Prerequisite:

Create an s3 bucket and upload the training and validation data set.

A screenshot of a computer

Description automatically generated

**Task1: Parallel training on 4 ec2 Instances**

* 1. Lauch an EMR cluster.

General Configuration:

Launch Mode : Cluster

Software Configuration:

Applications: Spark: Spark 2.4.8 on Hadoop 2.10.1 YARN and Zeppelin 0.10.0

Hardware configuration

Number of instances:4

Security and access

EC2 key pair: Select an existing Key pair or create a new pair if you don’t have one

Change a new rule in inbound rules of master node’s security group and add new rule with SSH and custom TC as the masters public IP address in order to enable sshing.

Graphical user interface, text, application, email

Description automatically generated

* 1. Copy the IP and login to master from terminal

Text

Description automatically generated

* 1. Create a py new file

$ nano winequalitytrain.py

Write your code ,save and close

Install all necessary libraries using pip install

Eg: pip install findspark

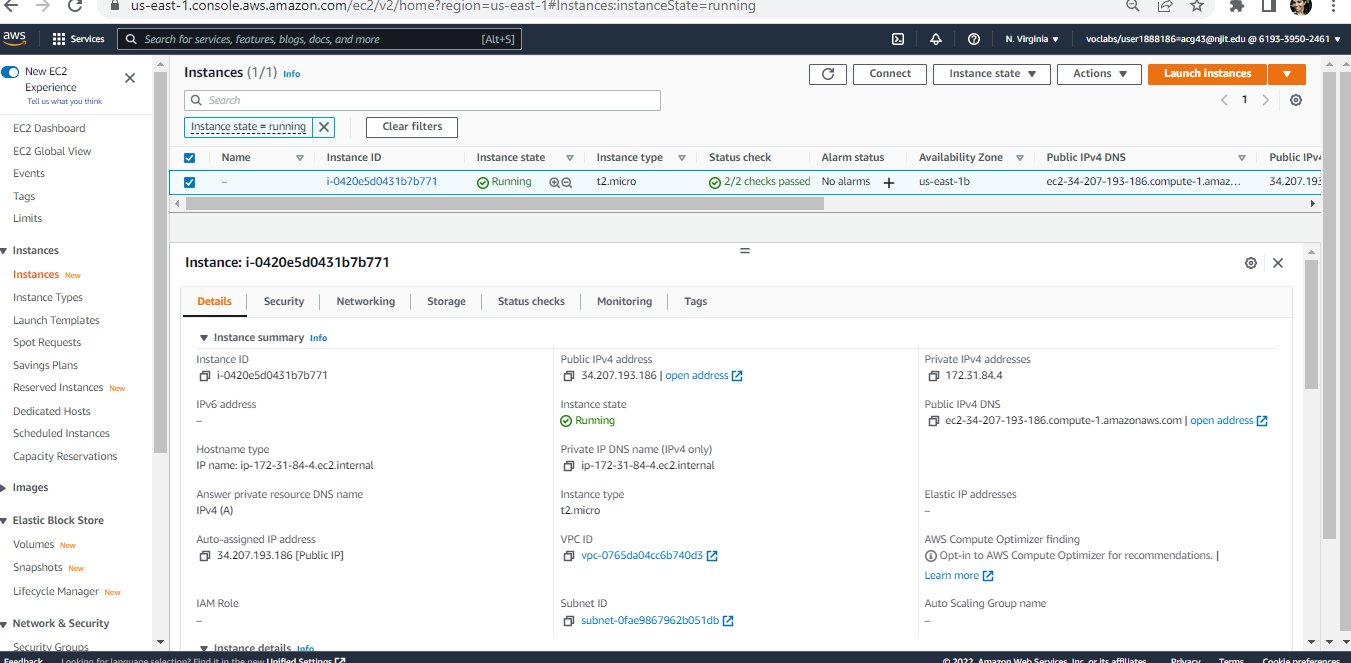
* 1. Run your code using ‘spark-submit winequalitytrain.py
  2. Go to application interface and select spark history server and check status of your job.

Graphical user interface, table

Description automatically generated

**Task2: Single machine prediction application**

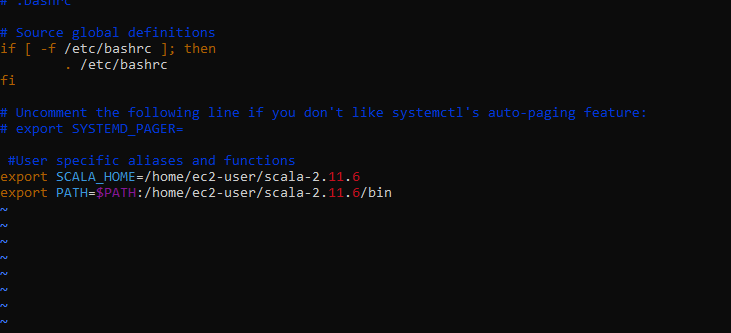
2.1 Launch an ec2 instance.



2.2 ssh to new instance from your local terminal

2.3Install 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



* 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

2.5Check java –version

Install if not its not there. Steps in below link

https://techviewleo.com/install-java-openjdk-on-amazon-linux-system/

2.5 check python version

If not available install python

2.6Install all necessary libraries using pip install. If pip fails run below command

$ sudo yum -y install python-pip

2.7 Open a new .py file with command nano winequalitypredict.py

Insert python code.

2.8 spark-submit winepred.py run your code in ec2.

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

Install docker in ec2.

3.1 Install docker

$sudo amazon-linux-extras install docker -y

$ sudo service docker start

$ sudo chmod 666 /var/run/docker.sock

$docker login

$ docker pull garvit007/wine-qualitytrain:tag

$ docker run -p 4000:80 garvit007/wine-qualitytrain:tag