CS634 Programming Assignment 2

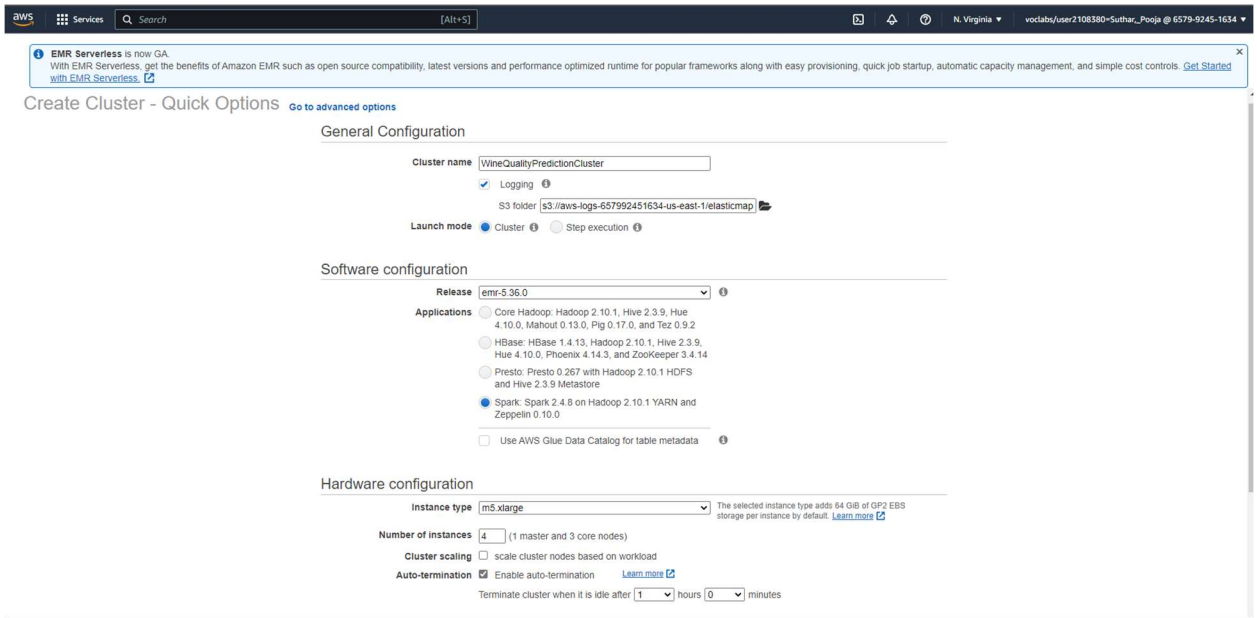
This assignment is about to develop parallel machine learning (ML) application in Amazon AWS cloud platform and use Apache Spark to train ML model in parallel on 4 EC2 instances and save the model.

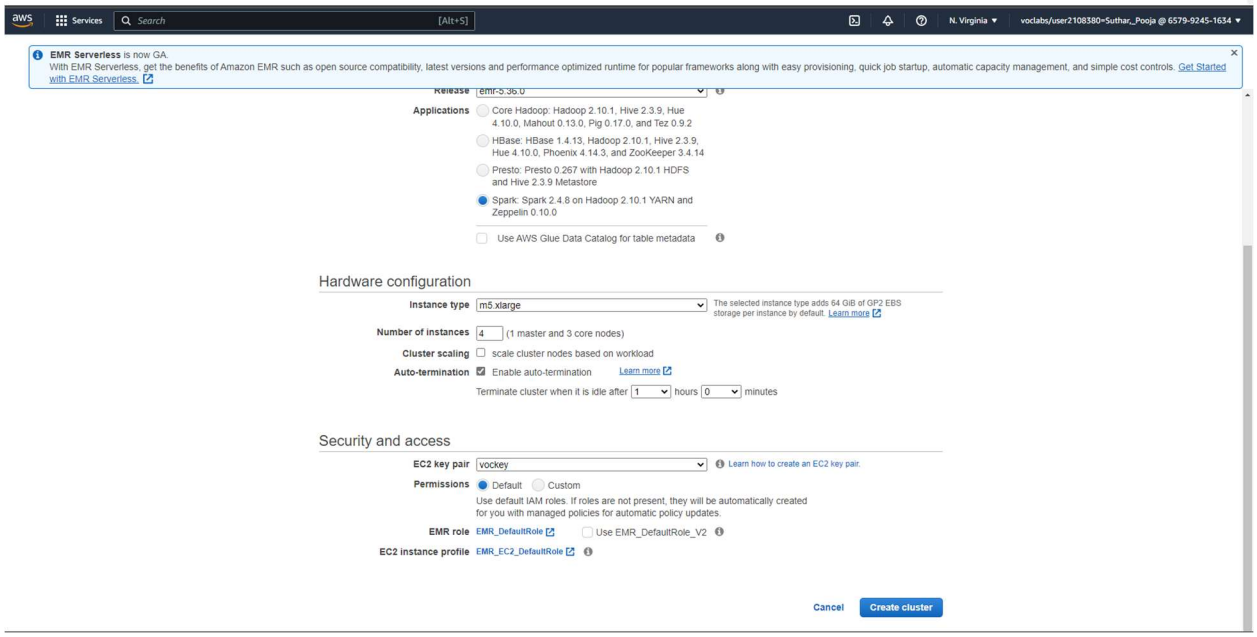
Then load the trained model in Spark application and run prediction app on one EC2 instance. After that do model deployment using Docker i.e., run prediction app on docker container. The link to GitHub and Docker Hub is provided at end.

