Programming Assignment 2
Cloud Computing

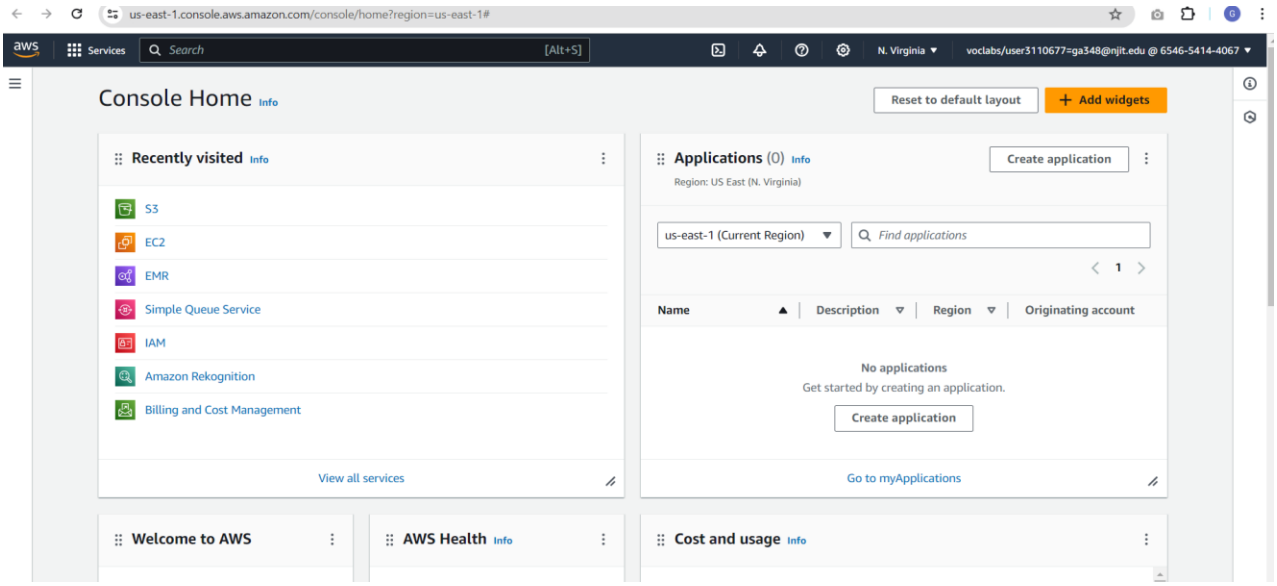
Name – Gayatri Aavula
UCID – ga348@njit.edu

GitHub – https://github.com/gayatriaavula/winepred_quality.git

DockerHub - **https://hub.docker.com/repository/docker/ga348/cs643-programming-assignment-2/**

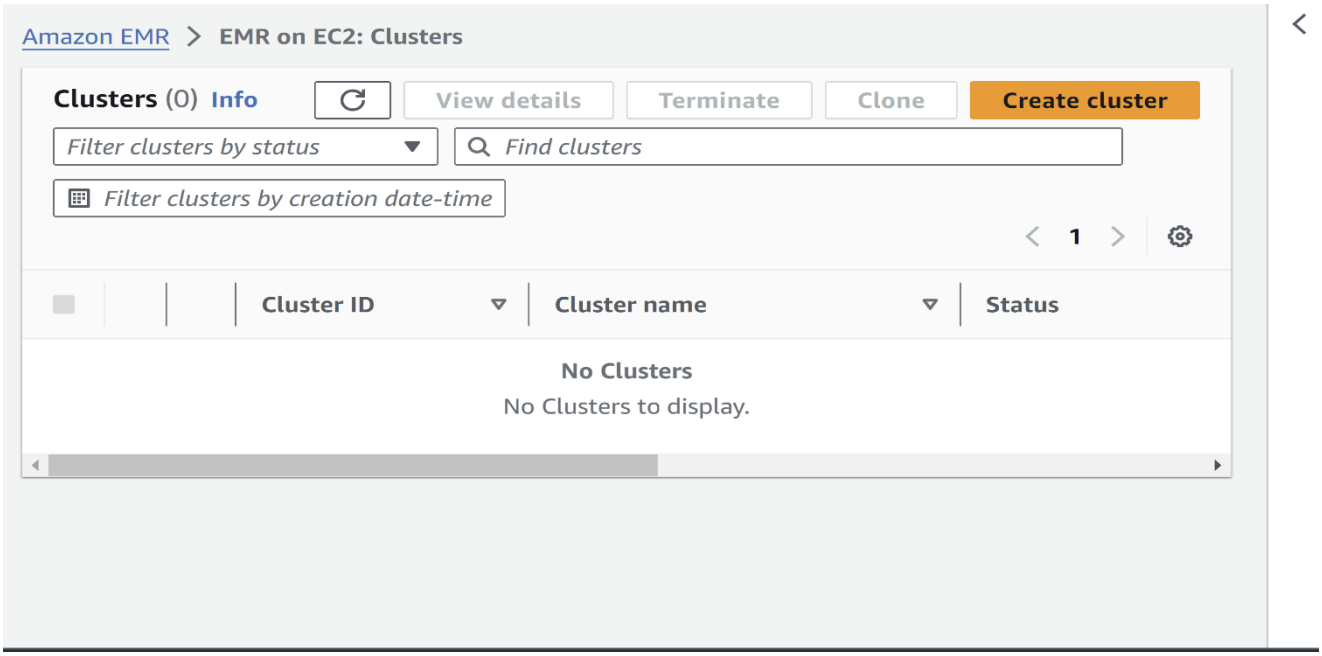
1.Log into the AWS Management Console.

