CS643 - PROGRAMMING ASSIGNMENT 2

Wine Quality Prediction ML Model

This project involves creating a Python application that utilizes the PySpark interface. 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 predict wine quality. Docker is used to produce a container image for the trained machine learning model, simplifying the deployment process.

The primary Python source files in this project are:

1. winequilityprediction.py: Reads the training dataset from S3 and trains the model in parallel on an EMR Spark cluster. Once trained, the model can be executed on provided test data via S3. The program stores the trained model in the S3 bucket.

2. winequilitytestdataprediction.py: Loads the trained model and executes it on a given test data file. This program prints the F1 score as a metric for the accuracy of the trained model.

GitHub: https://github.com/Nissi-Prabhatha/Cloud_Computing

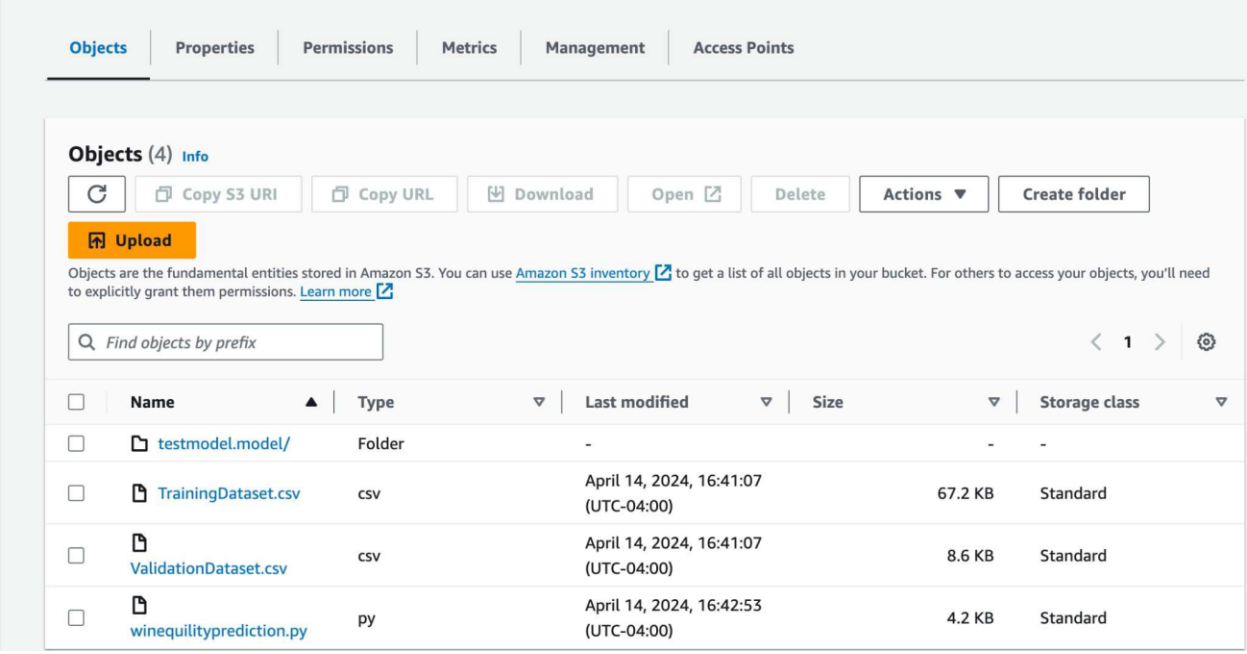
Docker: https://hub.docker.com/r/nissig/pa2-docker

Steps to Launch an EMR Cluster on AWS and Train the ML Model without Docker:

- Create an S3 bucket and upload the following files:

winequalityprediction.py, TrainingDataset.csv and ValidationDataset.csv in

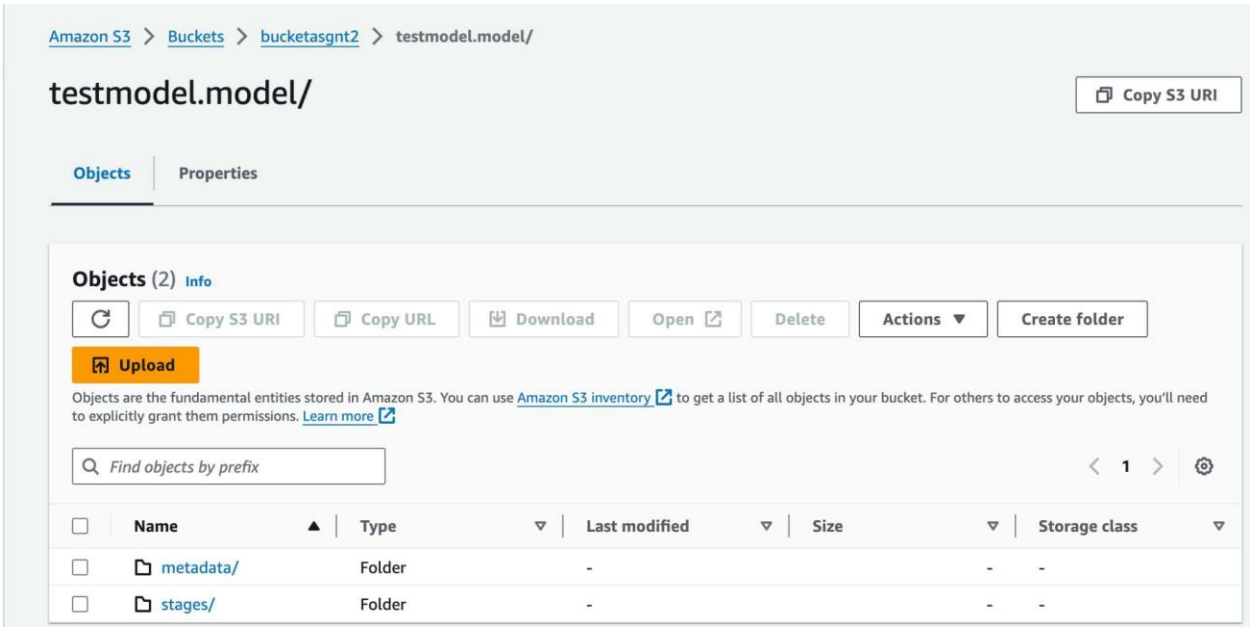
the bucket.



- Navigate to the EMR Service and configure the EMR cluster.
    a. Provide the details such as Cluster Configuration: (Spark, Hadoop).
    b. Select the EC2 instance type for the cluster nodes.
    c. Select the number of instances for the cluster.
    d. Provide the EC2 key pair that will be used to SSH connect with the instance.
- Connect to the EC2 instance using the SSH command and use the key pair defined above.
- After the connection is successful, submit the task for execution.

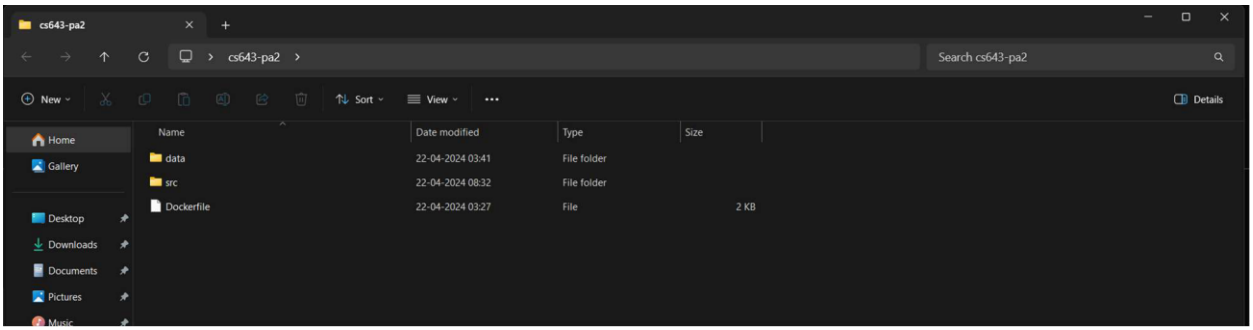
  Command: spark-submit s3://bucketasgnt2/winequilityprediction.py

- Once the task execution is complete the model will be generated in the S3 bucket defined above.