TASK 1: Parallel training on 4 ec2 Instances

How to create EMR Cluster?

- Log into AWS console,
- Go to EMR Service & Create Cluster
    - Enter cluster name
    - Launch Mode cluster
    - Vendor Amazon
    - Release emr-5.3.10
    - Select spark application – version 2.4.5
    - Hardware Configurations
        - select the instance type
        - number of instances to 4(1 master 3 slaves)
    - select the ec2 key pair or generate one to access the master node.
    - click on create cluster.





Upload files to EMR Cluster Master node

1) Download the ppk key (as I am using PuTTY) from aws learner lab.
    a. Start the lab.
    b. Go to AWS Details tab in Learner lab.
    c. Click on ‘Download PPK’

2) I used PuTTY to connect with ec2 instances. So, in PuTTY upload the downloaded ppk key and start ssh session for master node instance.

3) I used WinSCP to upload files to master node ec2-instance. Create a directory 'Data' under /home/hadoop/ and upload TrainingDataSet.csv and ValidationDataset.csv to Data folder and upload app.jar in /home/hadoop.

Copy files to HDFS :

- Now all files are on our master node we want to move them to HDFS so that all slave nodes can also access them, and we don't have to manually copy them to all ec2 nodes.
- Use this command to copy files from Master node to HDFS.

```
[hadoop@ip-172-31-31-220 ~]$ hadoop fs -put Data/ValidationDataset.csv /user/hadoop/ValidationDataset.csv
```

```
[hadoop@ip-172-31-31-220 ~]$ hadoop fs -put Data/TrainingDataset.csv /user/hadoop/TrainingDataset.csv
```

- Use this command to verify if files are successfully copied to HDFS

```
[hadoop@ip-172-31-31-220 ~]$ hdfs dfs -ls /user/hadoop/
Found 2 items
-rw-r--r--   1 ec2-user hdfsadmingroup      67450 2022-12-09 15:59 /user/hadoop/TrainingDataset.csv
-rw-r--r--   1 ec2-user hdfsadmingroup       8527 2022-12-09 15:59 /user/hadoop/ValidationDataset.csv
```

Launch ModelTrainer application :

- Now everything is done, we want to launch Apache-spark application on EMR cluster.
- Execute following command to run application

```
[hadoop@ip-172-31-31-220 ~]$ sudo spark-submit --class CS643_Programming_Assignment2.Wine_Quality_Training app.jar
```

I used three different Machine Learning model for training

- LogisticRegression Model
- DecisionTreeClassifier Model
- RandomForestClassifier Model

I used Validation Dataset to check the training model F1 score and accuracy score.

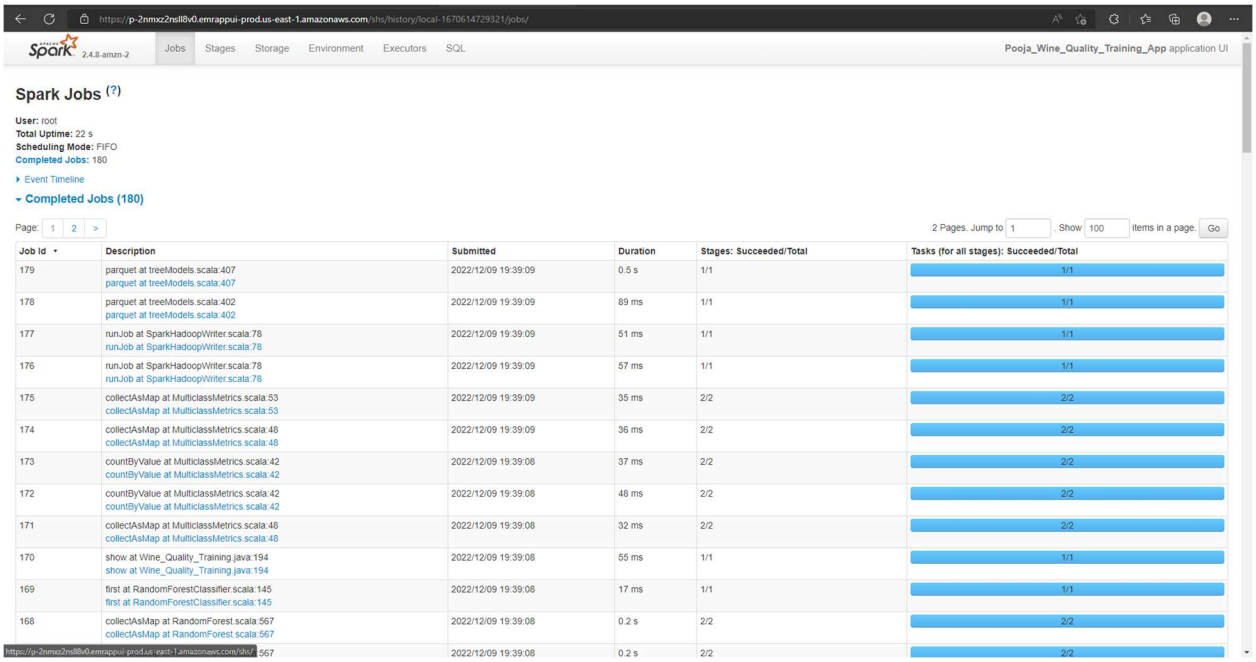
Output:

```
Wine Quality Training
Training Dataset

+-----+-------+---------+----+-------+-------------------+--------------------+---------+--------------+-----------+----------------+-------------+
|label|alcohol|sulphates|  pH|density|free sulfur dioxide|total sulfur dioxide|chlorides|residual sugar|citric acid|volatile acidity|fixed acidity|
+-----+-------+---------+----+-------+-------------------+--------------------+---------+--------------+-----------+----------------+-------------+
|    6|    9.4|     0.53|3.39| 0.9968|               29.0|                60.0|    0.077|           1.8|       0.48|            0.22|          8.9|
|    5|    9.7|     0.65|3.52| 0.9982|               23.0|                71.0|    0.082|           2.3|       0.31|            0.39|          7.6|
|    5|    9.5|     0.91|3.17| 0.9966|               10.0|                37.0|    0.106|           1.6|       0.21|            0.43|          7.9|
|    5|    9.4|     0.53|3.17| 0.9968|                9.0|                67.0|    0.084|           2.3|       0.11|            0.49|          8.5|
|    6|    9.7|     0.63|3.43| 0.9968|               21.0|                40.0|    0.085|           2.4|       0.14|             0.4|          6.9|
+-----+-------+---------+----+-------+-------------------+--------------------+---------+--------------+-----------+----------------+-------------+
only showing top 5 rows

Validation Dataset

+-----+-------+---------+----+-------+-------------------+--------------------+---------+--------------+-----------+----------------+-------------+
|label|alcohol|sulphates|  pH|density|free sulfur dioxide|total sulfur dioxide|chlorides|residual sugar|citric acid|volatile acidity|fixed acidity|
+-----+-------+---------+----+-------+-------------------+--------------------+---------+--------------+-----------+----------------+-------------+
|    5|    9.4|     0.56|3.51| 0.9978|                 11|                  34|    0.076|           1.9|        0.0|             0.7|          7.4|
|    5|    9.8|     0.68| 3.2| 0.9968|                 25|                  67|    0.098|           2.6|        0.0|            0.88|          7.8|
