**New Jersey Institute of Technology**

Fall 22 CS 643

Programming Assignment -II

Professor: Zunnurhain, Kazi (PhD)

Students:

Muhammad Imran Burhanullah (mib3@njit.edu)

**OBJECTIVE:**

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.

**STEP-BY-STEP HOW-TO SETUP THE CLOUD ENVIORNMENT AND RUNNING THE MODEL TRAINING AND THE APPLICATION PREDICTION:**

I used Amazon EMR (Amazon Elastic MapReduce) to create a Spark cluster with Hadoop for my application purpose. First of all, log into Amazon Web Services (AWS) Account

Graphical user interface, application

Description automatically generated

**Screenshot for Elastic Map Reduce > Opening EMR**

**Graphical user interface, text, application, email

Description automatically generated**

Creating Cluster with required details of Configurations: name, select, Spark as

Application, ec2 key pair, select number of instances > click create cluster

Graphical user interface, text, application, email

Description automatically generated

Note I have only as number of instance so it act as a EC2 instance with all configuration of

Spark required for running our prediction application.

Graphical user interface, text, application

Description automatically generated

I have also upload the 2 files in S3 bucket names [s3://[wineprgpredict](https://s3.console.aws.amazon.com/s3/buckets/wineprgpredict?region=us-east-1)](https://s3.console.aws.amazon.com/s3/buckets/myawss32022?region=us-east-1)

Graphical user interface, text, application

Description automatically generated

Cluster starting ….

Graphical user interface, text, application, email

Description automatically generated

Command : **hadoop@ec2-3-87-237-159.compute-1.amazonaws.com**

Graphical user interface, text

Description automatically generated

Text

Description automatically generated with medium confidence

Run pip install findspark

Run pip install pandas

Run pip install scikit-learn

Run spark-submit prgwinetrain.py

Run pip install sklearn

Table

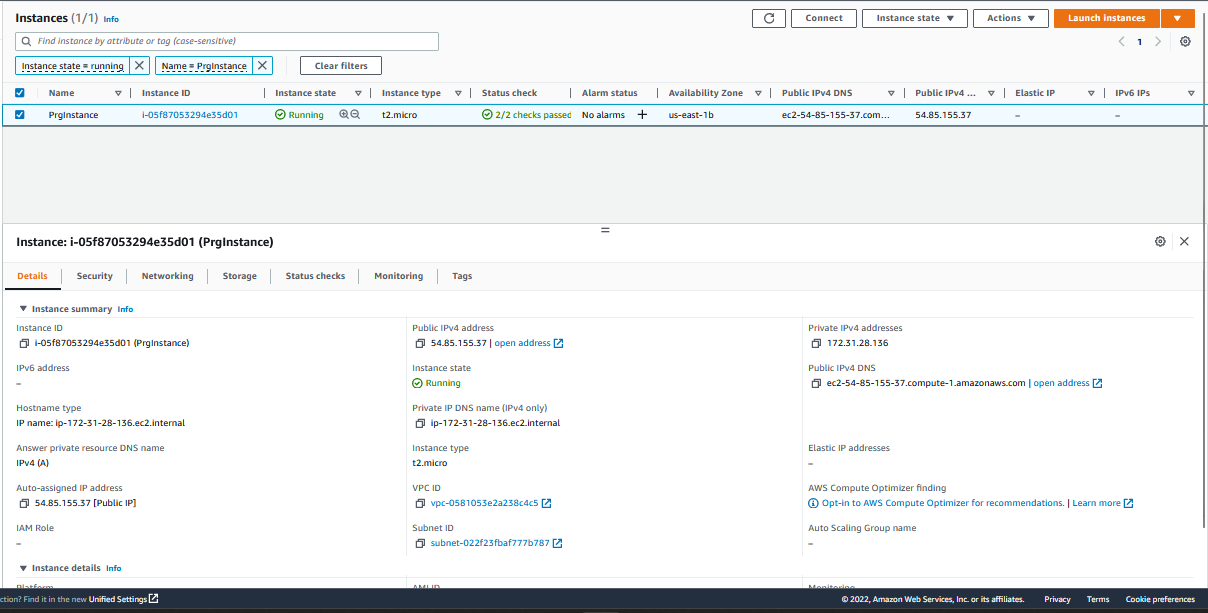
Description automatically generated

Text

Description automatically generated

**Task2:**

Launch an ec2 instance.



2.2 ssh to new instance from your local terminal

Text

Description automatically generated

2.3Installing 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:
  + - 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

Text

Description automatically generated

Text

Description automatically generated

**Task 3: Prediction by using docker image.**

Install docker in ec2.

3.1 Install docker

$sudo yum install -y docker

$ sudo service docker start

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

#mkdir predict

#mkdir -p dataset

#nano test.py

#docker build -t predict .

Text

Description automatically generated