Wine Quality Prediction System using PySpark:

This project implements a distributed machine learning system for wine quality prediction using Apache Spark ML on AWS EC2. The system includes training and prediction components, containerized with Docker for easy deployment.

Dockerhub Repository : https://hub.docker.com/repository/docker/chandra459/wine-predictor/general

Launch EC2 Instances:

- **Instance Type**: Select an instance type like t2.large or m5.large for sufficient resources.
- **Ensure instances are in the same VPC for network connectivity.**

Environment Setup (install in all 4 ec2 instances)

Python Dependencies

- PySpark
- NumPy
- Pandas

SSH into the Instances
SSH into each EC2 instance using:
bash
Copy code
ssh -i "your-key.pem" ubuntu@<instance-public-ip>

Passphrase-less SSH
Generate a pair of authentication keys on each instance using:
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

Append each instance's public key(id_rsa.pub) to other instances' authorized_keys
cat ~/.ssh/id_rsa.pub
nano ~/.ssh/authorized_keys

Correct Command to Install OpenJDK 17
bash
Copy code
sudo apt update && sudo apt upgrade -y
sudo apt install openjdk-17-jdk wget unzip -y
java -version # Verify Java installation

Install Hadoop

Download Hadoop

bash

Copy code

```
wget https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
tar -xvzf hadoop-3.3.6.tar.gz
sudo mv hadoop-3.3.6 /usr/local/hadoop
```

Configure Hadoop Environment Variables

Add the following lines to your ~/.bashrc file:

bash

Copy code

```
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

Reload the environment variables:

bash

Copy code

```
source ~/.bashrc
```

Configure Spark

Edit Spark's conf/spark-env.sh:

bash

Copy code

```
cp $SPARK_HOME/conf/spark-env.sh.template $SPARK_HOME/conf/spark-env.sh
nano $SPARK_HOME/conf/spark-env.sh
```

Add:

bash

Copy code

```
export SPARK_MASTER_HOST=<master-node-private-ip>
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop
```

Configure Worker Nodes

Edit conf/slaves on the master node and add the private IPs of all worker nodes:

bash

Copy code

```
<worker-node-1-private-ip>
<worker-node-2-private-ip>
```

Start Spark
Start the Spark master:
bash
Copy code
start-master.sh

Start Spark workers on all worker nodes:
bash
Copy code
start-slave.sh spark://<master-node-private-ip>:7077

master:

Worker1:

Worker2:

Worker3:

Upload dataset to each instance using below cmd:
scp -i "your-key.pem" file-to-upload ubuntu@<instance-public-ip>:<remote-path>

Train the model:
bash
Copy
 # On master node
Python3 wine_training.py

Build and run Docker container for prediction:
bash
Copy
 # Build Docker image
docker build -t wine-predictor.py .

 # Run prediction

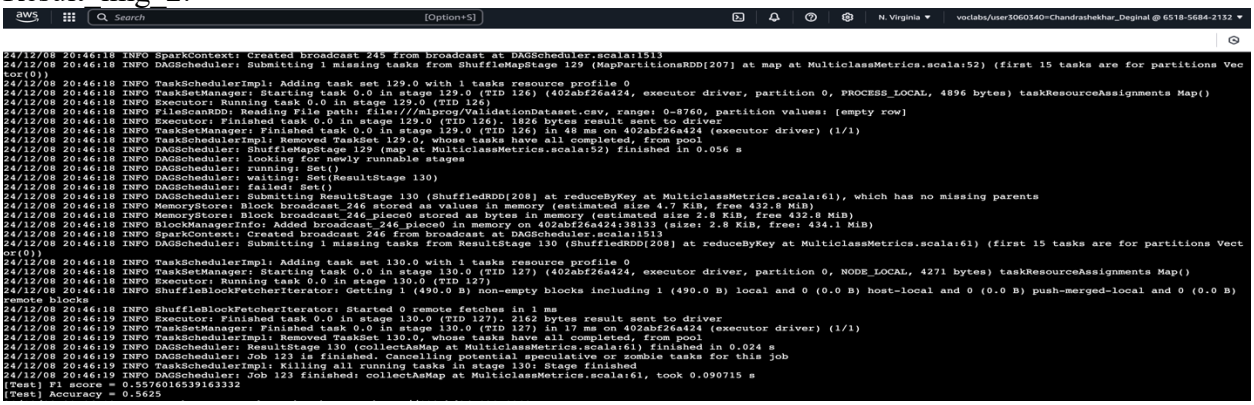
docker run – wine-predictor.py .

Docker_Built_Successfull_Image:

Wine_Predictor_Results:

Result_img_1:

Result_img_2:



Docker push

docker push <dockerhub-username>/wine-predictor:latest