CS643 Programming Assignment -2
Wine-Quality-Prediction

- Goal

  The purpose of this individual assignment is to learn how to develop parallel machine learning (ML) applications in Amazon AWS cloud platform. Specifically, you will learn:

  1. How to use Apache Spark to train an ML model in parallel on multiple EC2 instances.
  2. How to use Spark's MLlib to develop and use an ML model in the cloud.
  3. How to use Docker to create a container for your ML model to simplify model deployment.

- Link to the Code

  Wine Quality Prediction Github Link

- Link to container in DockerHub

  Wine Quality Prediction Docker Link

- Step by Step Instruction on how to set AWS EMR Cluster

1. How to create Spark cluster in AWS EMR cluster.
   - Go to AWS account.
   - Create key-pair for EMR cluster.Download .ppk file for Windows. Also Add SSH to EC2 instance.
   - Go to EMR cluster.
   - Create new cluster.
   - Add Spark option to Software configuration.
   - Add 4 number of instance to Hardware configuration.
   - In security and access tab, add your newly created key-pair.
   - It will create new cluster. Wait for 5-10 minutes to fully functional cluster.

2. Create Key-pair security key for EMR cluster (.ppk). Download the .ppk file.

3. Open Putty application on Windows. Login with IP address and Key-pair(.ppk).

4. Open WinSCP application on Windows. Login with IP address and Key-pair(.ppk).

5.Copy your ValidationDataset.csv, TrainingDataset.csv and winequalityprediction.jar file to WinSCP server path `/home/hadoop`.

6. Run the following commands in Putty: -
- Run command `hadoop fs -put * .` this will add you current file to server.
-Run command
`spark-submit --deploy-mode cluster –class com.applipred.WineQualityPrediction.ApplicationPredictionModel windqualityprediction.jar hdfs://ip-172-31-85-193.ec2.internal:8020/user/hadoop/TrainingDataset.csv hdfs://ip-172-31-85-193.ec2.internal:8020/user/hadoop/ValidationDataset.csv hdfs://ip-172-31-85-193.ec2.internal:8020/user/hadoop/finaldata/`

this will deploy all the files with spark and creating final data directory for the output. This will run ApplicationPredictionModel file. You will use this dataset to train the model in parallel on multiple EC2 instances.
- Run command `hadoop fs -ls` this will bind the data set with spark and Log factory.
- From EMR cluster open "Spark History server". Here there is 2 applications are deployed to the cluster with your App Names.
- Run command
`spark-submit --class com.applipred.WineQualityPrediction.PredictionModel windqualityprediction.jar
hdfs://ip-172-31-85-193.ec2.internal:8020/user/hadoop/ValidationDataset.csv
hdfs://ip-172-31-85-193.ec2.internal:8020/user/hadoop/finaldata/`
this will deploy the prediction data set and will create performace metric with Validation dataset. You will use this dataset to validate the model and optimize its performance.

7. Set up Docker and Run prediction using Spark cluster: -
   - Install Docker with container.
   - Create Docker Image. Run command `docker pull yash290397/winequalityprediction`.
   - Place your Test data set csv file to the folder which is associate with docker container.
   - Run command
   `docker run -v {directory path for data set}:/code/data/csv yash290397/winequalityprediction {test data set csv name} `
   -Example: -
   `docker run -v /Users/mjpatel/Document/windqualityprediction/data/csv:/code/data/csv yash290397/winequalityprediction ValidationDataset.csv`