CS643861 Programming Assignment 2(mr936)

Setup and Instructions:

1. Setup an EMR cluster with 1 primary node(master), 3 core nodes and 1 task node (slave nodes) with c4.xlarge type
2. Select the cluster termination option as manual, application type – custom (spark, Hadoop, EMR-latest version)
3. All the other settings are defaults
4. ssh into master node : ssh -i "C:\Users\user\Downloads\assn2.pem" ec2-user@ec2-3-220-232-215.compute-1.amazonaws.com
5. copy the datasets into the local ec2 env

   scp -i "C:\Users\user\Downloads\assn2.pem" "C:\Users\user\Downloads\ValidationDataset.csv" ec2-user@ec2-3-220-232-215.compute-1.amazonaws.com:~

   scp -i "C:\Users\user\Downloads\assn2.pem" "C:\Users\user\Downloads\TrainingDataset.csv" ec2-user@ec2-3-220-232-215.compute-1.amazonaws.com:~
6. maven installation
   wget https://archive.apache.org/dist/maven/maven-3/3.6.3/binaries/apache-maven-3.6.3-bin.tar.gz
   tar -xvf apache-maven-3.6.3-bin.tar.gz
   echo "export M2_HOME=/home/hadoop/apache-maven-3.6.3" >> ~/.bashrc
   echo "export PATH=\$M2_HOME/bin:\$PATH" >> ~/.bashrc
   source ~/.bashrc
   mvn –version
7. versions:
   Spark: 3.5.2-amzn-1
   OpenJDK: 1.8.0_432
   Maven: Apache Maven 3.6.3
8. mkdir -p WineQualityPrediction/src/main/java/com/example- for java program
9. Write the java files in the /example, WineQualityPrediction.java for training the model and WineQualityEvaluation.java for prediction using the pretrained model saved in the WineQualityPrediction.java
10. Create a pom.xml file in the WineQualityPrediction/ common for both the java classes after checking the required dependencies (according to the installed versions on your ec2)
11. Build the maven project individually for both the java applications

    For prediction:
    mvn clean package –Pprediction
    Run on cluster
    spark-submit --class com.example.WineQualityPrediction --master yarn /home/ec2-user/WineQualityPrediction/target/wine-quality-project-1.0-SNAPSHOT.jar

```
[ec2-user@ip-172-31-8-247 target]$ spark-submit --class com.example.WineQualityPrediction --master yarn /home/ec2-user/WineQualityPrediction/target/wine-qua
lity-project-1.0-SNAPSHOT.jar
```

```
24/12/08 20:27:05 INFO CodeGenerator: Code generated in 5.227924 ms
24/12/08 20:27:05 INFO SQLExecution: Generating and posting SparkListenerSQLExecutionObfuscatedInfo...
24/12/08 20:27:05 INFO SQLExecution: Posted SparkListenerSQLExecutionObfuscatedInfo in 0 ms
+-------+----------+
|quality|prediction|
+-------+----------+
|      0|       1.0|
|      1|       1.0|
|      0|       0.0|
|      1|       1.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
|      0|       0.0|
+-------+----------+
only showing top 20 rows

24/12/08 20:27:05 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/08 20:27:05 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-8-247.ec2.internal:4040
24/12/08 20:27:05 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/08 20:27:05 INFO MemoryStore: MemoryStore cleared
24/12/08 20:27:05 INFO BlockManager: BlockManager stopped
24/12/08 20:27:05 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/08 20:27:05 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/08 20:27:05 INFO SparkContext: Successfully stopped SparkContext
24/12/08 20:27:05 INFO ShutdownHookManager: Shutdown hook called
24/12/08 20:27:05 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-3b71adfa-4dee-41b9-bfa6-5f0878edf1ee
24/12/08 20:27:05 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-2d2ba89c-2147-411b-a448-8f9f424ac6a1
[ec2-user@ip-172-31-8-247 ~]$
```

For evaluation:

