**CS 643 102, Cloud Computing - Spring ‘20**

**Programming Assignment 2**

**NEEL PATEL (** [**nap73@njit.edu**](mailto:nap73@njit.edu) **)**

**Prof. Borcea Christian**

**Aim :**

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.

Description :

I used AWS EMR to create a spark cluster with Hadoop for my application purpose. I used python as my programming language to write the application programs. The input of the model creation application is Training Dataset which is a csv file of Wine Data which I stored in my S3 bucket. I tried several machine learning approaches starting from linear regression, logistic regression, randon forest and decision tree classifier. Experimenting on the above mention approaches , I found out that Random Forest has better accuracy among all of them. Hence I implemented random forest as my selected approach for my model creation application. The output of my model creation application is a trained model name ‘modrf.model’ which is stored in S3. Next stage was to create a prediction application that can take validation dataset as an input and use the previously trained model to predict the output. I fetched both my inputs from my s3 bucket. The output of this application is the accuracy and f1 score of my model.

SETTING ENVIRONMENT FOR MODEL CREATION AND TRAINING

Login into AWS Account

A screenshot of a cell phone

Description automatically generated

Search for EMR in service > Open EMR

A screenshot of a cell phone

Description automatically generated

Click create cluster

A screenshot of a cell phone screen with text

Description automatically generated

Provide required details of configurations : name, select spark as application, ec2 key pair, select number of instances > click create cluster

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Cluster created .. in waiting phase

A screenshot of a social media post

Description automatically generated

Click ssh to get login details for Master node of our EMR cluster > Save it.

In the security and access summary ,, click on the security group for master to change inbound rules for traffic.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Click on master’s security group > add ssh inbound rules and save it.

A screenshot of a cell phone

Description automatically generated

Cluster ready in waiting stage.

A screenshot of a cell phone

Description automatically generated

After successful creation of AWS EMR Spark cluster, we know login into the cluster using terminal.

Command : ssh -i key.pem hadoop@user

A screenshot of a computer screen

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Copy program file to cluster from s3/respective location

Command : aws s3 cp s3://bucket\_name/file\_name

A screenshot of a cell phone

Description automatically generated

Run the program using ‘ spark-submit ./file\_name.py.



Output of the model creation program is a model (modrf in our case ) stored in destination folder in S3.

A screenshot of a social media post

Description automatically generated

ENVIRONMENT SETUP FOR PREDICTION APPLICATION

Login into AWS Account

A screenshot of a cell phone

Description automatically generated

Search for EMR in service > Open EMR

A screenshot of a cell phone

Description automatically generated

Click create cluster

A screenshot of a cell phone screen with text

Description automatically generated

Provide required details of configurations : name, select spark as application, ec2 key pair, select number of instances > click create cluster

A screenshot of a social media post

Description automatically generated

Note : here I have used only 1 as number of instance so that it acts as an ec2 instance with all the configuration of spark, Hadoop etc required for running our PREDICITION application

A screenshot of a cell phone

Description automatically generated

Cluster created .. in starting stage

A screenshot of a cell phone

Description automatically generated

Click ssh to get login details for Master node of our EMR cluster > Save it.

A screenshot of a cell phone

Description automatically generated

After successful creation of AWS EMR Spark cluster, we know login into the cluster using terminal.

Command : ssh -i key.pem hadoop@user

A screenshot of a cell phone

Description automatically generated

Copy prediction file to cluster from s3/respective location

Command : aws s3 cp s3://bucket\_name/file\_name



Run the program using ‘ spark-submit ./file\_name.py.

A screenshot of a cell phone

Description automatically generated

Output of our prediction application is accuracy , recall , precision & f1 score

A screenshot of a cell phone

Description automatically generated

Hence clearly from the above screenshot we can see that the accuracy of our model is 56.87 % for the given validation dataset .The f1 score are 0.575

GITHUB LINK TO MY PROGRAM FILES :

<https://github.com/nap73/cs643>

LEARNING OUTCOMES OF THE PROJECT :

Setting up aws emr cluster with spark and Hadoop.

Programming in python using ML libraries for spark cluster.

Using github.

THANK YOU !

I tried my best to set up docker container as a part my deployment step but couldn’t succeed in it. I would surely try my hands on it coming days as I found it so much interesting. I thank the professor for giving such an interesting project work as a part of my course and also the ta for guiding me throughout the development process.