|    5|    9.8|     0.65|3.26|  0.997|                 15|                  54|    0.092|           2.3|       0.04|            0.76|          7.8|
|    6|    9.8|     0.58|3.16|  0.998|                 17|                  60|    0.075|           1.9|       0.56|            0.28|         11.2|
|    5|    9.4|     0.56|3.51| 0.9978|                 11|                  34|    0.076|           1.9|        0.0|             0.7|          7.4|
+-----+-------+---------+----+-------+-------------------+--------------------+---------+--------------+-----------+----------------+-------------+
only showing top 5 rows


Logistic Regression Model

Training DataSet Metrics
Accuracy: 1.0
F-measure: 1.0

 Validation Training Set Metrics
+-----------------------------------------------------------+-----+----------+
|features                                                   |label|prediction|
+-----------------------------------------------------------+-----+----------+
|[5.0,9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5    |5.0       |
|[5.0,9.8,0.68,3.2,0.9968,25.0,67.0,0.098,2.6,0.0,0.88,7.8]  |5    |5.0       |
|[5.0,9.8,0.65,3.26,0.997,15.0,54.0,0.092,2.3,0.04,0.76,7.8] |5    |5.0       |
|[6.0,9.8,0.58,3.16,0.998,17.0,60.0,0.075,1.9,0.56,0.28,11.2]|6    |6.0       |
|[5.0,9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5    |5.0       |
+-----------------------------------------------------------+-----+----------+
only showing top 5 rows


The accuracy of the model is 1.0
F1: 1.0
Decision Tree Model


 Validation Training Set Metrics
+-----------------------------------------------------------+-----+----------+
|features                                                   |label|prediction|
+-----------------------------------------------------------+-----+----------+
|[5.0,9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5    |5.0       |
|[5.0,9.8,0.68,3.2,0.9968,25.0,67.0,0.098,2.6,0.0,0.88,7.8]  |5    |5.0       |
|[5.0,9.8,0.65,3.26,0.997,15.0,54.0,0.092,2.3,0.04,0.76,7.8] |5    |5.0       |
|[6.0,9.8,0.58,3.16,0.998,17.0,60.0,0.075,1.9,0.56,0.28,11.2]|6    |6.0       |
|[5.0,9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5    |5.0       |
+-----------------------------------------------------------+-----+----------+
```