2. the AWS EMR service from the list of available services. Then, choose EMR on EC2 Clusters.

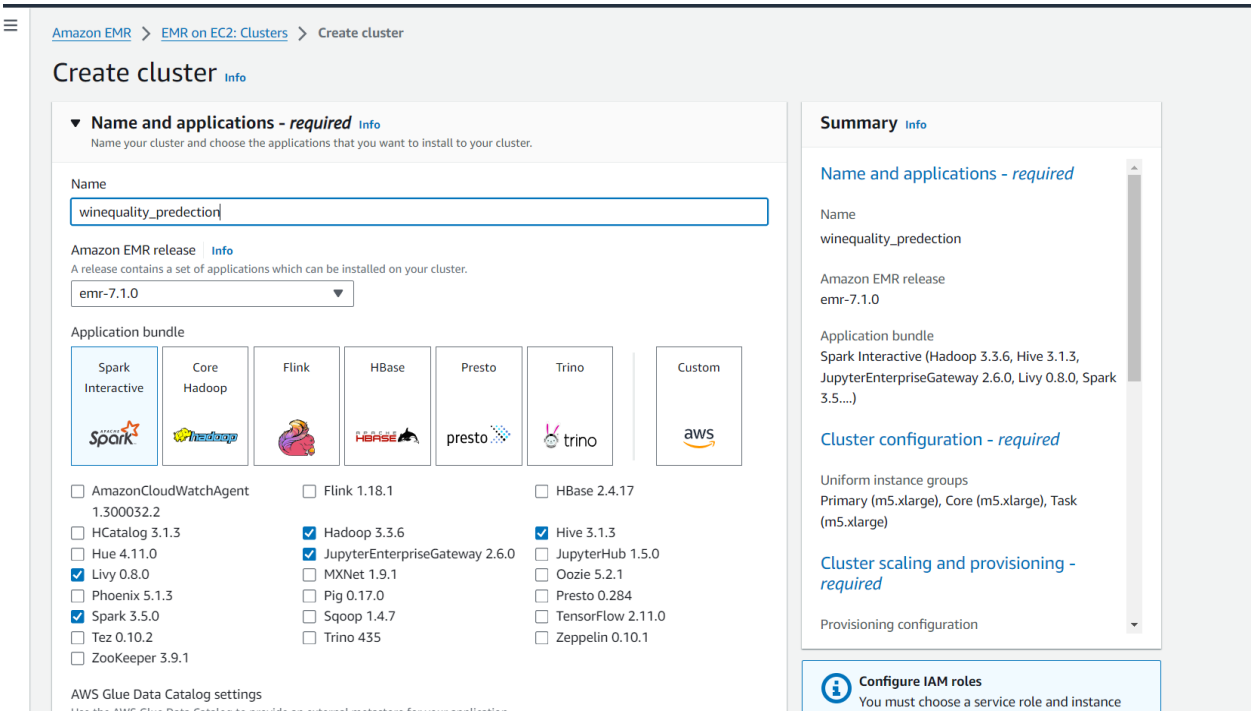


3.You'll be taken to the Clusters page where you'll see that there are currently no active clusters.

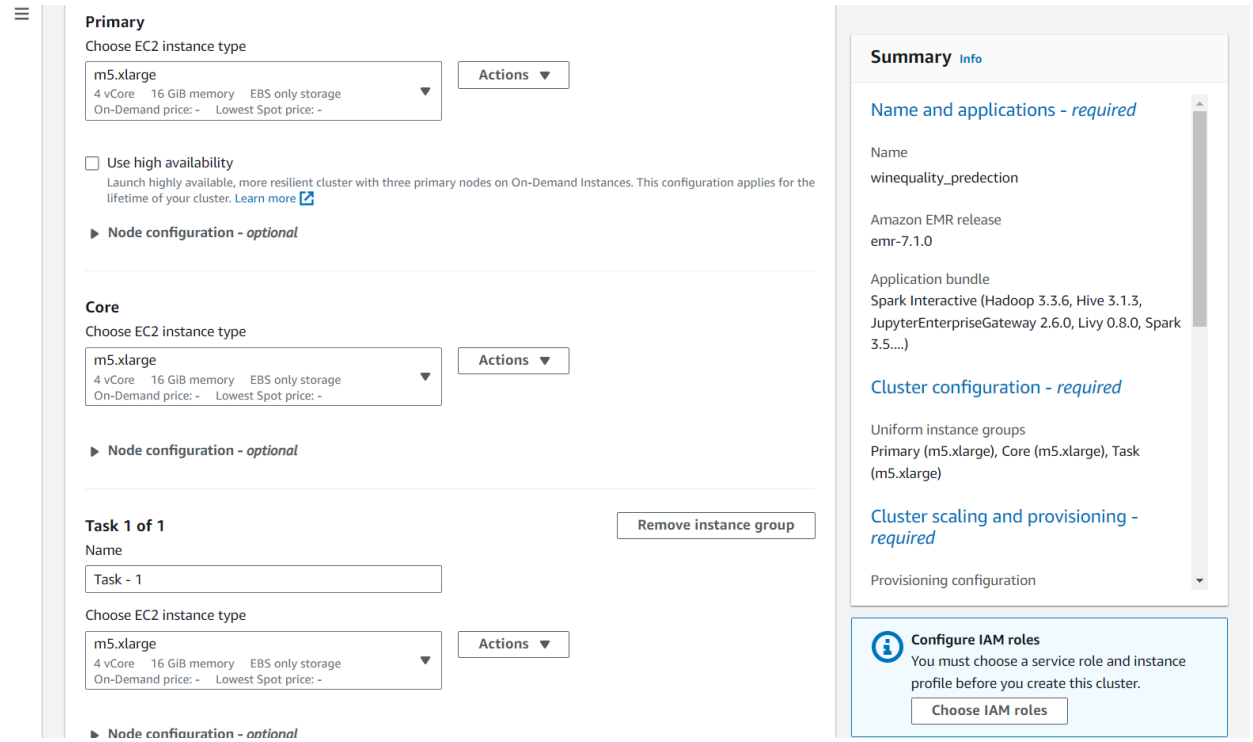
4.To create a new cluster, click on the "Create cluster" button.



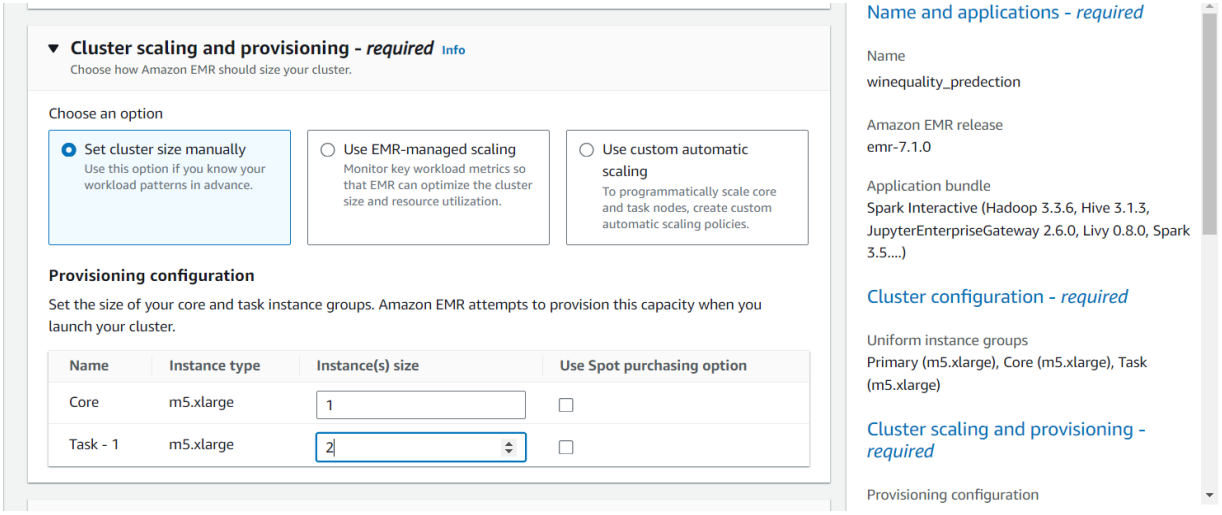
5.Give any name you want to the cluster. Additionally, ensure that you select the latest version of EMR available and choose 'Spark Interactive' under the Application bundle options.



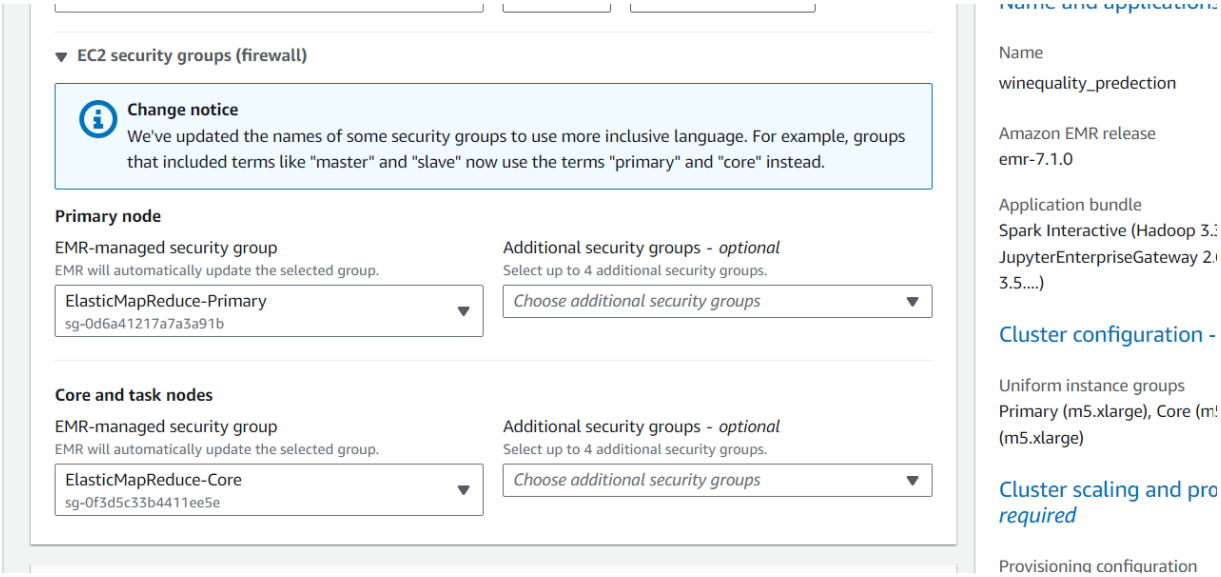
6.I selected the instance type m5.xlarge for the Primary, Core, and Task instance groups. You have the flexibility to choose any instance type that suits your needs.



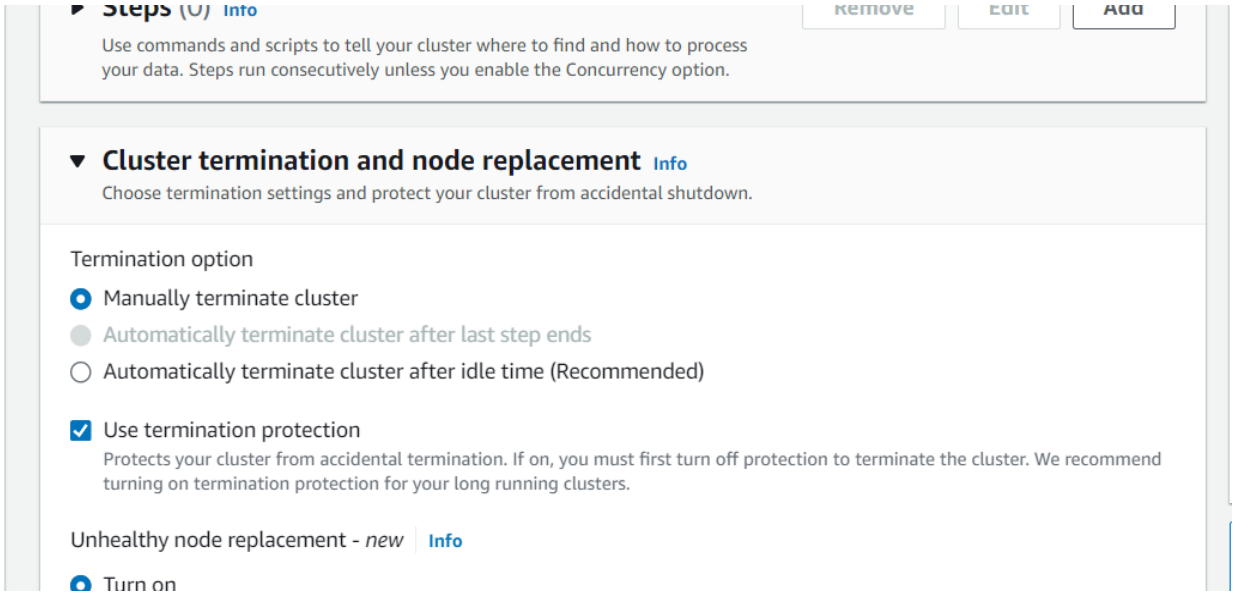
7. To ensure cluster scalability and provisioning, configure Core's instance size to 1 and Task-1's instance size to 2.



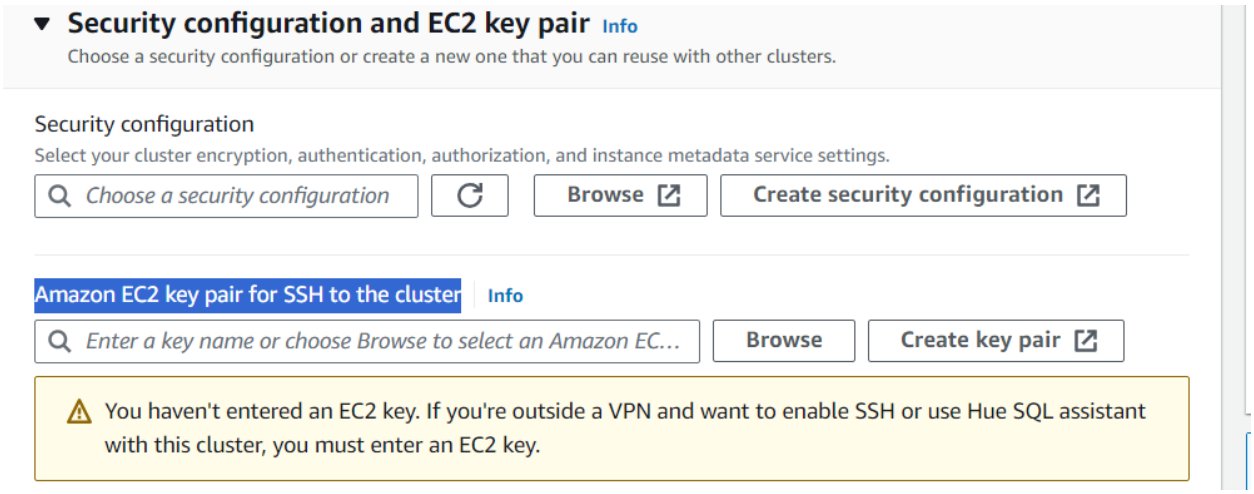
8. In order to pick the security groups for EC2's Primary node, Core, and task nodes, select the security groups that are displayed below.



