**Readme File**

**Github Link:** https://github.com/ap2876-Ashutosh/cs643-PG2.git

**Docker Hub Link:** https://hub.docker.com/repository/docker/ap2876/wine-prediction/general

Steps to create cluster and execute code:

1. Create EMR Cluster:

* Go to amazon console and select EMR option.
* Select latest versions for the cluster.

A screenshot of a computer

Description automatically generated

* For the primary and core instances select m.5.xlarge as instance type.

A screenshot of a computer

Description automatically generated

* Under the add the instance group add 4 task nodes (slave nodes).

A screenshot of a computer

Description automatically generated

* Select the cluster termination as manual so as to keep the cluster alive even when there is no activity.

A screenshot of a computer

Description automatically generated

* Select the EC2 key pair for connection using SSH. Select the key-pair created, else create one for connecting to the EMR cluster.

A screenshot of a computer

Description automatically generated

* Choose the default roles for IAM roles, EC2 instance profile and automatic scaling role.

A screenshot of a computer

Description automatically generated

* After selecting create cluster wait for the cluster status to be *waiting* indicating the cluster is waiting to be trained.

A screenshot of a computer

Description automatically generated

1. Connecting to the Primary node using PUTTY:

* Enter the host-name in the putty configuration

A computer screen shot of a computer

Description automatically generated

* To provide access to the SSH port (22) edit the inbound rules for the primary node.

A screenshot of a computer

Description automatically generated

* Under Connection 🡪 Auth 🡪 Credentials provide the private key configuration provided during the cluster creation process.

A screenshot of a computer

Description automatically generated

* A screenshot of a computer

  Description automatically generatedOnce connected the cluster looks like this:

1. Uploading Files to S3 bucket:

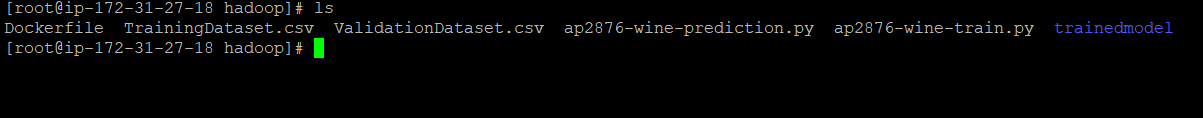
* Create a bucket ap2876-wine-predictior with all the settings to default.
* Upload all the required files in the S3 bucket. This includes the python files and docker file.

A screenshot of a computer

Description automatically generated

1. Training the Model:

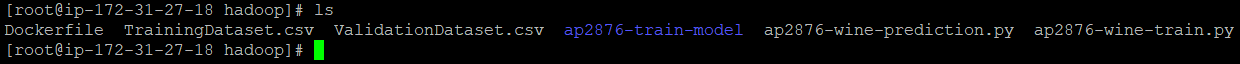
* Before training the model, we need to sync all the files in the instance. The following command is used to sync all the files in the Ec2-instance:
  + aws s3 sync s3://ap2876-wine-predictor/ .
* To verify the list of files in the instance use the command *‘ls’* :



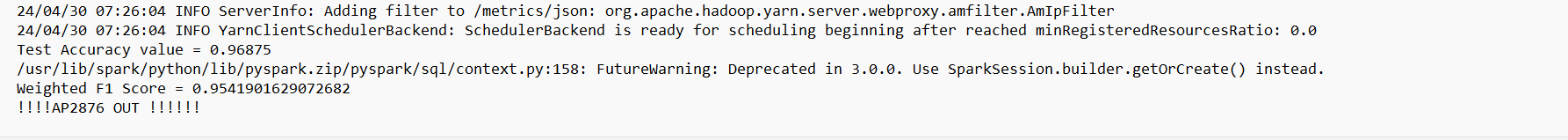
* To avoid any issues, switch to root user and install the dependencies like numpy using the command.
  + sudo su
  + pip install numpy
* Run the training script using the following command:
  + spark-submit ap2876-wine-train.py

1. Running Model for prediction in the EC2-Instance:

* Since the model is already trained in Master EC2-instance we can use the prediction model in a similar instance.
* Given the file system may be lacking the Hadoop core, we may need to copy the file in the Hadoop file system. We can use the following commands.
  + hadoop fs -copyFromLocal TrainingDataset.csv hdfs://ip-172-31-27-18.ec2.internal:8020/user/root/
  + hadoop fs -copyFromLocal ValidationDataset.csv hdfs://ip-172-31-27-18.ec2.internal:8020/user/root/
* We also might need to copy the model from HDFS to root directory using the following command:
  + hadoop fs -ls hdfs://ip-172-31-27-18.ec2.internal:8020/user/root/
  + hadoop fs -get hdfs://ip-172-31-27-18.ec2.internal:8020/user/root/ap2876-train-model .
* Check if the model is imported in the root directory using command ‘ls’:



* Run the model for prediction using the following command:
  + spark-submit ap2876-wine-prediction.py
* Check the test score and the weighted-F1 score:



1. Running the prediction model using Docker:

* Login into the docker using the following command:
  + docker login

*Note: The docker logins are stored in the /root/.docker/config.json so we only need to input the command once.*

A screen shot of a computer

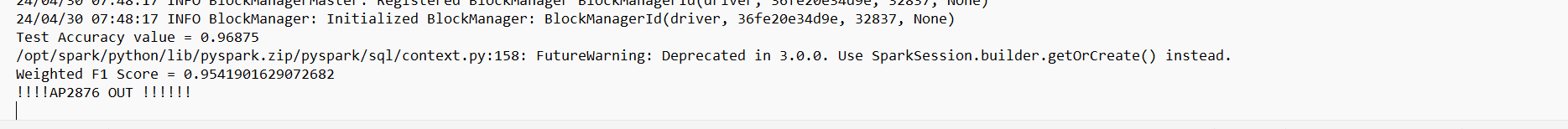
Description automatically generated

* Using the Docker file create the docker image:
  + docker build -t ap2876/wine-prediction .

A computer screen with blue text

Description automatically generated

* Run the image using the following command:
  + docker run ap2876/wine-prediction



* Push the image in the docker hub using the following command:
  + docker push ap2876/wine-prediction

A screen shot of a computer

Description automatically generated