**Cloud Computing Programming Assignment 2**

**Wine Quality Prediction in AWS using EMR Cluster and S3 Bucket using Python Code**

GitHub Repository:

<https://github.com/ar2637/cccs643_pa>2

Docker Hub:

<https://hub.docker.com/repository/docker/ar2637/cccs643_pa2/general>

AWS SETUP :

Let’s Start with setting up AWS environment.

1. Get AWS learner module access and setup account.
2. Start AWS Lab and Launch Console.

EMR Cluster SETUP :

1. Go to services and select “EMR”.
2. For EMR Configuration, select
   1. “EMR 5.36.0 “
   2. Custom (Hadoop 2.10.1, Spark 2.4.8, Zeppelin 0.10.0)
   3. Setup Cluster Configuration Primary as m4.large, core as

m4.large and Tasks as m4.large. Select instance size as 4 for both core and tasks.

1. Create Two such instances.
2. Then Create Cluster. Now wait for the status to go from “starting” to “waiting”.

Network firewall Setup.

1. Go to Security Firewall and select “Primary”.

2. Go to Inbound Network Rules and select Edit option.

3. Create a new rule to allow “SSH” Connection from myIP.

Connecting SSH Connection

1. Create a folder and move python files and key pair file inside folder.
2. Open Terminal and Set the current directory using cd command. Move into the directory with “.pem” and “.py” files.

Command : cd /user/path/folder

1. Check key pair file access. Command : chmod 400 keypair.pem
2. Connect to EMR Cluster using below command.

ssh -i /filepath/keypair.pem hadoop@<<IPaddress of EC2 instance>>

Publishing python files to EMR Instances

1. Hosting the training and prediction python codes in EMR
2. Connect to EC2 using sftp command:

sftp -i "keypair.pem" hadoop@<<IPaddress of EC2 instance>>

Publish python files using put command:

put /filepath/to/your/pythonfile/filename.py

1. Connect to EMR using ssh command and verify the files using Pwd and ls commands

Create S3 Bucket

1. Go to services and select “S3”.
2. Choose a bucket name, and go with default settings
3. Choose Create Bucket
4. Upload both the datasets, training.csv and validation.csv
5. Use this bucket to save the model after running the model through EMR Cluster.

Create Docker

1. docker build -t ar2637/prediction.
2. docker push ar2637/prediction:latest
3. dockerrun -v /home/ec2-user/code/data/csv/:code/data/csv ar2637/wine-prediction validation.csv

Train ML model in Spark cluster with 4 ec2 instances in parallel

1. Now when cluster is ready to accept jobs, to submit one you can either use step button to add steps or submit manually.

To submit manually, Perform SSH to master of cluster using below command: ssh -i "ec2key.pem" <<User>>@<<Public IPv4 DNS>>

1. On successful login to master , change to root user by running command:

sudo su

1. Submit job using following command:

spark-submit s3://wine-data-ar2637-14/prediction.py

1. You can trace status of this job in EMR UI application logs. Once status is succeded a test.model will be created in s3 bucket

Trained ML model locally without docker.

1. Clone this repository.
2. Make sure you have spark environment setup locally for running this. To setup one follow link https://spark.apache.org/docs/latest
3. Navigate to awssparkwineappk/src folder
4. Place your testdata in 'awssparkwineappk/data/csv' folder
5. Install pyspark, you can use pip -m install pyspark. Or install conda package.
6. Once setup is ready execute command

Run ML model using Docker

1.Install docker where you want to run this container

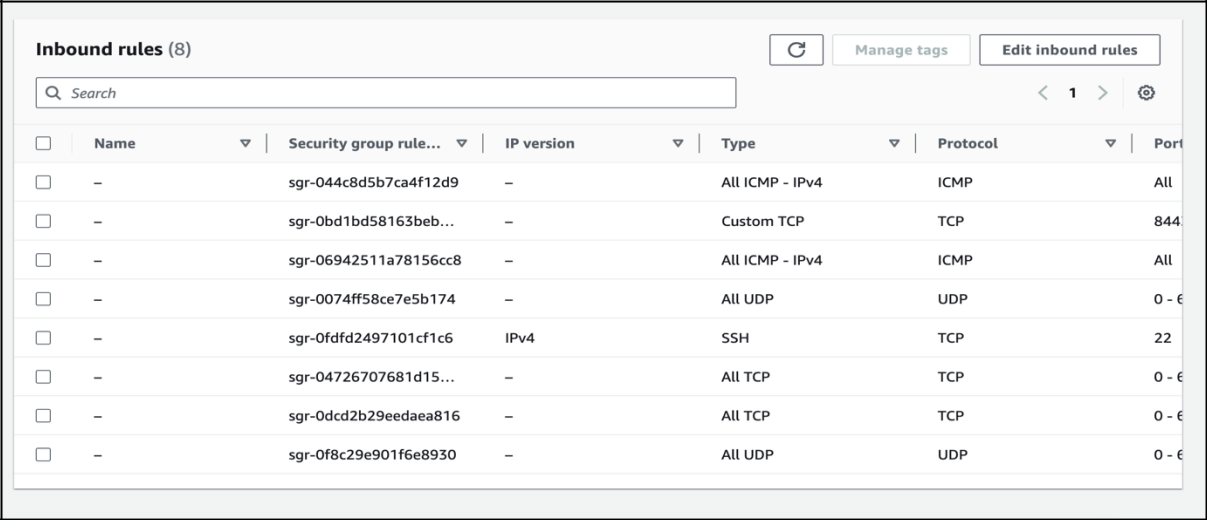
2.A public image has been created and posted on DockerHub. Use the command Docker pull /awssparkwineapp to get the image on your machine.

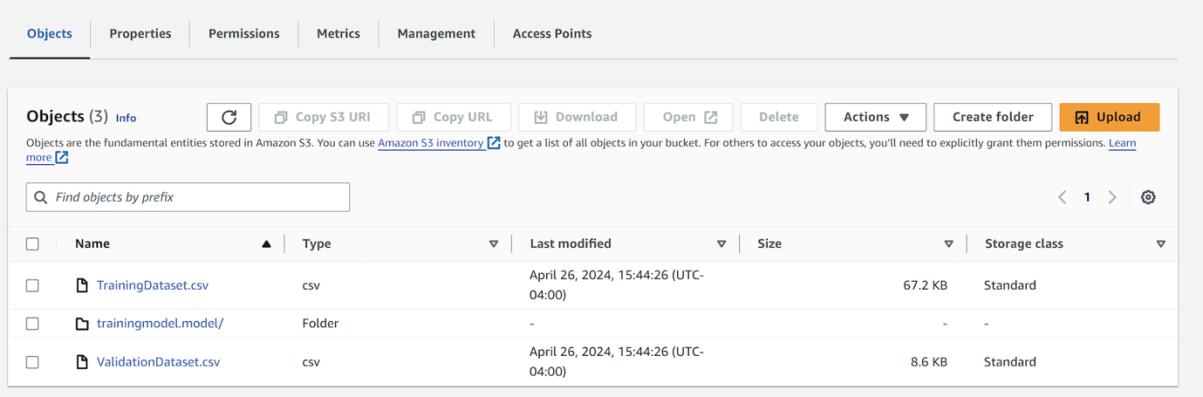
3.Place your test data file in a folder which you will mount with docker container.

4.docker run -v {directory path} code/data/csv /awssparkwineapp {testdata file name}

docker run -v /Users/<username>/<path-to-folder>/awssparkwineapp/data/csv:/code/data/csv/awssparkwineapp

New Inbound Rules for EMR

****

S3 Bucket

DockerHub