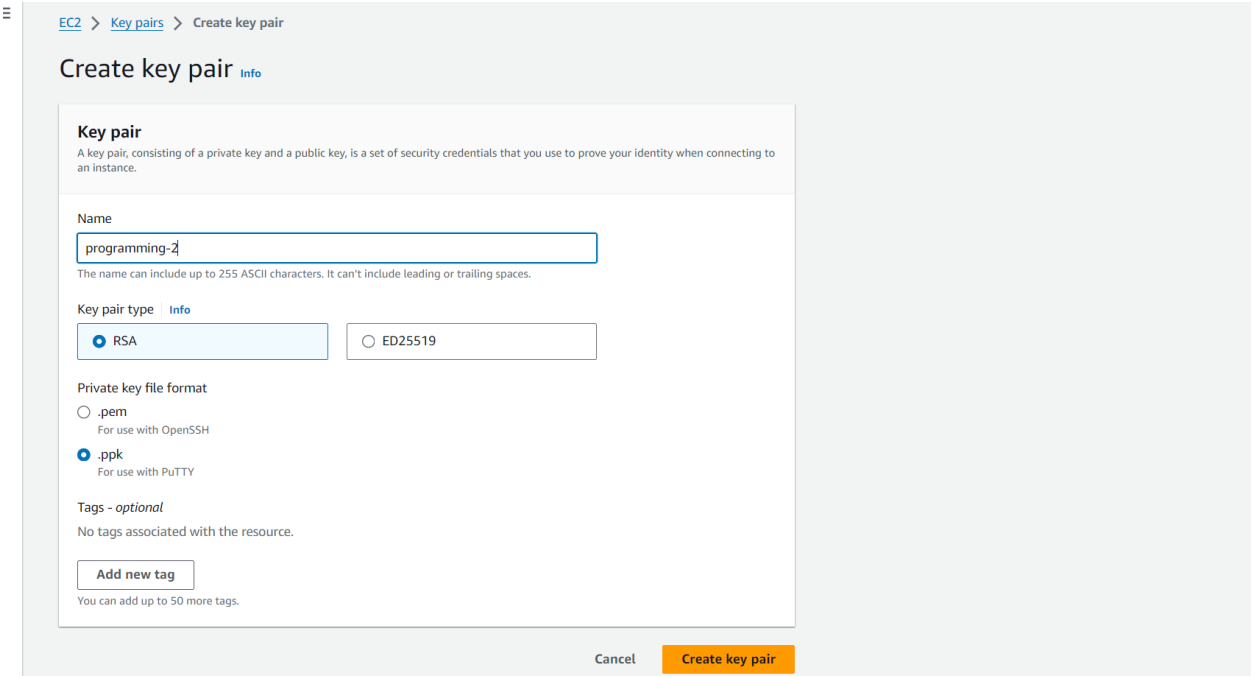
9. To prevent the cluster from being terminated automatically, make sure that the option to manually end the cluster is selected.