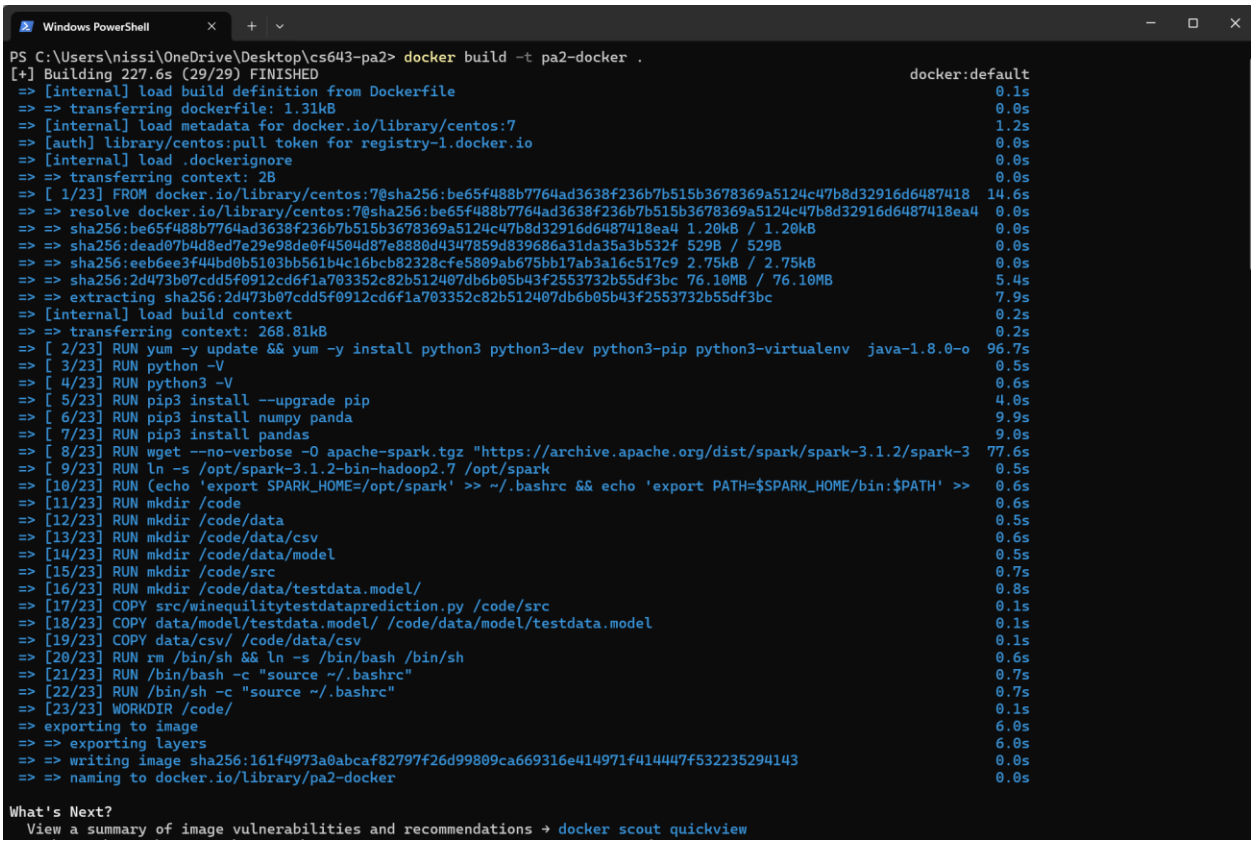


Steps for running the ML Model with Docker:

- Open Terminal and navigate to the folder where the Docker File are present.