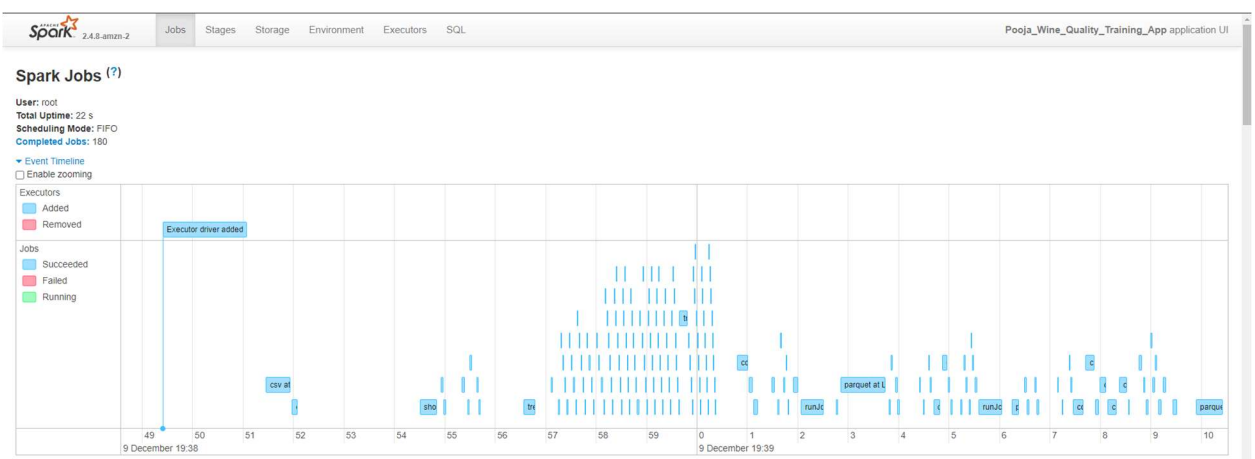
```
Random Forest Model

Training DataSet Metrics
Accuracy: 0.9788897576231431
F-measure: 0.9752474797311071

 Validation Training Set Metrics
+-----------------------------------------------------------+-----+----------+
|features                                                   |label|prediction|
+-----------------------------------------------------------+-----+----------+
|[5.0,9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5    |5.0       |
|[5.0,9.8,0.68,3.2,0.9968,25.0,67.0,0.098,2.6,0.0,0.88,7.8]  |5    |5.0       |
|[5.0,9.8,0.65,3.26,0.997,15.0,54.0,0.092,2.3,0.04,0.76,7.8] |5    |5.0       |
|[6.0,9.8,0.58,3.16,0.998,17.0,60.0,0.075,1.9,0.56,0.28,11.2]|6    |6.0       |
|[5.0,9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5    |5.0       |
+-----------------------------------------------------------+-----+----------+
only showing top 5 rows


The accuracy of the model is 0.95
F1: 0.9356975772765247
[ec2-user@ip-172-31-31-220 hadoop]$
```