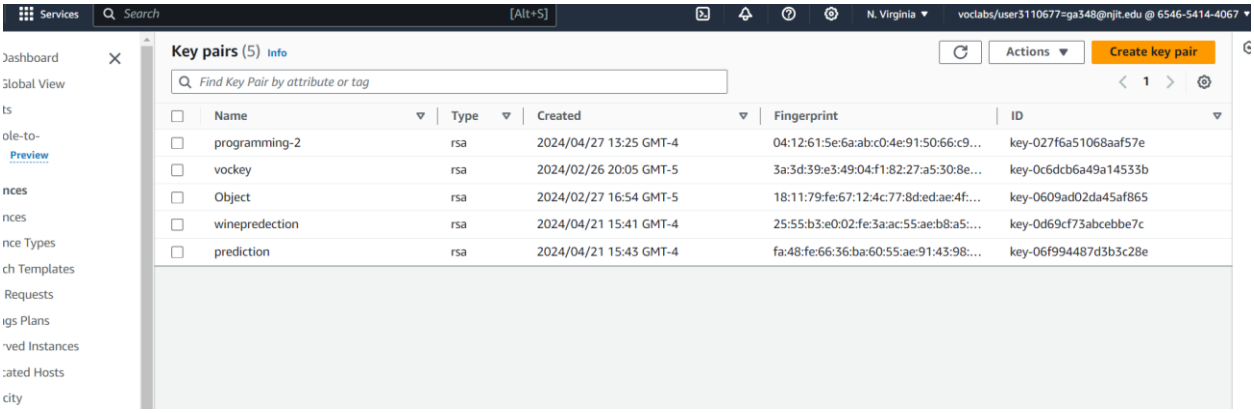
10. create a new key pair



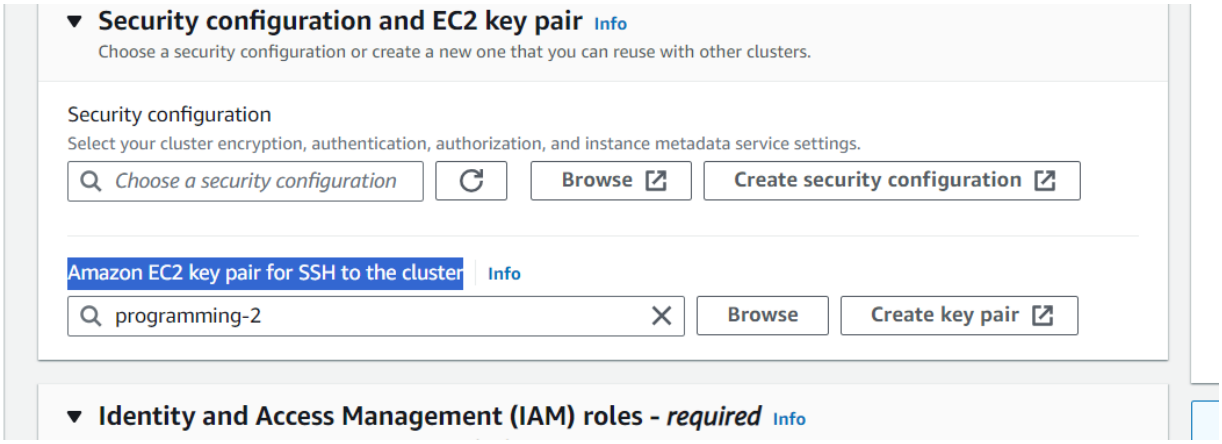
11. Provide a custom name for the key pair and select 'ppk' as the file format when creating the key pair



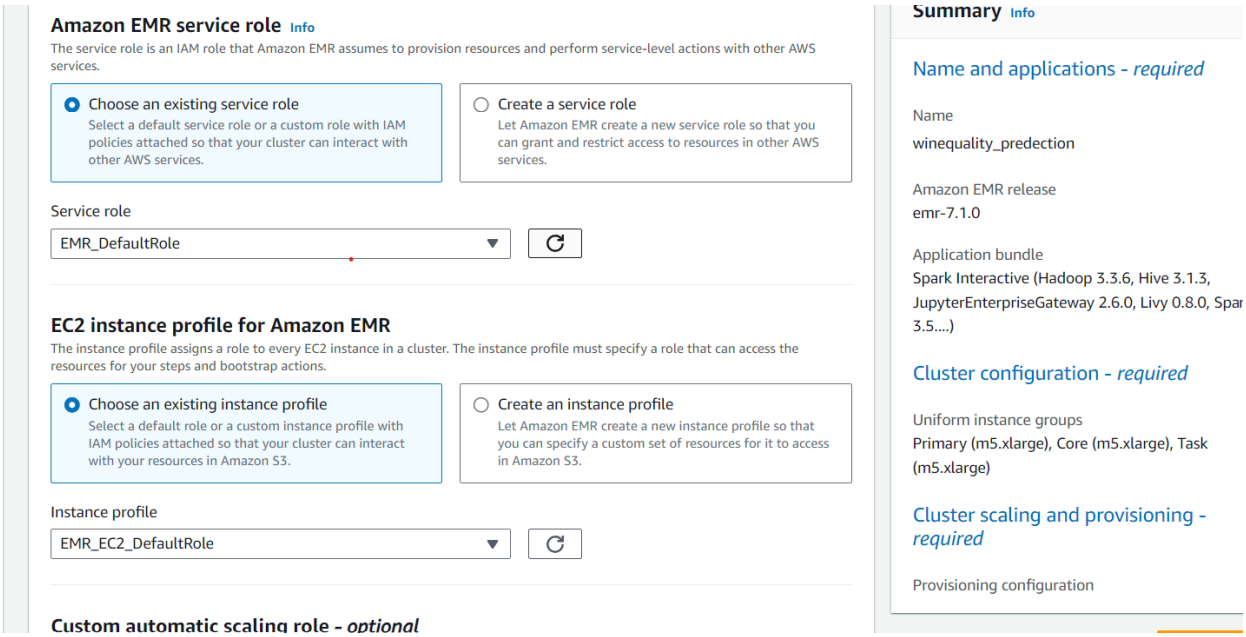
12. The key pair named 'programming-2' has been created, and the key has been downloaded and stored on the local system for connecting to the cluster using PuTTY as the SSH server.