- Run the Docker build command to build the Docker image.
  Command: docker build -t pa2-docker.



- Now login to Docker hub to push the image.
  Command: docker login -u nissig

- Run the below command to push the image on Docker hub.

  Command:

  docker tag NissiG/pa2-docker nissig/pa2-docker

  docker push nissig/pa2-docker

```
PS C:\Users\nissi\OneDrive\Desktop\cs643-pa2> docker tag NissiG/pa2-docker nissig/pa2-docker
PS C:\Users\nissi\OneDrive\Desktop\cs643-pa2> docker push nissig/pa2-docker
Using default tag: latest
The push refers to repository [docker.io/nissig/pa2-docker]
5f70bf18a086: Mounted from srujit12091997/apache-test
e44ea1f0c9ca: Pushed
fe3682496427: Pushed
6665f85289ea: Pushed
5e6e7ff91d20: Pushed
347e49e2a168: Pushed
ec8b4a8557b1: Pushed
85d1eeedef2f: Pushed
a12f2a597cc4: Pushed
b7ac8de9baf4: Pushed
24a2082d1a00: Pushed
b16995428680: Pushed
588a846cbccd: Pushed
c5b22b030e2d: Pushed
488320b74288: Pushed
ef2227c84052: Pushed
b97c229c29ce: Pushed
1903a2fdcd39: Pushed
90e3f225bd8c: Pushed
bc03fd35189d: Pushed
99f9f232871b: Pushed
174f56854903: Mounted from library/centos
latest: digest: sha256:5b08cf5edee5782647243000e58d264211d898214820f7e777911d7426bc8818 size: 5109
```

- Now pull the image from the docker hub on the machine where you want to run the Docker image.

Command: docker pull nissig/pa2-docker

```
PS C:\Users\nissi\OneDrive\Desktop\cs643-pa2> docker pull nissig/pa2-docker
Using default tag: latest
latest: Pulling from nissig/pa2-docker
Digest: sha256:5b08cf5edee5782647243000e58d264211d898214820f7e777911d7426bc8818
Status: Image is up to date for nissig/pa2-docker:latest
docker.io/nissig/pa2-docker:latest

What's Next?
  View a summary of image vulnerabilities and recommendations → docker scout quickview nissig/pa2-docker
```

- Now run the Docker run command to execute the image and see the results.