mvn clean package –Pevaluation

Run on single node

spark-submit --class com.example.WineQualityEvaluation --master local /home/ec2-user/WineQualityPrediction/target/wine-quality-project-1.0-SNAPSHOT.jar

```
[hadoop@ip-172-31-4-164 WineQualityPrediction]$ spark-submit --class com.example.WineQualityEvaluation --master local /home/hadoop/WineQualityPrediction
/target/wine-quality-project-1.0-SNAPSHOT.jar
```

```
24/12/08 22:39:44 INFO TaskSetManager: Finished task 0.0 in stage 16.0 (TID 11) in 103 ms on ip-172-31-8-247.ec2.internal (executor driver) (1/1)
24/12/08 22:39:44 INFO TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed, from pool
24/12/08 22:39:44 INFO DAGScheduler: ResultStage 16 (show at WineQualityEvaluation.java:78) finished in 0.120 s
24/12/08 22:39:44 INFO DAGScheduler: Job 9 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/08 22:39:44 INFO TaskSchedulerImpl: Killing all running tasks in stage 16: Stage finished
24/12/08 22:39:44 INFO DAGScheduler: Job 9 finished: show at WineQualityEvaluation.java:78, took 0.140885 s
24/12/08 22:39:44 INFO CodeGenerator: Code generated in 7.368638 ms
24/12/08 22:39:44 INFO SQLExecution: Generating and posting SparkListenerSQLExecutionObfuscatedInfo...
24/12/08 22:39:44 INFO SQLExecution: Posted SparkListenerSQLExecutionObfuscatedInfo in 0 ms
+-------+----------+--------------------+
|quality|prediction|         probability|
+-------+----------+--------------------+
|      0|       0.0|[0.99341230595121...|
|      0|       0.0|[0.99291135976049...|
|      0|       0.0|[0.98975302687522...|
|      0|       0.0|[0.92688563095772...|
|      0|       0.0|[0.99341230595121...|
|      0|       0.0|[0.99241246585409...|
|      0|       0.0|[0.99229042475609...|
|      1|       0.0|[0.98882667841722...|
|      1|       0.0|[0.98836434300442...|
|      0|       0.0|[0.92780032471469...|
|      0|       0.0|[0.99340735995123...|
|      0|       0.0|[0.92780032471469...|
|      0|       0.0|[0.98998507790332...|
|      0|       0.0|[0.94822443087016...|
|      0|       0.0|[0.98316413872534...|
|      0|       0.0|[0.98134471338674...|
|      1|       0.0|[0.85588000980082...|
|      0|       0.0|[0.95581146291418...|
|      0|       0.0|[0.99459261608876...|
|      0|       0.0|[0.94115823402374...|
+-------+----------+--------------------+
only showing top 20 rows

24/12/08 22:39:44 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/08 22:39:44 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-8-247.ec2.internal:4040
24/12/08 22:39:44 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/08 22:39:45 INFO MemoryStore: MemoryStore cleared
24/12/08 22:39:45 INFO BlockManager: BlockManager stopped
24/12/08 22:39:45 INFO BlockManagerMaster: BlockManagerMaster stopped
```

```
24/12/09 02:00:49 INFO DAGScheduler: Job 7 finished: collectAsMap at MulticlassMetrics.scala:61, took 0.115257 s
F1 Score: 0.8104636591478698
24/12/09 02:00:49 INFO FileSourceStrategy: Pushed Filters:
```

```
24/12/09 02:00:48 INFO DAGScheduler: Job 6 finished: collectAsMap at MulticlassMetrics.scala:61, took 0.252416 s
Accuracy: 0.84375
24/12/09 02:00:48 INFO FileSourceStrategy: Pushed Filters:
```

12. Upload the datasets, .java files and .jar file to s3 bucket

aws s3 cp /home/hadoop/WineQualityPrediction/target/wine-quality-project-1.0-SNAPSHOT.jar s3://njitcs643/

```
[hadoop@ip-172-31-4-164 .aws]$ aws s3 ls
2024-12-08 00:05:27 aws-emr-studio-294737378300-us-east-1
2024-12-07 20:52:27 aws-logs-294737378300-us-east-1
2024-10-20 18:59:13 njitcs643
[hadoop@ip-172-31-4-164 .aws]$ aws s3 cp /home/hadoop/WineQualityPrediction/target/wine-quality-project-1.0-SNAPSHOT.jar s3://njitcs643/
upload: ../WineQualityPrediction/target/wine-quality-project-1.0-SNAPSHOT.jar to s3://njitcs643/wine-quality-project-1.0-SNAPSHOT.jar
[hadoop@ip-172-31-4-164 .aws]$ aws s3 ls s3://njitcs643/
                           PRE images/
2024-10-20 19:20:52       9632 1.jpg
2024-10-20 19:27:26      62871 10.jpg
2024-10-20 19:26:42      52101 2.jpg
2024-10-20 19:26:48      39816 3.jpg
2024-10-20 19:26:54     117860 4.jpg
2024-10-20 19:26:59      87437 5.jpg
2024-10-20 19:27:04     129299 6.jpg
2024-10-20 19:27:10    1013668 7.jpg
2024-10-20 19:27:15      13781 8.jpg
2024-10-20 19:27:19     527810 9.jpg
2024-11-21 00:38:14      68804 TrainingDataset.csv
2024-11-21 00:09:23       8760 ValidationDataset.csv
2024-12-09 02:34:09  258903436 wine-quality-project-1.0-SNAPSHOT.jar
[hadoop@ip-172-31-4-164 .aws]$
```

```
[hadoop@ip-172-31-4-164 ~]$ aws s3 ls s3://njitcs643/
                           PRE data/
                           PRE images/
                           PRE metadata/
2024-10-20 19:20:52       9632 1.jpg
2024-10-20 19:27:26      62871 10.jpg
2024-10-20 19:26:42      52101 2.jpg
2024-10-20 19:26:48      39816 3.jpg
2024-10-20 19:26:54     117860 4.jpg
2024-10-20 19:26:59      87437 5.jpg
2024-10-20 19:27:04     129299 6.jpg
2024-10-20 19:27:10    1013668 7.jpg
2024-10-20 19:27:15      13781 8.jpg
2024-10-20 19:27:19     527810 9.jpg
2024-11-21 00:38:14      68804 TrainingDataset.csv
2024-11-21 00:09:23       8760 ValidationDataset.csv
2024-12-09 03:35:35       3860 WineQualityEvaluation.java
2024-12-09 03:36:00       3612 WineQualityPrediction.java
2024-12-09 02:34:09  258903436 wine-quality-project-1.0-SNAPSHOT.jar
[hadoop@ip-172-31-4-164 ~]$
```

