Name : Keval Kavle
Assignment : Programming Assignment-2
Subject : Cloud Computing CS643

Docker Image :
https://hub.docker.com/repository/docker/kkdockernet/wine-quality-prediction

Github Link to Code :
https://github.com/KKGITHUBNET/Cloud-Computing

➢ Installation process for cloud setup and running application without docker :

1. **Create EMR cluster on AWS using following configurations.**

```
EMR version 5.20.0

In advance option :
          1.1 Check spark, hadoop and uncheck others
          1.2 Paste below code in configurations :
                [
                  {
                     "Classification": "spark-env",
                     "Configurations": [
                       {
                         "Classification": "export",
                         "Properties": {
                            "PYSPARK_PYTHON": "/usr/bin/python3"
                          }
                       }
                    ]
                  }
                ]
```

2. **EMR Master configurations :**
   Add following in bound traffic rule :
   Protocol : ssh port : 22

3. **Copy code from github on EMR master using ssh client(winscp / filezilla)**

   TrainingDataset.csv
   traintest.py
   dockerfile

   Add TestDataset.csv file to this EMR(for training use ValidationDataset.csv)

4. **EMR setup :**
   Run following commands on EMR master.

   sudo yum update –y

   sudo nano /etc/sudoers
   Update this line -> Defaults secure_path = /sbin:/bin:/usr/sbin:/usr/bin
   add “/usr/local/bin” to this PATH

```
# Install pip, python, and other libraries

sudo easy_install pip
sudo pip install --upgrade pip
sudo pip install wheel
sudo pip install findspark
sudo pip install pyspark
sudo pip install numpy



##### Following commands were run on ec2 – not needed on EMR #####
 sudo yum install java-1.8.0
 export JAVA_HOME=/etc/alternatives/jre
 #JAVA_HOME=/etc/alternatives/jre
 pwd
 export PATH=$PATH:<pwd>
```

5. Run Program :

```
# Program for validation dataset  (training)
    >  python train.py
# Program to run for testdataset  (predictions)
    >  python traintest.py
```

- **<u>Docker setup and running application with docker :</u>**

1. Run following command on EC2(master) or EMR(master) :

1.1 Install docker and start docker service

```
sudo yum install -y docker
sudo service docker start
```

1.2 Following commands are run for creating and pushing docker image to dockerhub (not to be run for running the application) :

```
sudo docker login -u kkdockernet
sudo docker build . -f dockerfile -t kkdockernet/wine-quality-prediction
sudo docker run -t kkdockernet/wine-quality-prediction
sudo docker push  kkdockernet/wine-quality-prediction
```

1.3 To run from anywhere(linux or docker desktop) where docker is installed :

```
sudo docker pull kkdockernet/wine-quality-prediction
```

 # Please check following guideline to run the docker image :
 # sudo docker run –v <Testfile> -t <DockerImage> <path for testfile (only path and not the name.)>

```
sudo docker run -v TestDataset.csv -t kkdockernet/dockertrial ./
```

Following screenshot demonstrates running the docker image on ec2(not on emr.) This is only for reference. Docker image could be run anywhere where docker is installed.(EMR/ linux machine etc.)