So, after seeing the F1 score, Decision Tree Model score is high than other models.

Job execution on Spark.



Event Timeline of Spark:



After running this, the 'Model' folder should have been created under /home/hadoop.
Copy that folder to local machine as we must use those models for prediction.

TASK 2: Predict wine quality on single ec2 instance

At this stage we are interested in executing prediction code on single ec2 instance. For that we need TestDataset.csv, prediction-app.jar and models folder (from task1)

Ec2 instance Create:

- After logging into AWS console,
- Go to EC2 -> launch instance, select AMI
- Select keypair and launch it.

```
ec2-user@ip-172-31-0-61:~
login as: ec2-user
Authenticating with public key "imported-openssh-key"
Last login: Fri Dec  9 22:35:52 2022 from pool-173-72-59-148.cmdnnj.fios.verizon
.net

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
19 package(s) needed for security, out of 31 available
Run "sudo yum update" to apply all updates.
[ec2-user@ip-172-31-0-61 ~]$
```

- **Install JAVA:**

  You must update Java to latest version in each PuTTY terminal.

  - sudo yum install java-1.8.0-openjdk
- **Install SCALA:**
  - wget http://downloads.typesafe.com/scala/2.11.6/scala-2.11.6.tgz
  - tar -xzvf scala-2.11.6.tgz

```
[ec2-user@ip-172-31-0-61 ~]$ wget http://downloads.typesafe.com/scala/2.11.6/scala-2.11.6.tgz
--2022-12-09 23:04:07--  http://downloads.typesafe.com/scala/2.11.6/scala-2.11.6.tgz
Resolving downloads.typesafe.com (downloads.typesafe.com)... 18.165.98.46, 18.165.98.116, 18.165.98.127, ...
Connecting to downloads.typesafe.com (downloads.typesafe.com)|18.165.98.46|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 27130723 (26M) [application/octet-stream]
Saving to: 'scala-2.11.6.tgz'

100%[===================================================================================>] 27,130,723  88.0MB/s   in 0.3s

2022-12-09 23:04:07 (88.0 MB/s) - 'scala-2.11.6.tgz' saved [27130723/27130723]
```

```
[ec2-user@ip-172-31-0-61 ~]$ tar -xzvf scala-2.11.6.tgz
scala-2.11.6/
scala-2.11.6/man/
scala-2.11.6/man/man1/
scala-2.11.6/man/man1/scala.1
scala-2.11.6/man/man1/scalap.1
scala-2.11.6/man/man1/fsc.1
scala-2.11.6/man/man1/scaladoc.1
scala-2.11.6/man/man1/scalac.1
scala-2.11.6/bin/
scala-2.11.6/bin/scalac
scala-2.11.6/bin/fsc
scala-2.11.6/bin/fsc.bat
scala-2.11.6/bin/scala
scala-2.11.6/bin/scalap
scala-2.11.6/bin/scaladoc.bat
scala-2.11.6/bin/scaladoc
scala-2.11.6/bin/scalac.bat
scala-2.11.6/bin/scala.bat
scala-2.11.6/bin/scalap.bat
scala-2.11.6/doc/
scala-2.11.6/doc/tools/
scala-2.11.6/doc/tools/index.html
scala-2.11.6/doc/tools/scalap.html
scala-2.11.6/doc/tools/images/
scala-2.11.6/doc/tools/images/scala_logo.png
scala-2.11.6/doc/tools/images/external.gif
scala-2.11.6/doc/tools/scala.html
scala-2.11.6/doc/tools/css/
scala-2.11.6/doc/tools/css/style.css
scala-2.11.6/doc/tools/fsc.html
scala-2.11.6/doc/tools/scalac.html
scala-2.11.6/doc/tools/scaladoc.html
scala-2.11.6/doc/README
scala-2.11.6/doc/LICENSE.md
scala-2.11.6/doc/licenses/
scala-2.11.6/doc/licenses/mit_jquery-ui.txt
scala-2.11.6/doc/licenses/mit_sizzle.txt
scala-2.11.6/doc/licenses/apache_jansi.txt
scala-2.11.6/doc/licenses/bsd_asm.txt
scala-2.11.6/doc/licenses/mit_tools.tooltip.txt
scala-2.11.6/doc/licenses/mit_jquery-layout.txt
scala-2.11.6/doc/licenses/mit_jquery.txt
scala-2.11.6/doc/licenses/bsd_jline.txt
scala-2.11.6/doc/License.rtf
scala-2.11.6/lib/
scala-2.11.6/lib/scala-parser-combinators_2.11-1.0.3.jar
scala-2.11.6/lib/scala-reflect.jar
scala-2.11.6/lib/akka-actor_2.11-2.3.4.jar
scala-2.11.6/lib/scala-continuations-library_2.11-1.0.2.jar
scala-2.11.6/lib/config-1.2.1.jar
scala-2.11.6/lib/scalap-2.11.6.jar
scala-2.11.6/lib/scala-xml_2.11-1.0.3.jar
scala-2.11.6/lib/scala-continuations-plugin_2.11.5-1.0.2.jar
scala-2.11.6/lib/scala-actors-migration_2.11-1.1.0.jar
scala-2.11.6/lib/jline-2.12.1.jar
scala-2.11.6/lib/scala-library.jar
scala-2.11.6/lib/scala-compiler.jar
scala-2.11.6/lib/scala-actors-2.11.0.jar
scala-2.11.6/lib/scala-swing_2.11-1.0.1.jar
```

- Update PATH environment variable:
  - vim ~/.bashrc
  - copy following lines into file and then save it
    - export SCALA_HOME=/home/ec2-user/scala-2.11.6
    - export PATH=$PATH:/home/ec2-user/scala-2.11.6/bin
  - source ~/.bashrc

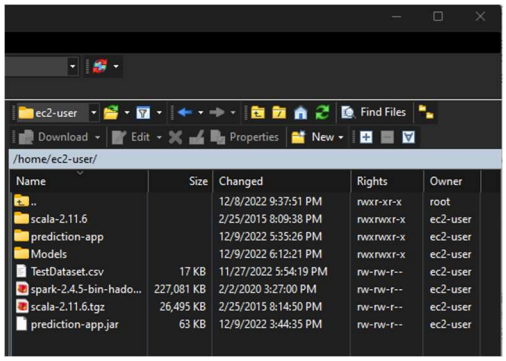
```
[ec2-user@ip-172-31-0-61 ~]$ vim ~/.bashrc
[ec2-user@ip-172-31-0-61 ~]$ source ~/.bashrc
[ec2-user@ip-172-31-0-61 ~]$ cat ~/.bashrc
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

# User specific aliases and functions
export SCALA_HOME=/home/ec2-user/scala-2.11.6
export PATH=$PATH:/home/ec2-user/scala-2.11.6/bin
```

