Objective:

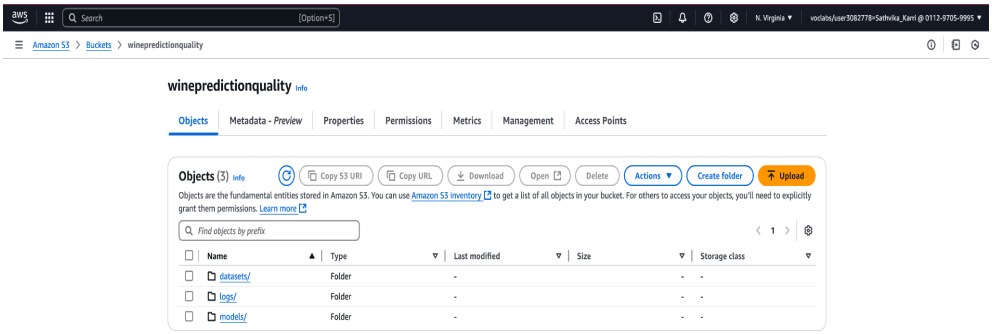
The goal is to develop a Python application using PySpark for predicting wine quality. This application is deployed on an AWS Elastic MapReduce (EMR) cluster, where training occurs in parallel across multiple EC2 instances using available datasets. A Docker container is created to streamline the deployment of the trained machine learning model.

Repositories:

- **GitHub Repository**: https://github.com/SathvikaKarri/CC-Programming-Assignment-2
- **Docker:** **https://hub.docker.com/repository/docker/sathvikarri/wineprediction/general**

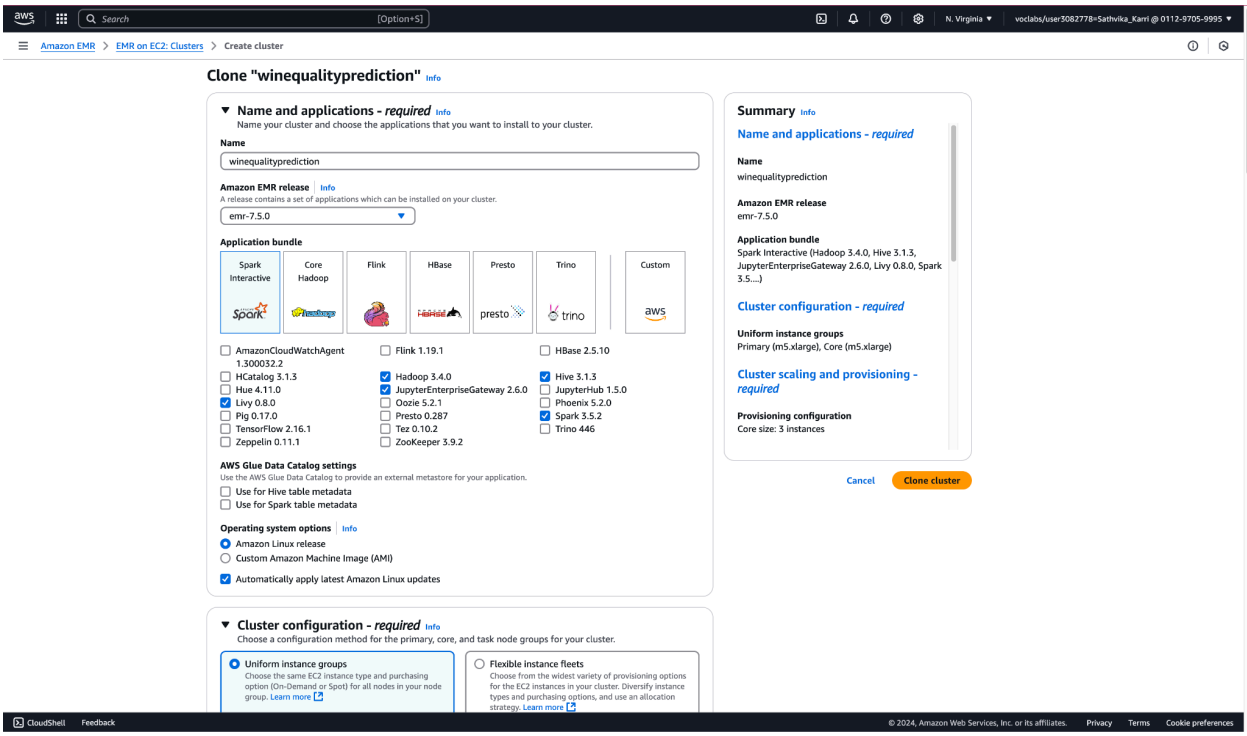
Execution Steps:

1. **Create Key-Pair for the EMR Cluster**:
    - Go to EC2 > Key Pairs.
    - Generate a key pair named sathvi.pem and download it in .pem format.
2. **Create an S3 Bucket**:
    - Create an S3 bucket with the name winepredictionquality

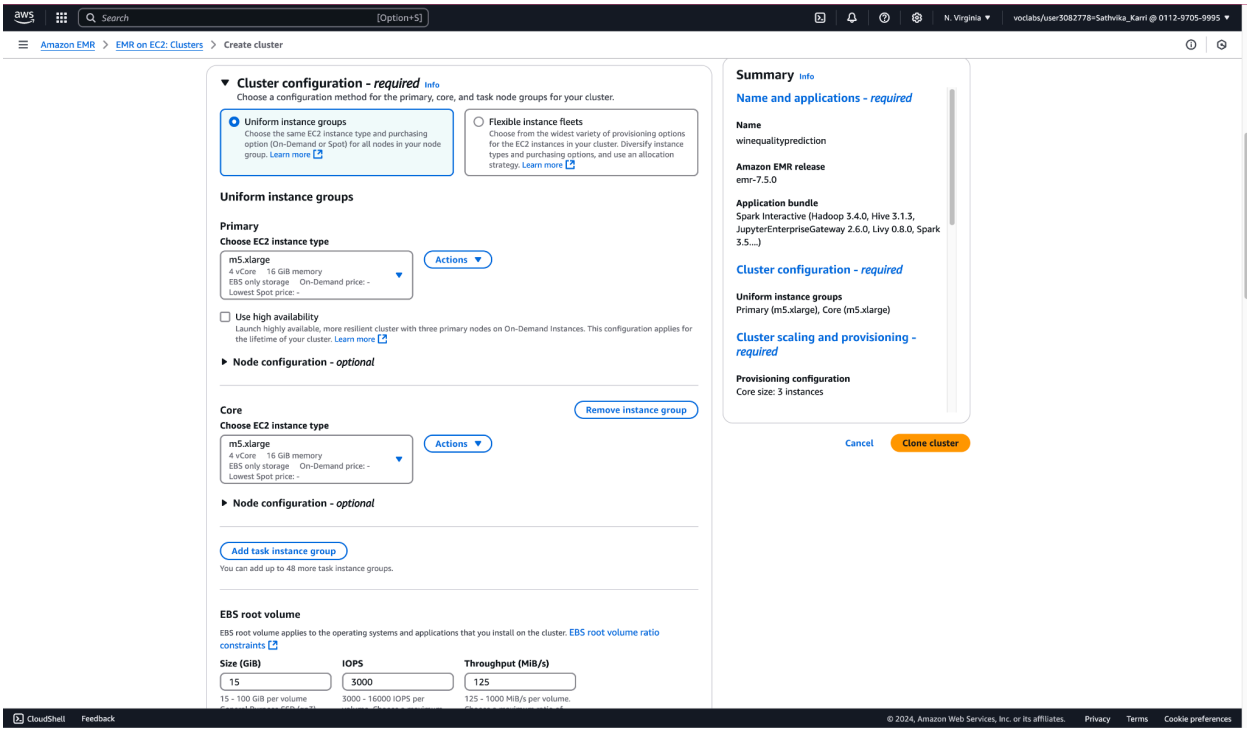


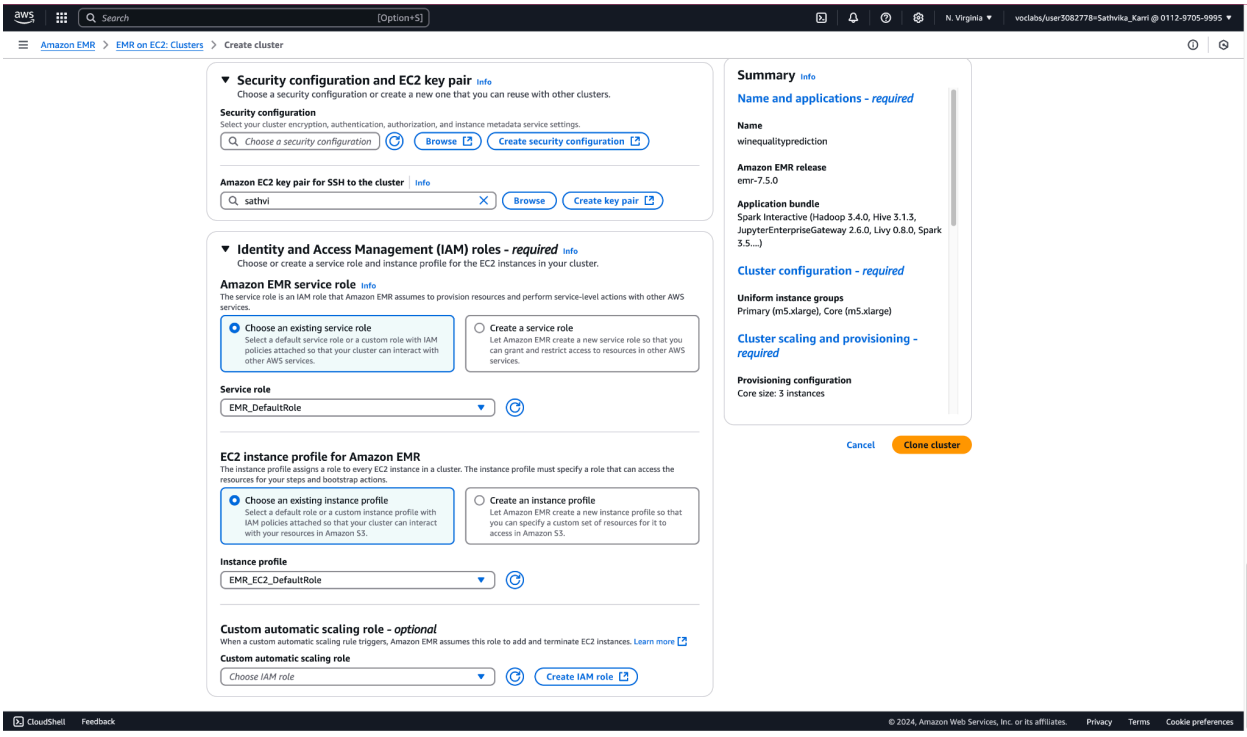
3. **Set Up the EMR Cluster**:
    - Navigate to the EMR Console and create a cluster named winepredictionquality.

- Use the EMR release version emr-7.5.0, including Hadoop 3.4.0 and Spark 3.5.2.



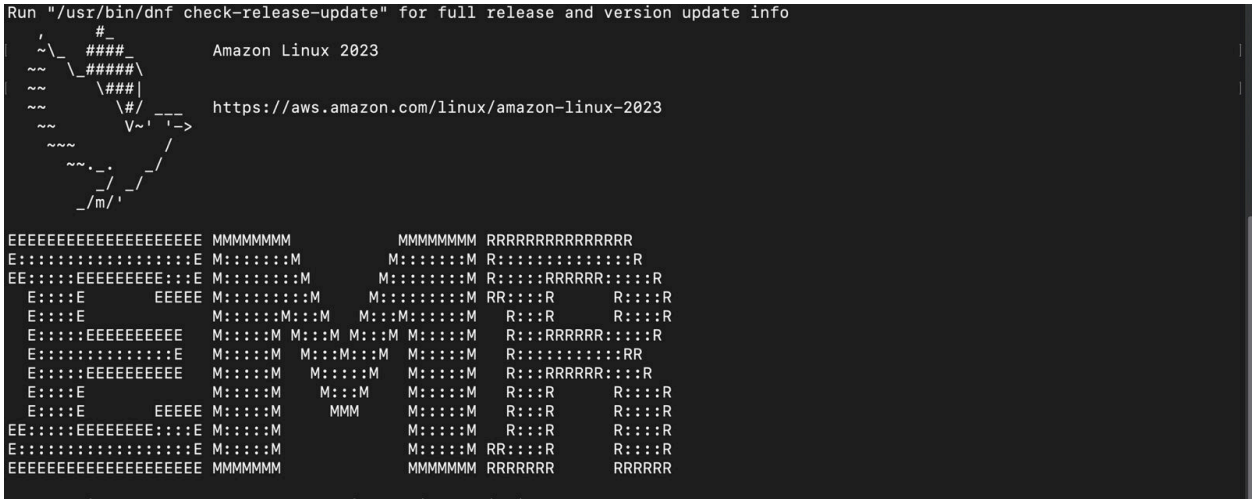
4. **Configure the Spark Cluster**:
   - Create a cluster.
   - Configure cluster scaling, provisioning, networking, termination policies, and security using IAM roles and the sathvi.pem EC2 key pair.





