Docker Hub Link:

<https://hub.docker.com/r/ap2876/wine-prediction/tags>

Github Link:

<https://github.com/ap2876-Ashutosh/cs643-PG2.git>

Instructions for Use:

1. \*\*Create Spark Cluster in AWS:\*\*

- Create a Spark cluster using the EMR console provided by AWS. Follow the steps to create one with 4 EC2 instances (increase instances based on workload).

- Generate a Key-Pair for the EMR cluster using the navigation `EC2 -> Network & Security -> Key-pairs`. Download the `.ppk` file.

- Navigate to the Amazon EMR console and create a cluster:

- Fill in the required sections:

- General Configuration -> Cluster Name

- Software Configuration -> Choose EMR 5.33 with Spark 2.4.7

- Hardware Configuration -> Set instance count to 4

- Security Access -> Provide the `.ppk` key created earlier.

- The cluster status should be 'Waiting' upon successful creation.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

2. \*\*Train ML Model in Spark Cluster with 4 EC2 Instances in Parallel:\*\*

- Once the cluster is ready, SSH into the master node using putty or WINSCP:

A computer screen shot of a computer

Description automatically generated

- Change to the root user:

sudo su

- Submit the job:

spark-submit ap2876-wine-train.py

- Track the job status in the EMR UI application logs. Upon successful completion, a `test.model` will be created in the S3 bucket (s3://tanvisbucket).

3. \*\*Run ML Model using Docker:\*\*

- Install Docker.

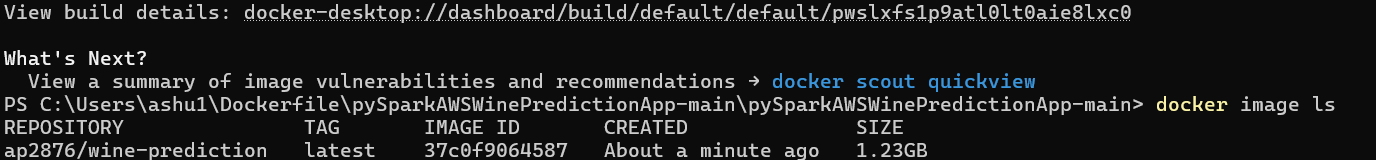
- Build the Docker image using the following command:

docker build -t ap2876/wine-prediction .

A screen shot of a computer

Description automatically generated

- View the image: docker image ls



- Push the image to Docker Hub repository:

docker push ap2876/wine-prediction:latest

A screen shot of a computer

Description automatically generated

- Pull the public image from DockerHub:

docker pull ap2876/wine-prediction:latest

A screen shot of a computer

Description automatically generated

- Place your test data file in a folder and mount it with the Docker container. Run the following command:

docker run -v /home/ec2-user/code/data/csv/:code/data/csv ap2876/wine-prediction testdata.csv

**Task2: Single machine prediction application**

2.1 Launch an ec2 instance.

A screenshot of a computer

Description automatically generated

2.2 Connect to Ec2 instance using putty

2.3 Install scala.

* wget http://downloads.typesafe.com/scala/2.11.6/scala-2.11.6.tgz
* tar -xzvf scala-2.11.6.tgz
* 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

A computer screen with text on it

Description automatically generated

* 1. Install Spark

1. wget <https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz>
2. sudo tar xvf spark-2.4.5-bin-hadoop2.7.tgz -C /opt
3. sudo chown -R ec2-user:ec2-user /opt/spark-2.4.5-bin-hadoop2.7
4. sudo ln -fs spark-2.4.5-bin-hadoop2.7 /opt/spark
5. 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

2.5 Check java –version

* 1. Install Java : sudo yum install -y jdk-8u141-linux-x64.rpm
  2. sudo yum install java-1.8.0
  3. Install Python

1. sudo yum install python
2. pip: sudo yum -y install python-pip
3. Install dependent Libraries: pip install numpy, pip install pandas,pip install sciki-learn, pip install py4j, pip install findspark.
   1. Crete python file and run the job using command: spark-submit ap2876-wine-train.py