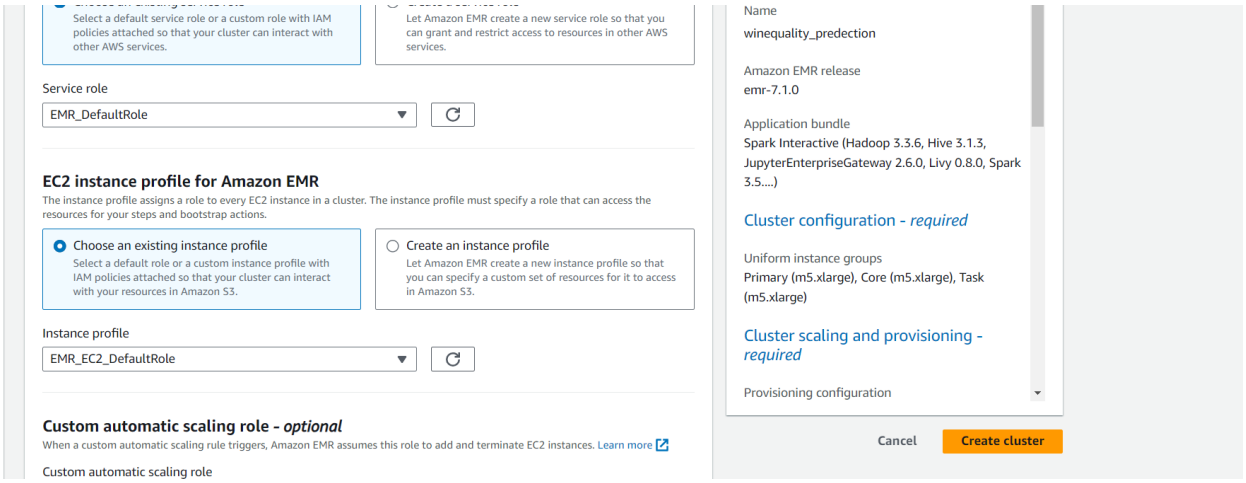
13. The key pair named 'programming-2' has been created, and you can now browse the key .



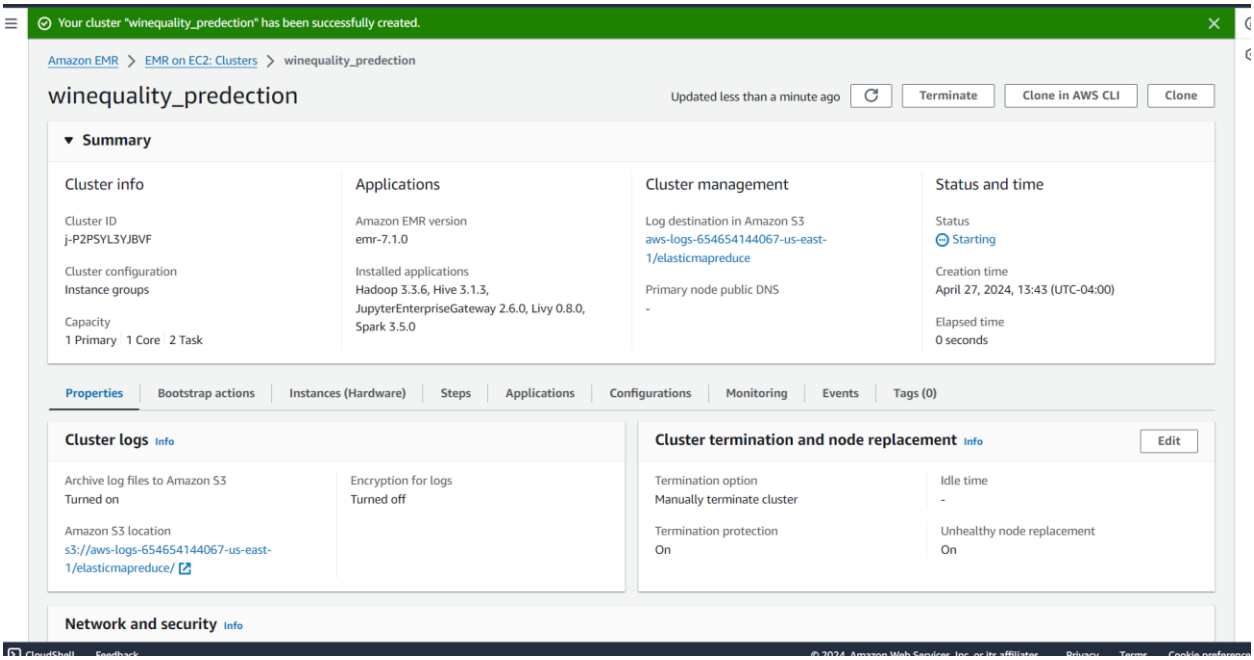
14. Go to the IAM roles and choose them as indicated below.



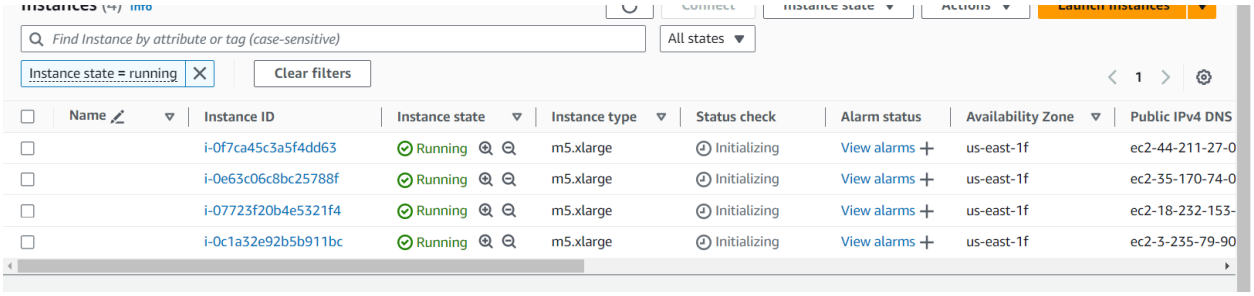
15. Click the button labeled "Create Cluster" to initiate the process of creating the clusters.

