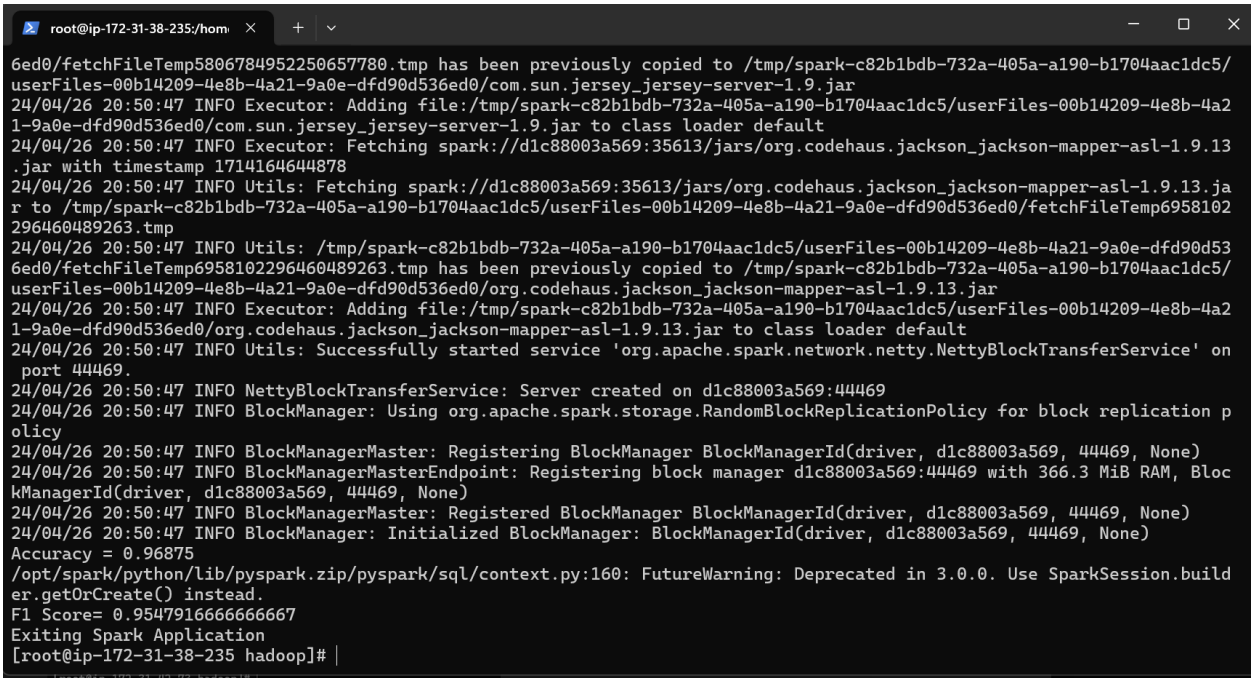
- Get the image from docker repo using the below command.
  - sudo docker pull abishek183/wine_predict:train
  - sudo docker pull abishek183/wine_predict:predict

- Run the train tag image to create a ML model using the below command.
  - sudo docker run -v /home/ec2-user/:/job abishek183/wine_predict:train



- Run the image with predict tag to get the accuracy.
    - sudo docker run -v /home/ec2-user/:/job abishek183/wine_predict:predict ValidationDataset.csv

- we can infer from the below image the accuracy is 95%.



Docker Image: