Programming Assignment-2
CS643-Cloud Computing

Submitted By:
Akash Shrivastava
UCID: AS57

Section-1

GitHub URLs:

Parallel training implementation-
https://github.com/akatast/Wine_quality_prediction/blob/main/wineQPredModelTraining.py

Single machine prediction application-
https://github.com/akatast/Wine_quality_prediction/blob/main/wineQPredModelValidation.py

Docker hub URL:

Docker container for prediction application-

https://hub.docker.com/repository/docker/as5721/as57dockerpublic

Command to execute docker container using input file-

```
sudo docker run -it -v `pwd`/TestDataset.csv:/dataset/TestDataset.csv as5721/as57dockerpublic:test-wine-qp /dataset/TestDataset.csv
```

Please make sure docker is started and input file TestDataset.csv is available at the same directory where this command is being submitted.

Section-2

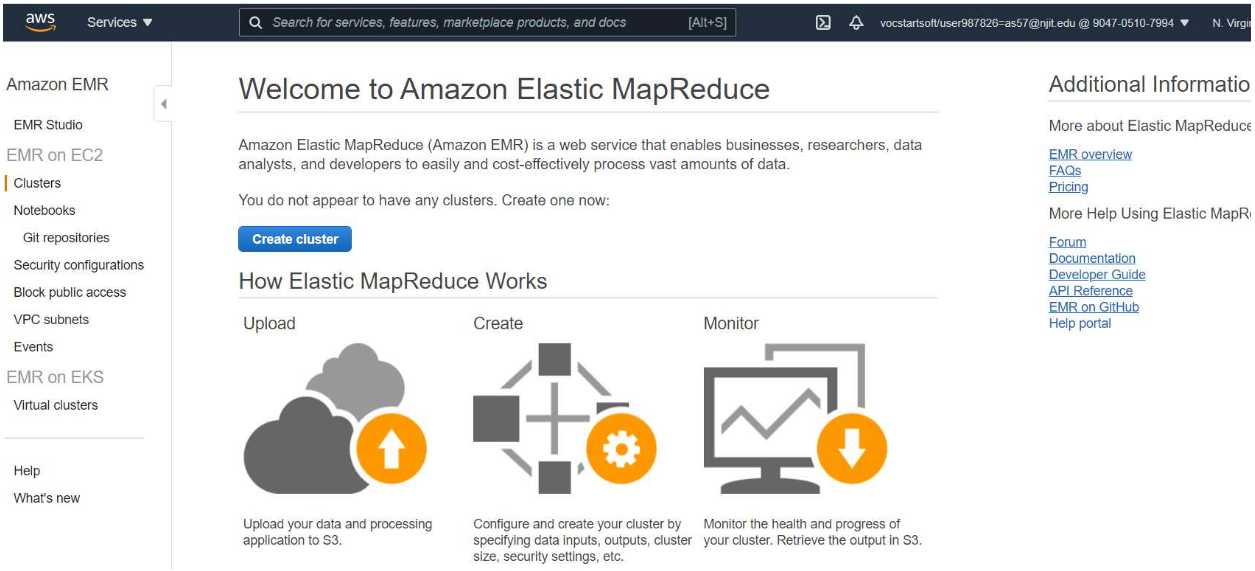
AWS Cloud environment set-up

2.1 AWS educate account:

As very first step, login to AWS educate classroom (vocareum) and click on AWS Console.

2.2 Create EMR cluster:

Step 1: Search for the EMR (Amazon Elastic MapReduce) and click on create cluster:



Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

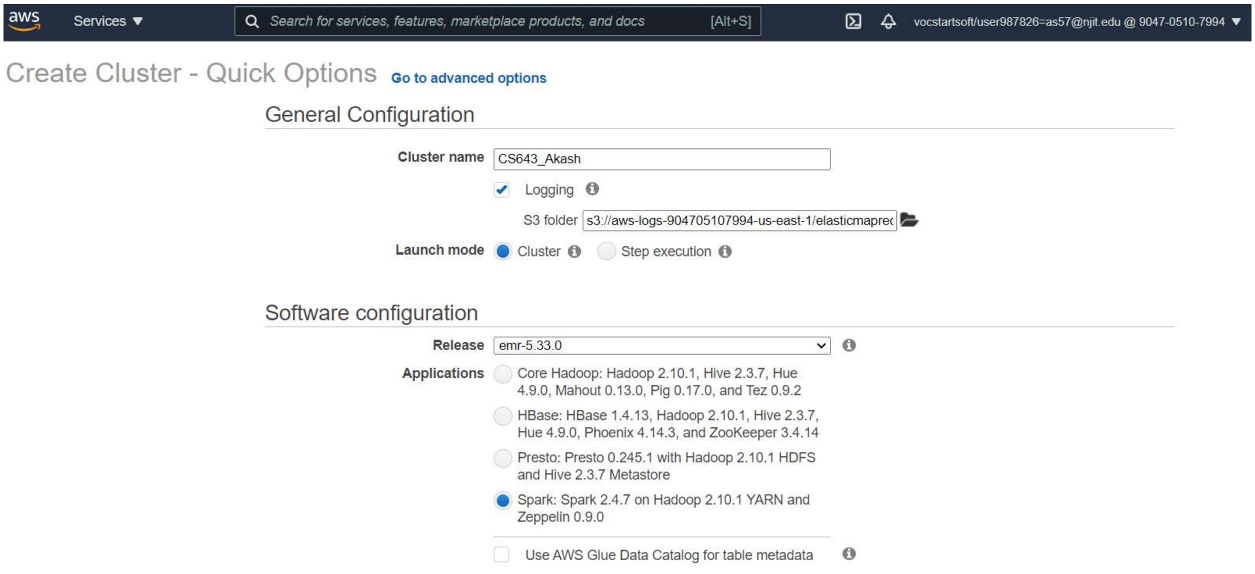
How Elastic MapReduce Works

Upload your data and processing application to S3.

Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.

Monitor the health and progress of your cluster. Retrieve the output in S3.

Step 2: Give the cluster name and select Spark as application:


aws
Services ▼
Search for services, features, marketplace products, and docs [Alt+S]
vocstartsoft/user987826=as57@njit.edu @ 9047-0510-7994 ▼
Go to advanced options

Create Cluster - Quick Options

General Configuration

Cluster name: CS643_Akash

☑ Logging

S3 folder: s3://aws-logs-904705107994-us-east-1/elasticmaprec

Launch mode: ● Cluster ○ Step execution

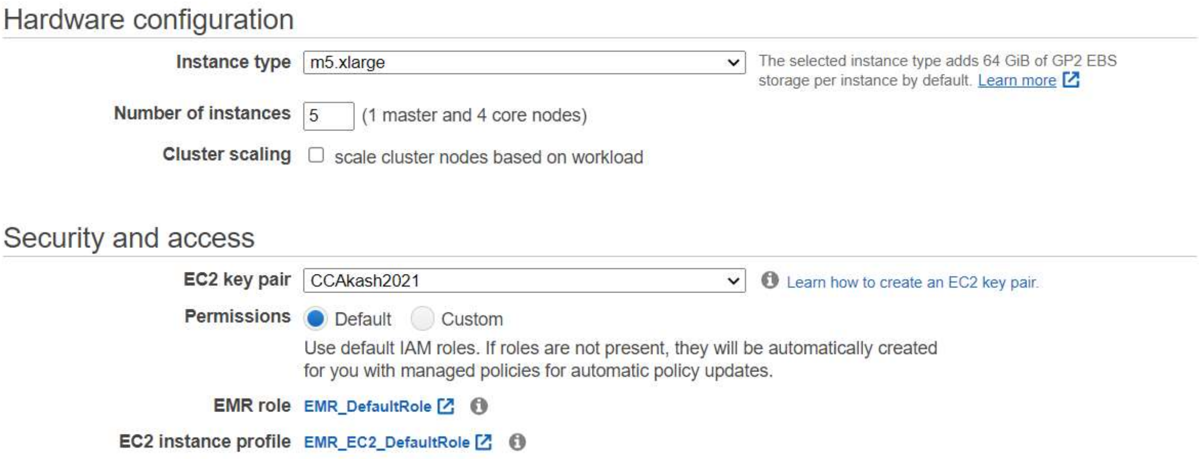
Software configuration

Release: emr-5.33.0

Applications:
- ○ Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ○ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- ○ Presto: Presto 0.245.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore
- ● Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0

□ Use AWS Glue Data Catalog for table metadata

As mentioned in the requirement, use number of instances as 4 (one master and four core) and select the key pair:



Hardware configuration

Instance type: m5.xlarge
The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. Learn more

Number of instances: 5 (1 master and 4 core nodes)

Cluster scaling: □ scale cluster nodes based on workload

Security and access

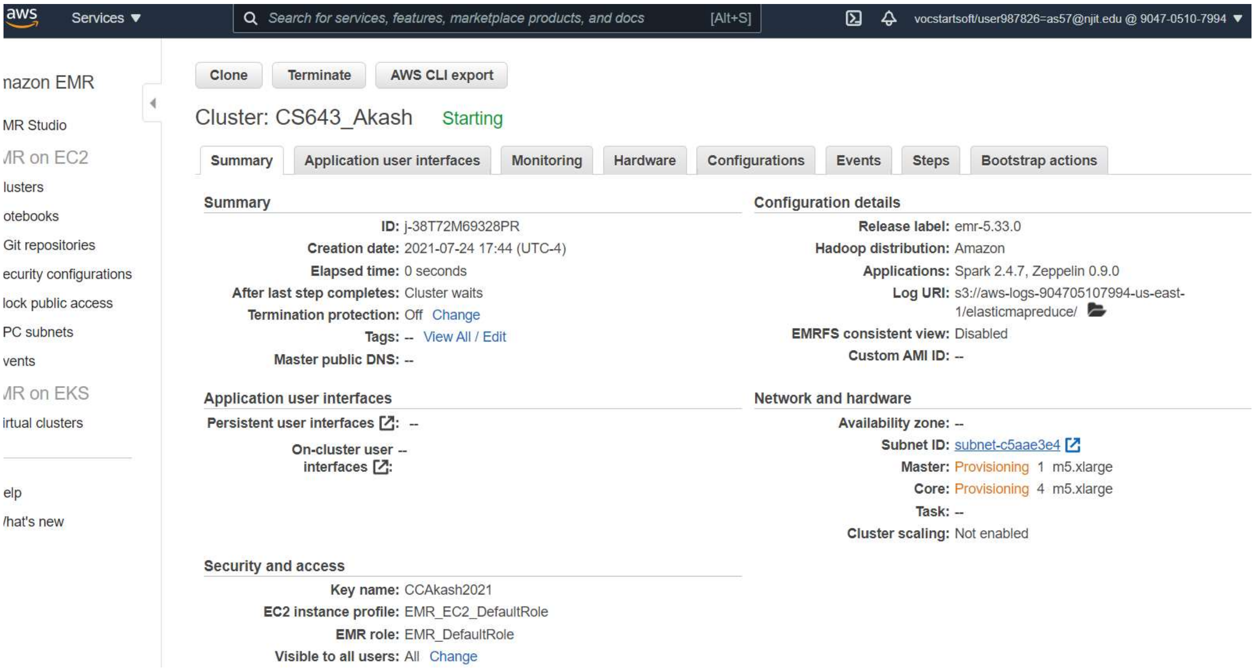
EC2 key pair: CCAkash2021
Learn how to create an EC2 key pair.

Permissions: ● Default ○ Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole

EC2 instance profile: EMR_EC2_DefaultRole

Step 3: Then click on create cluster and you'll get following screen with "Starting" message:



Cluster: CS643_Akash Starting

Summary

ID: j-38T72M69328PR
Creation date: 2021-07-24 17:44 (UTC-4)
Elapsed time: 0 seconds
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: --

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: --

Security and access

Key name: CCAkash2021
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change

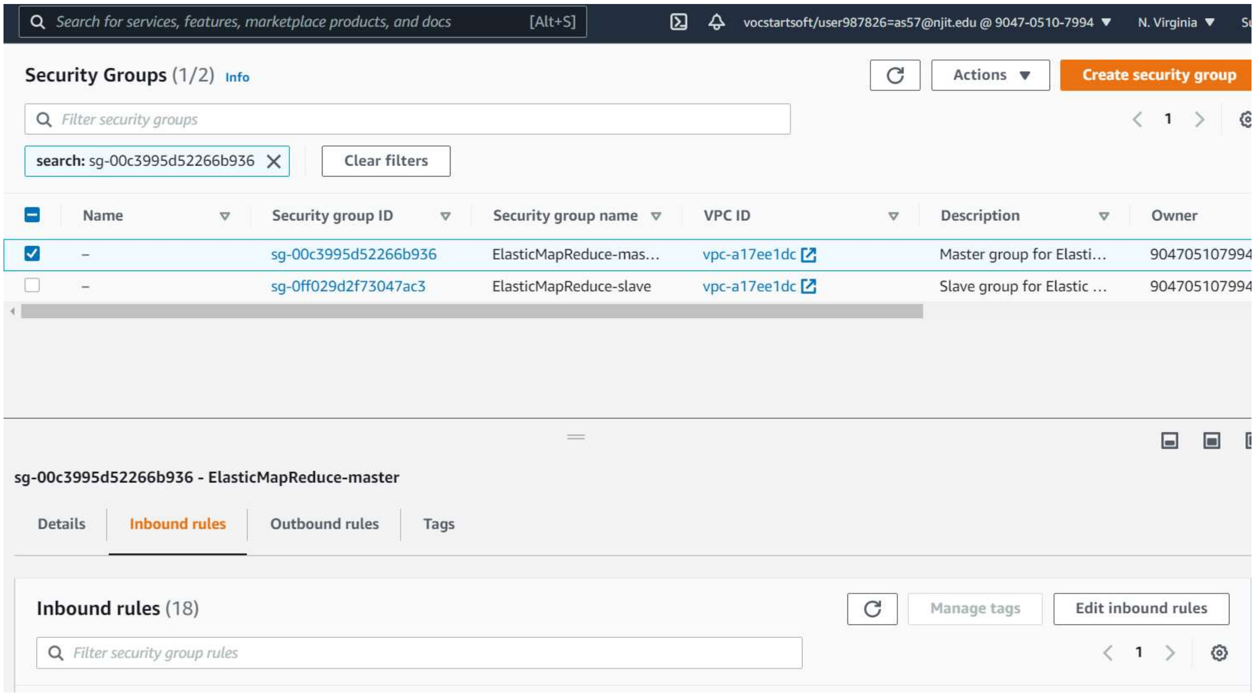
Configuration details

Release label: emr-5.33.0
Hadoop distribution: Amazon
Applications: Spark 2.4.7, Zeppelin 0.9.0
Log URI: s3://aws-logs-904705107994-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: --
Subnet ID: subnet-c5aae3e4
Master: Provisioning 1 m5.xlarge
Core: Provisioning 4 m5.xlarge
Task: --
Cluster scaling: Not enabled

Step 4: Go to security group of master and click on inbound rules:


Search for services, features, marketplace products, and docs
[Alt+S]
vocstartsoft/user987826=as57@njit.edu @ 9047-0510-7994
N. Virginia

Security Groups (1/2) Info

Actions
Create security group
Filter security groups

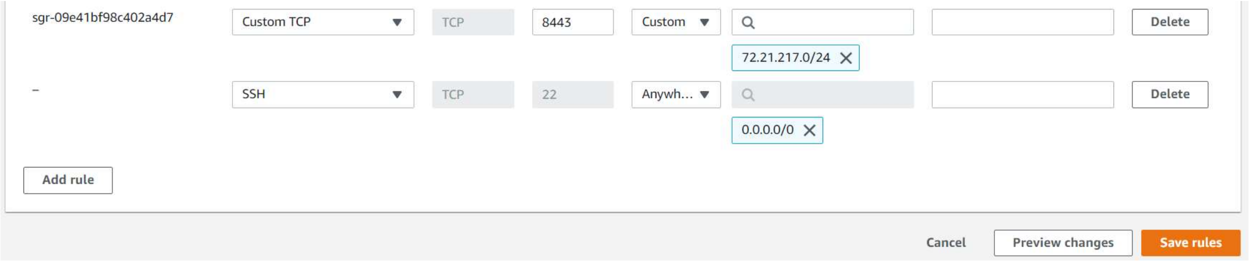
search: sg-00c3995d52266b936

	Name	Security group ID	Security group name	VPC ID	Description	Owner
☑	–	sg-00c3995d52266b936	ElasticMapReduce-mas…	vpc-a17ee1dc	Master group for Elasti…	904705107994
□	–	sg-0ff029d2f73047ac3	ElasticMapReduce-slave	vpc-a17ee1dc	Slave group for Elastic …	904705107994

sg-00c3995d52266b936 - ElasticMapReduce-master

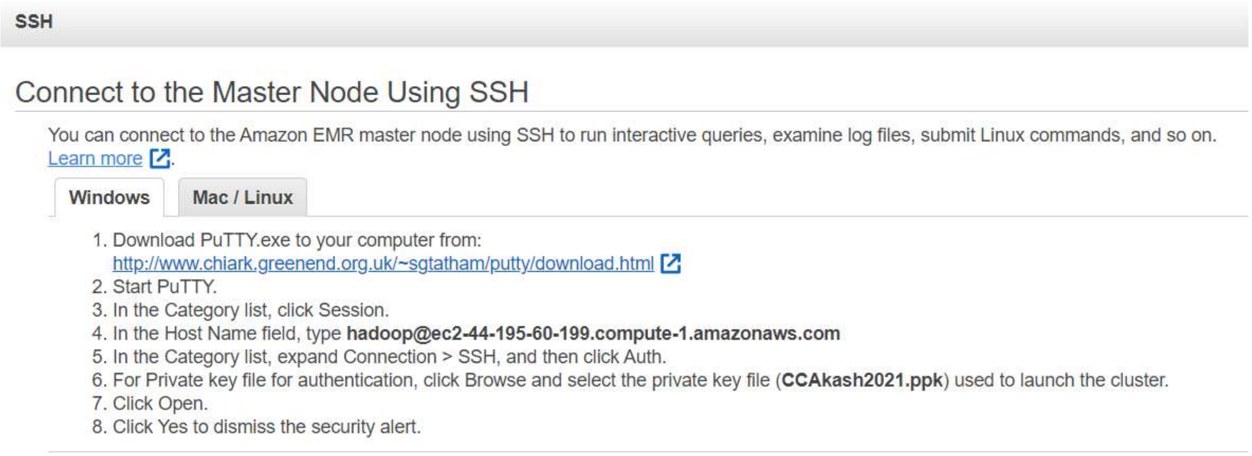
Inbound rules (18)

Click on 'Edit inbound rules' and add a new rule with type = SSH, port range=22 and source = anywhere



sgr-09e41bf98c402a4d7	Custom TCP	TCP	8443	Custom	72.21.217.0/24	Delete
–	SSH	TCP	22	Anywh…	0.0.0.0/0	Delete

Step 5: AWS gives the steps to Connect to the Master Node Using SSH. Use the following steps to connect to the master node:

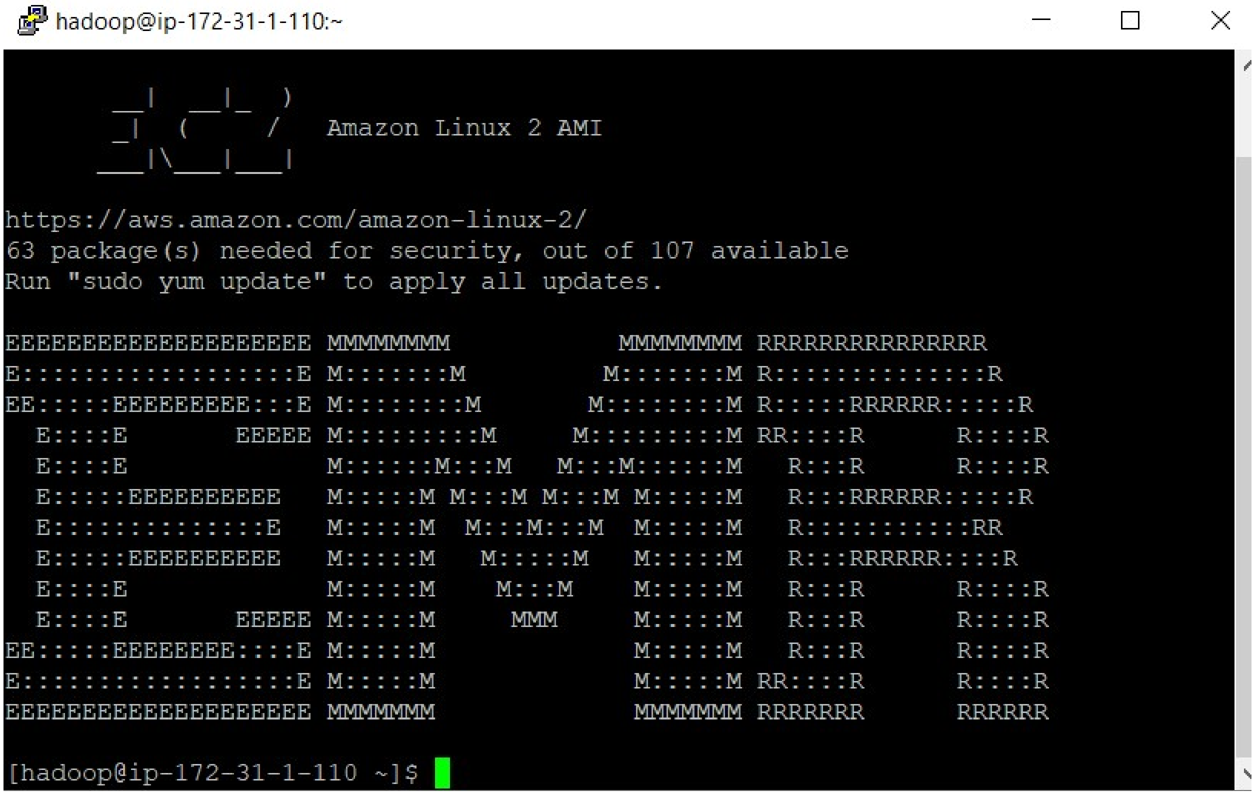


Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. Learn more.

1. Download PuTTY.exe to your computer from: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
2. Start PuTTY.
3. In the Category list, click Session.
4. In the Host Name field, type **hadoop@ec2-44-195-60-199.compute-1.amazonaws.com**
5. In the Category list, expand Connection > SSH, and then click Auth.
6. For Private key file for authentication, click Browse and select the private key file (**CCAkash2021.ppk**) used to launch the cluster.
7. Click Open.
8. Click Yes to dismiss the security alert.

Step 6: After successful connection we'll get following screen:

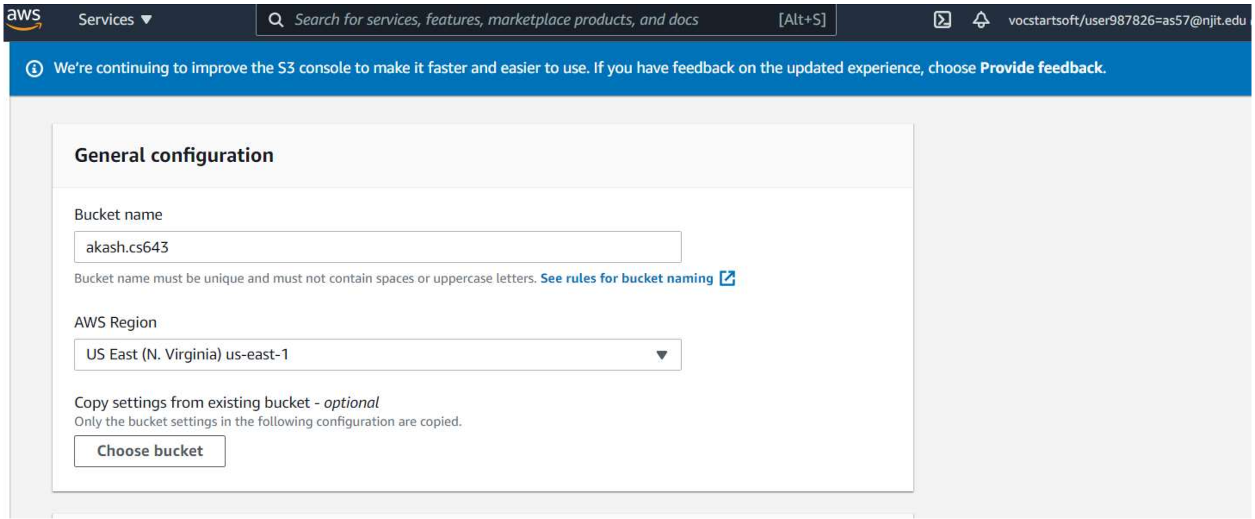


```
       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
63 package(s) needed for security, out of 107 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-110 ~]$
```

Section-3
S3 bucket and file storage

After termination of EMR cluster we lose all the data so we can use S3 bucket to store the files beyond one session/longer-term.

3.1 Create a new S3 bucket:
Create a S3 bucket to store the files-


aws
Services ▼
Search for services, features, marketplace products, and docs
[Alt+S]
vocstartsoft/user987826=as57@njit.edu
Choose bucket

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

General configuration

Bucket name: akash.cs643

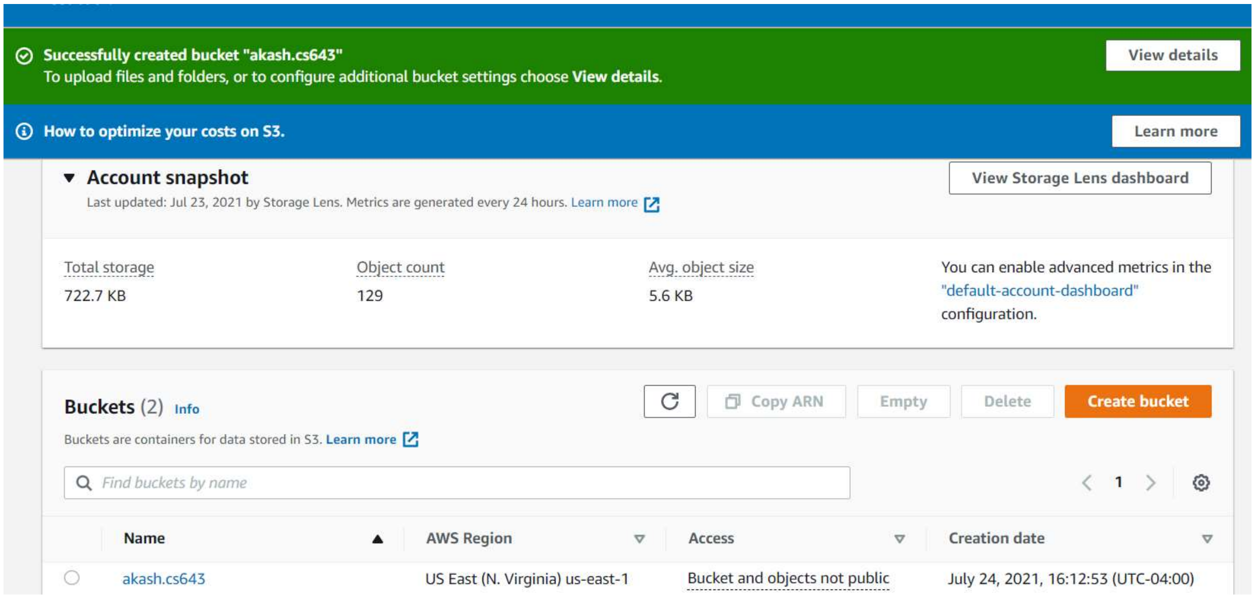
Bucket name must be unique and must not contain spaces or uppercase letters. See rules for bucket naming

AWS Region: US East (N. Virginia) us-east-1

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

We'll get following success message-



Successfully created bucket "akash.cs643"

To upload files and folders, or to configure additional bucket settings choose **View details**.

How to optimize your costs on S3.

Account snapshot

Last updated: Jul 23, 2021 by Storage Lens. Metrics are generated every 24 hours. Learn more

Total storage	Object count	Avg. object size
722.7 KB	129	5.6 KB

You can enable advanced metrics in the "default-account-dashboard" configuration.

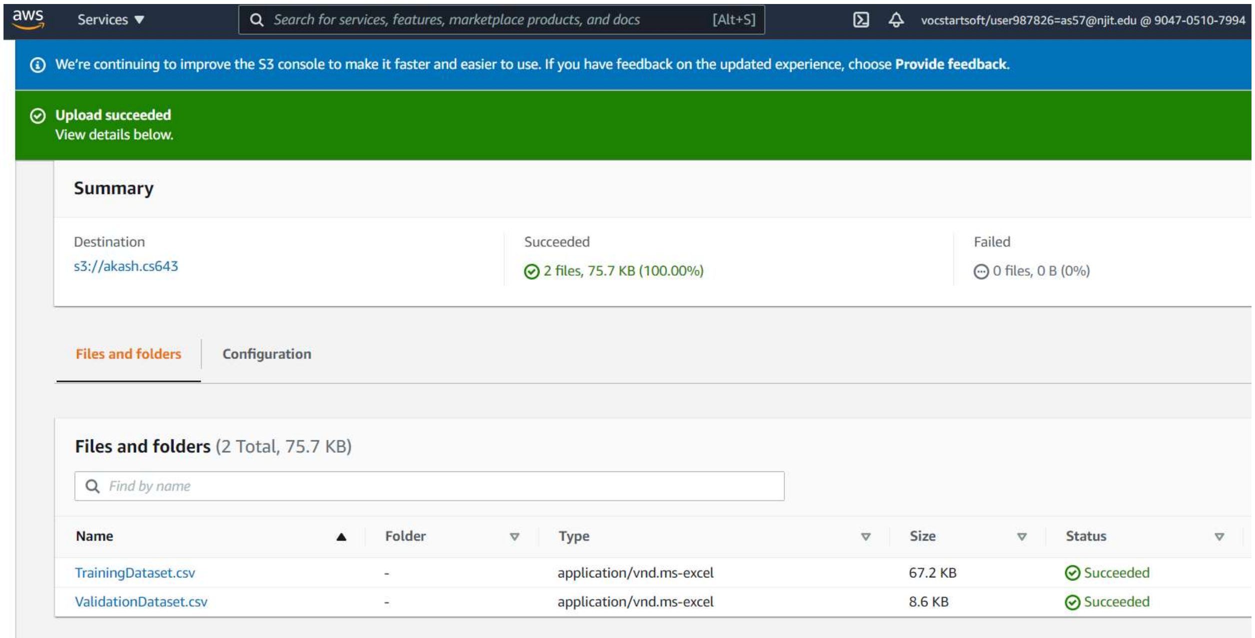
Buckets (2)

Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	akash.cs643	US East (N. Virginia) us-east-1	Bucket and objects not public	July 24, 2021, 16:12:53 (UTC-04:00)

3.2 Upload the files:

Upload the training and validation datasets to S3 bucket-


aws
Services ▼
Search for services, features, marketplace products, and docs
[Alt+S]
vocstartsoft/user987826=as57@njit.edu @ 9047-0510-7994
Files and folders
Configuration
Find by name

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Upload succeeded
View details below.