Command: docker run -v C:/Users/nissi/OneDrive/Desktop/cs643-pa2/data:/data pa2-docker ValidationDataset.csv

```
PS C:\Users\nissi\OneDrive\Desktop\cs643-pa2> docker run -v C:/Users/nissi/OneDrive/Desktop/cs643-pa2/data:/data pa2-docker ValidationDataset.csv
24/04/29 01:39:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
24/04/29 01:39:41 INFO SparkContext: Running Spark version 3.1.2
24/04/29 01:39:42 INFO ResourceUtils: ==============================================================
24/04/29 01:39:42 INFO ResourceUtils: No custom resources configured for spark.driver.
24/04/29 01:39:42 INFO ResourceUtils: ==============================================================
24/04/29 01:39:42 INFO SparkContext: Submitted application: cs643_wine_prediction
24/04/29 01:39:42 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory ->
name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/04/29 01:39:42 INFO ResourceProfile: Limiting resource is cpu
24/04/29 01:39:42 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/04/29 01:39:42 INFO SecurityManager: Changing view acls to: root
24/04/29 01:39:42 INFO SecurityManager: Changing modify acls to: root
24/04/29 01:39:42 INFO SecurityManager: Changing view acls groups to:
24/04/29 01:39:42 INFO SecurityManager: Changing modify acls groups to:
24/04/29 01:39:42 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(root); groups with view pe
rmissions: Set(); users  with modify permissions: Set(root); groups with modify permissions: Set()
24/04/29 01:39:42 INFO Utils: Successfully started service 'sparkDriver' on port 43403.
24/04/29 01:39:42 INFO SparkEnv: Registering MapOutputTracker
24/04/29 01:39:42 INFO SparkEnv: Registering BlockManagerMaster
24/04/29 01:39:42 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/04/29 01:39:42 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/04/29 01:39:42 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/29 01:39:42 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-722ea975-be20-4722-969e-f38bf0a606a1
24/04/29 01:39:42 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
24/04/29 01:39:42 INFO SparkEnv: Registering OutputCommitCoordinator
24/04/29 01:39:43 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/29 01:39:43 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://00af7a2d8d2b:4040
24/04/29 01:39:43 INFO Executor: Starting executor ID driver on host 00af7a2d8d2b
24/04/29 01:39:43 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 45217.
24/04/29 01:39:43 INFO NettyBlockTransferService: Server created on 00af7a2d8d2b:45217
24/04/29 01:39:43 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/29 01:39:43 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 00af7a2d8d2b, 45217, None)
24/04/29 01:39:43 INFO BlockManagerMasterEndpoint: Registering block manager 00af7a2d8d2b:45217 with 366.3 MiB RAM, BlockManagerId(driver, 00af7a2d8d2b, 45217,
None)
24/04/29 01:39:43 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 00af7a2d8d2b, 45217, None)
24/04/29 01:39:43 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 00af7a2d8d2b, 45217, None)
24/04/29 01:39:44 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('file:/code/spark-warehouse').
24/04/29 01:39:44 INFO SharedState: Warehouse path is 'file:/code/spark-warehouse'.
Test data for Input file
```

```
Test data for Input file
data/csv/ValidationDataset.csv
+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+-----+---
-----------------+--------------------+----------+
|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density|  pH|sulphates|alcohol|quality|            features|label|
    rawPrediction|         probability|prediction|
+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+-----+---
-----------------+--------------------+----------+
|          7.4|             0.7|        0.0|           1.9|    0.076|               11.0|                34.0| 0.9978|3.51|     0.56|    9.4|    5.0|[7.4,0.7,0.0,1.9,...|  0.0|[47
.8700004929938...|[0.95740000985987...|       0.0|
|          7.8|            0.88|        0.0|           2.6|    0.098|               25.0|                67.0| 0.9968| 3.2|     0.68|    9.8|    5.0|[7.8,0.88,0.0,2.6...|  0.0|[46
.3984230232075...|[0.92796846046415...|       0.0|
|          7.8|            0.76|       0.04|           2.3|    0.092|               15.0|                54.0|  0.997|3.26|     0.65|    9.8|    5.0|[7.8,0.76,0.04,2....|  0.0|[44
.5162339884992...|[0.89032467976998...|       0.0|
|         11.2|            0.28|       0.56|           1.9|    0.075|               17.0|                60.0|  0.998|3.16|     0.58|    9.8|    6.0|[11.2,0.28,0.56,1...|  1.0|[1.
1911066631060l...|[0.02382213326212...|       1.0|
|          7.4|             0.7|        0.0|           1.9|    0.076|               11.0|                34.0| 0.9978|3.51|     0.56|    9.4|    5.0|[7.4,0.7,0.0,1.9,...|  0.0|[47
.8700004929938...|[0.95740000985987...|       0.0|
+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+-----+---
-----------------+--------------------+----------+
only showing top 5 rows

None
Wine prediction model Test Accuracy =  0.9625
Wine prediction model for Weighted f1 score =  0.9479401629072682
PS C:\Users\nissi\OneDrive\Desktop\cs643-pa2>
```