5. Train the Machine Learning Model on EC2:

ssh -i “sathvi.pem" ec2-user@<ec2-public-dns>



Submit the training job using:

spark-submit winequality.py
s3://winepredictionquality/datasets/TrainingDataset.csv
s3://winepredictionquality/datasets/ValidationDataset.csv

This splits the TrainingDataset.csv into 90% for training and 10% for testing. The test data is saved as TestDataset.csv in the S3 dataset folder.

```
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilte
r
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmI
pFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilte
r.AmIpFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.
AmIpFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfi
lter.AmIpFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFi
lter
24/12/08 22:36:43 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:36:43 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegiste
redResourcesRatio: 0.0
INFO:__main__:2024-12-08 22:36:43.774507 - Setting the Spark logging level to 'ERROR' for a cleaner output...
INFO:__main__:2024-12-08 22:36:43.776447 - Loading the training data from s3://winepredictionquality/datasets/TrainingDataset.
csv and preparing it for processing...
INFO:__main__:2024-12-08 22:36:50.455992 - Dividing the data into training and test sets (90/10 split)...
INFO:__main__:2024-12-08 22:36:50.480205 - Saving the test dataset to S3 as a single file for future reference...
INFO:botocore.credentials:Found credentials from IAM Role: EMR_EC2_DefaultRole
INFO:__main__:2024-12-08 22:36:53.746254 - Loading and cleaning the validation data from s3://winepredictionquality/datasets/V
alidationDataset.csv...
INFO:__main__:2024-12-08 22:36:54.164908 - Building the machine learning pipeline and setting up cross-validation...
INFO:__main__:2024-12-08 22:36:54.223450 - Commencing training with cross-validation...
INFO:py4j.clientserver:Closing down clientserver connection
INFO:py4j.clientserver:Closing down clientserver connection
INFO:py4j.clientserver:Closing down clientserver connection
INFO:py4j.clientserver:Closing down clientserver connection
INFO:py4j.clientserver:Closing down clientserver connection
INFO:__main__:2024-12-08 22:38:46.316940 - Evaluating the model performance on the validation data...
INFO:__main__:2024-12-08 22:38:47.595161 - Model performance: Accuracy = 0.5500
INFO:__main__:2024-12-08 22:38:47.595297 - Weighted F1 Score of the best model: 0.5361
INFO:__main__:2024-12-08 22:38:47.595340 - Saving the best performing model to s3://winepredictionquality/models/winemodel...
INFO:__main__:2024-12-08 22:38:54.206098 - Spark session terminated successfully.
INFO:py4j.clientserver:Closing down clientserver connection
[hadoop@ip-172-31-65-219 ~]$
```

Save the trained model to:

s3://winepredictionquality/models

Test the Model:

Run the test using:
spark-submit testwinequality.py
s3://winequalityprediction/datasets/TestDataset.csv
s3://winepredictionquality/models

```
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /stages/pool: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /stages/pool/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFil
ter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /storage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /storage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /storage/rdd: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /storage/rdd/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFil
ter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /environment: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFil
ter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilte
r
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmI
pFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilte
r.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.
AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfi
lter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFi
lter
24/12/08 22:52:18 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/08 22:52:18 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegiste
redResourcesRatio: 0.0
INFO:__main__:2024-12-08 22:52:18.605395 - Setting the Spark logging level to 'ERROR' for cleaner output...
INFO:__main__:2024-12-08 22:52:18.607213 - Loading the test data from s3://winepredictionquality/datasets/TestDataset.csv...
INFO:__main__:2024-12-08 22:52:25.951203 - Loading the trained model from s3://winepredictionquality/models/winemodel...
INFO:__main__:2024-12-08 22:52:43.168071 - Making predictions on the test dataset...
INFO:__main__:2024-12-08 22:52:43.498195 - Evaluating the model performance...
INFO:__main__:2024-12-08 22:52:46.700552 - Test Data Model Performance: Accuracy = 0.7342
INFO:__main__:2024-12-08 22:52:46.700729 - Test Data Weighted F1 Score: 0.7243
INFO:__main__:2024-12-08 22:52:47.232644 - Spark session terminated successfully.
INFO:py4j.clientserver:Closing down clientserver connection
[hadoop@ip-172-31-65-219 ~]$
```

Using Docker for Deployment:

Set Up Docker:

Create a Docker account and install Docker on your local machine.

Log in to Docker and build the image:

docker build -t wine-quality-test .

Push the Image:

Tag the image:

docker tag winepredictionsathvikarri/winepredictionsathvikarri

Push to Docker Hub:

docker push sathvikarri/wineprediction

Pull the Image:

Pull the Docker image:

docker pull sathvikarri/wineprediction

Run the Docker Container:

Run the container with the test dataset and saved model:

docker run --rm wineprediction /app/datasets/TestDataset.csv /app/models

```
C:\Users\Sathvika\Downloads\ProgAss2\ProgAss2>docker run --rm sathvikarri/wineprediction /app/datasets/TestDataset.csv /app/models
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/12/09 05:36:46 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where a
pplicable
Loading test dataset from /app/datasets/TestDataset.csv...
Loading model from /app/models...
Making predictions on the test dataset...
Model performance: Accuracy = 0.7342
Weighted F1 Score of the best model: 0.7243
```

Model Accuracy:

- **Accuracy**: 0.73