Summary

Destination	Succeeded	Failed
s3://akash.cs643	2 files, 75.7 KB (100.00%)	0 files, 0 B (0%)

Files and folders (2 Total, 75.7 KB)

Name	Folder	Type	Size	Status
TrainingDataset.csv	-	application/vnd.ms-excel	67.2 KB	Succeeded
ValidationDataset.csv	-	application/vnd.ms-excel	8.6 KB	Succeeded

Section-4

Model training and validation

4.1 Parallel training implementation (Train the model):

Connect to the Master node, move the model training program to Master node and give the execute permission.

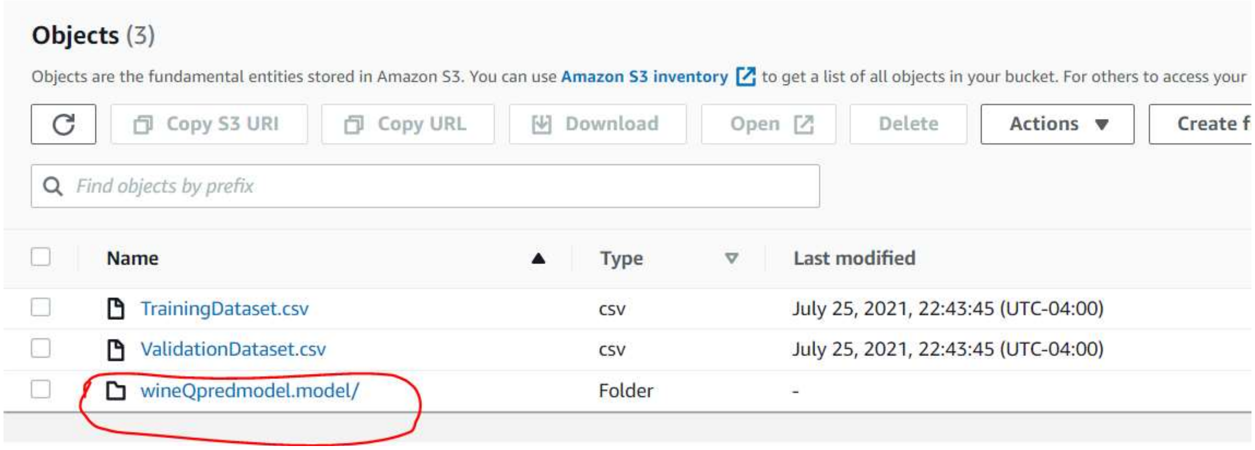
Please check page-1 for GitHub link of Parallel training implementation.

Execute the training program using following command-

```
Spark-submit wineQPredModelValidation.py
```

I am printing training error and F1 score after completion of training.

After successful execution, check the S3 bucket if model has been stored-



Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your

□	Name	Type	Last modified
□	TrainingDataset.csv	csv	July 25, 2021, 22:43:45 (UTC-04:00)
□	ValidationDataset.csv	csv	July 25, 2021, 22:43:45 (UTC-04:00)
□	wineQpredmodel.model/	Folder	-

4.2 Single machine prediction application (without docker):

Please check page-1 for GitHub link of Single machine prediction application.

Execute the validation program:

```
spark-submit wineQPredModelValidation.py
```

Section-5
Create Docker Container

Connect to EC-2 instance of master and perform following steps:

Install spark-

```
pip install --user pyspark
```

Set the path-

```
export PYSPARK_PYTHON=/usr/bin/python3
export PYSPARK_DRIVER_PYTHON=/usr/bin/python3
```

Install and start the Docker-

```
sudo yum update -y
sudo amazon-linux-extras install docker
sudo yum install docker
sudo service docker start
sudo usermod -a -G docker Hadoop
```

Change the permission-

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

Build the image (before executing this command, make sure the trained model, model testing program and the dockerfile are present on same directory)-

```
sudo docker build -t imagename .
```

List the available images and verify-

```
sudo docker image ls
```

Login to docker hub using following command, username, and password-

```
docker login
```

Go to dockerhub website and create a new public repo.

On the EC2 terminal tag the created docker image-

```
docker tag imagename:version username/repo:tagname
```

From EC2 push the image to docker hub repo-

```
docker push username/repo:tagname
```

Now docker container is ready, anyone can issue a pull request and execute it in same OS (Linux) using the input file TestDataset.csv

Docker container for prediction application:

Following is command to execute the container-

```
sudo docker run -it -v `pwd`/TestDataset.csv:/dataset/TestDataset.csv
as5721/as57dockerpublic:test-wine-qp /dataset/TestDataset.csv
```