- **Install SPARK:**
    - wget https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz
    - sudo tar xvf spark-2.4.5-bin-hadoop2.7.tgz -C /opt
    - sudo chown -R ec2-user:ec2-user /opt/spark-2.4.5-bin-hadoop2.7
    - sudo ln -fs spark-2.4.5-bin-hadoop2.7 /opt/spark
    - Update PATH Environment
        - vim ~/.bash_profile
        - copy following lines into file and then save it
            - export SPARK_HOME=/opt/spark
            - PATH=$PATH:$SPARK_HOME/bin
            - export PATH
        - source ~/.bash_profile

```
[ec2-user@ip-172-31-0-61 ~]$ wget https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz
--2022-12-09 23:08:16--  https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 232530699 (222M) [application/x-gzip]
Saving to: 'spark-2.4.5-bin-hadoop2.7.tgz'

100%[===================================================================>] 232,530,699 16.1MB/s   in 14s

2022-12-09 23:08:31 (15.4 MB/s) - 'spark-2.4.5-bin-hadoop2.7.tgz' saved [232530699/232530699]
```

```
[ec2-user@ip-172-31-0-61 ~]$ sudo chown -R ec2-user:ec2-user /opt/spark-2.4.5-bin-hadoop2.7
[ec2-user@ip-172-31-0-61 ~]$ sudo ln -fs spark-2.4.5-bin-hadoop2.7 /opt/spark
[ec2-user@ip-172-31-0-61 ~]$ vim ~/.bash_profile
[ec2-user@ip-172-31-0-61 ~]$ source  ~/.bash_profile
[ec2-user@ip-172-31-0-61 ~]$ cat ~/.bash_profile
# .bash_profile

# Get the aliases and functions
if [ -f ~/.bashrc ]; then
        . ~/.bashrc
fi

# User specific environment and startup programs
export SPARK_HOME=/opt/spark
PATH=$PATH:$SPARK_HOME/bin

#PATH=$PATH:$HOME/.local/bin:$HOME/bin

export PATH
```

After all this, copy the files through WinSCP to ec2-user as shown in figure.



- **Disable unnecessary log4j :**
    - cp $SPARK_HOME/conf/log4j.properties.template $SPARK_HOME/conf/log4j.properties
    - vi $SPARK_HOME/conf/log4j.properties
    - (on line 19 of the file, change the log level from INFO to ERROR)
    - log4j.rootCategory=ERROR, console
    - Save the file and exit the text editor

- **Run wine-predict application:**

```
[ec2-user@ip-172-31-0-61 ~]$ spark-submit --class cs643.predictionapp.Wine_Quality_Prediction prediction-app.jar
Prediction-app Running
Prediction Using Logistic Regression Model

TestDataset Metrics

F1 score 0.979338610950453
Accuracy score 0.98125
---------------------------------------------
Prediction-app Running
Prediction Using Descision Tree Classifier Model

TestDataset Metrics

F1 score 1.0
Accuracy score 1.0
---------------------------------------------
Prediction-app Running
Prediction Using Random Forest Classifier Model

TestDataset Metrics

F1 score 0.9667054139625472
Accuracy score 0.975
---------------------------------------------
[ec2-user@ip-172-31-0-61 bin]$
```

TASK 3: Predict wine quality using docker

For Predicting a wine quality on TestDataset.csv using docker. We need to provide local path of TestDataset.csv and pass it as an input argument to docker run command using -v flag. So that TestDataset.csv can be copied to docker container environment.

Test filename must be **TestDataset.csv,** and file must be placed under Data/ directory of container.

Two Ways for running my docker image, ruja2531/ps245_wine_quality:1.2, (found on docker-hub) :

1. Docker run -v <Host_path>/TestDataset.csv:/Data/TestDataset.csv ruja2531/ps245_wine_quality:1.2
2. One can find in the project submission, docker-compose.yml file. Change line 11 <Host_Path>/TestDataset.csv:/Data/TestDataset.csv

```
PS C:\Users\sutha\Docker\Wine_Quality> docker run ruja2531/ps245_wine_quality:1.2 -v myvolume/data.csv:/Data/TestDataset
.csv
Prediction-app Running
Prediction Using Logistic Regression Model

TestDataset Metrics

F1 score 0.979338610950453
Accuracy score 0.98125
----------------------------------------------
Prediction-app Running
Prediction Using Descision Tree Classifier Model

TestDataset Metrics

F1 score 1.0
Accuracy score 1.0
----------------------------------------------
Prediction-app Running
Prediction Using Random Forest Classifier Model

TestDataset Metrics

F1 score 0.9667054139625472
Accuracy score 0.975
----------------------------------------------
```

TASK 4 : Link to Github and DockerHub

GitHub Link: pookri/CS643_Programming_Assignment2 (github.com)

DockerHub Link : Image Layer Details - ruja2531/ps245_wine_quality:1.2 | Docker Hub