13. Building the docker application

nano Dockerfile
sudo docker build -t mr936/spark-prediction:1
sudo docker run --rm -m 4g -e BUCKET_NAME=dataset/ -v /home/hadoop:/home/hadoop mr936/spark-prediction:1
sudo docker tag mr936/spark-prediction:1 meghnareddi/spark-prediction:1
sudo docker push meghnareddi/spark-prediction:1

```
[hadoop@ip-172-31-4-164 WineQualityPrediction]$ nano Dockerfile
[hadoop@ip-172-31-4-164 WineQualityPrediction]$ sudo docker build -t mr936/spark-prediction:1 .
[+] Building 0.3s (9/9) FINISHED                                                             docker:default
 => [internal] load build definition from Dockerfile                                                   0.0s
 => => transferring dockerfile: 519B                                                                   0.0s
 => [internal] load metadata for docker.io/library/openjdk:8-jre-slim                                  0.2s
 => [auth] library/openjdk:pull token for registry-1.docker.io                                         0.0s
 => [internal] load .dockerignore                                                                      0.0s
 => => transferring context: 2B                                                                        0.0s
 => [1/3] FROM docker.io/library/openjdk:8-jre-slim@sha256:53186129237fbb8bc0a12dd36da6761f4c7a2a20233c20d4eb0d497e4045a4f5  0.0s
 => [internal] load build context                                                                      0.0s
 => => transferring context: 212B                                                                      0.0s
 => CACHED [2/3] WORKDIR /app                                                                          0.0s
 => CACHED [3/3] COPY target/wine-quality-project-1.0-SNAPSHOT.jar /app/wine-quality-project-1.0-SNAPSHOT.jar  0.0s
 => exporting to image                                                                                 0.0s
 => => exporting layers                                                                                0.0s
 => => writing image sha256:69837d0417731173da3f5fab38f196f62df2244b1fa1e417efacfb13ab29c40f           0.0s
 => => naming to docker.io/mr936/spark-prediction:1                                                    0.0s
```

```
[hadoop@ip-172-31-4-164 WineQualityPrediction]$ sudo docker run --rm -m 4g -e BUCKET_NAME=dataset/ -v /home/hadoop:/home/hadoop mr936/spark-prediction:1
Accuracy: 0.84375
F1 Score: 0.8104636591478698
+-------+----------+--------------------+
|quality|prediction|         probability|
+-------+----------+--------------------+
|      0|       0.0|[0.99341230595121...|
|      0|       0.0|[0.99291135976049...|
|      0|       0.0|[0.98975302687522...|
|      0|       0.0|[0.92688563095772...|
|      0|       0.0|[0.99341230595121...|
|      0|       0.0|[0.99241246585409...|
|      0|       0.0|[0.99229042475609...|
|      1|       0.0|[0.98882667841722...|
|      1|       0.0|[0.98836434300442...|
|      0|       0.0|[0.92780032471469...|
|      0|       0.0|[0.99340735995123...|
|      0|       0.0|[0.92780032471469...|
|      0|       0.0|[0.98998507790332...|
|      0|       0.0|[0.94822443087016...|
|      0|       0.0|[0.98316413872534...|
|      0|       0.0|[0.98134471338674...|
|      1|       0.0|[0.85588000980082...|
|      0|       0.0|[0.95581146291418...|
|      0|       0.0|[0.99459261608876...|
|      0|       0.0|[0.94115823402374...|
+-------+----------+--------------------+
only showing top 20 rows
```

```
[hadoop@ip-172-31-4-164 ~]$ sudo docker login
Log in with your Docker ID or email address to push and pull images from Docker Hub. If you don't have a Docker ID, head over to https://hub.docker.com/ to
create one.
You can log in with your password or a Personal Access Token (PAT). Using a limited-scope PAT grants better security and is required for organizations using
 SSO. Learn more at https://docs.docker.com/go/access-tokens/

Username: meghnareddi
Password:
WARNING! Your password will be stored unencrypted in /root/.docker/config.json.
Configure a credential helper to remove this warning. See
https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
[hadoop@ip-172-31-4-164 ~]$ sudo docker push meghnareddi/spark-prediction:1
The push refers to repository [docker.io/meghnareddi/spark-prediction]
0abd12158e9f: Pushed
23388dce9c4c: Pushed
b66078cf4b41: Mounted from library/openjdk
cd5a0a9f1e01: Mounted from library/openjdk
eafe6e032dbd: Mounted from library/openjdk
92a4e8a3140f: Mounted from library/openjdk
1: digest: sha256:8968cd42342166c9a2d6eeac1bcaf84d37cce7dde52660c27eb705613d6164cf size: 1578
```