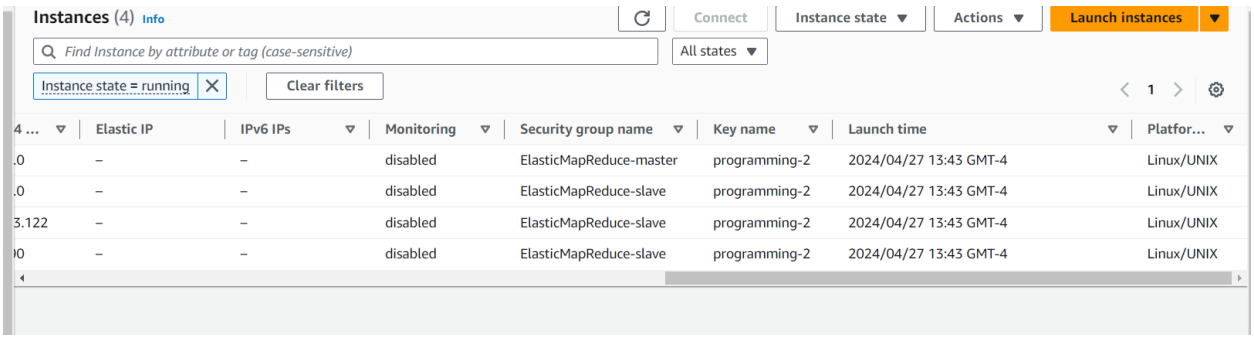


16. Cluster created successfully.

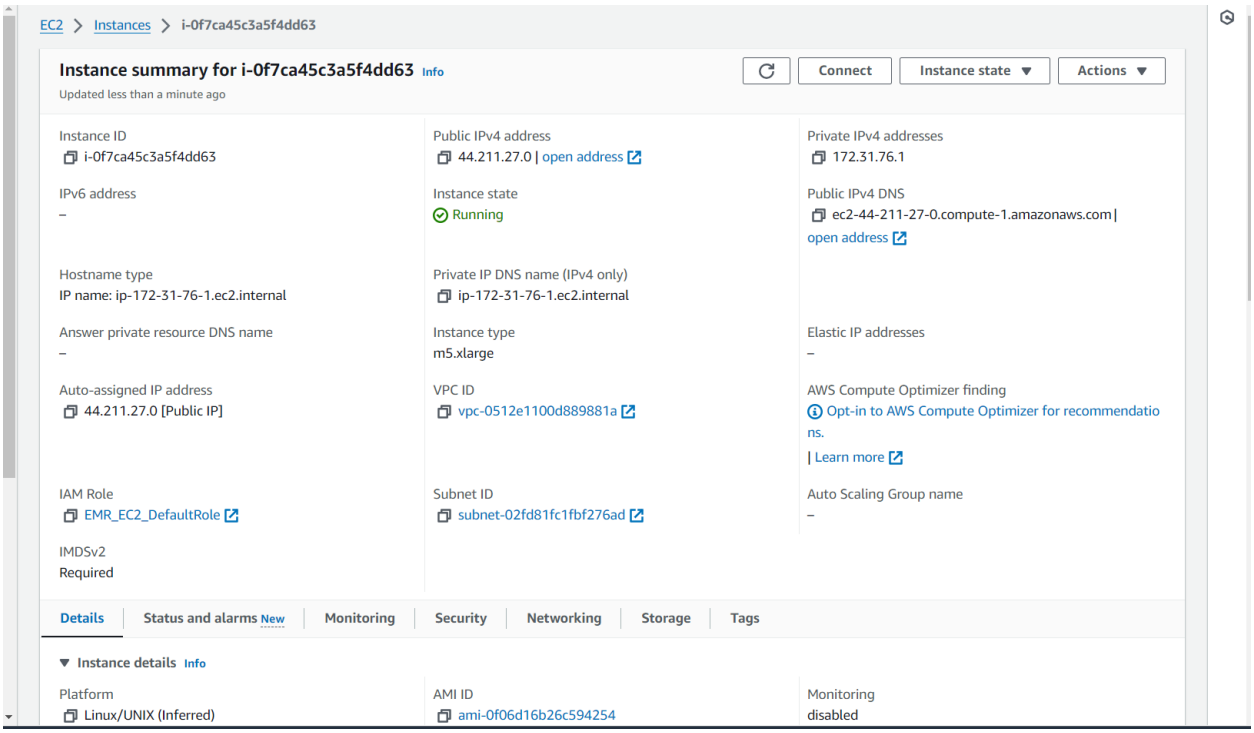


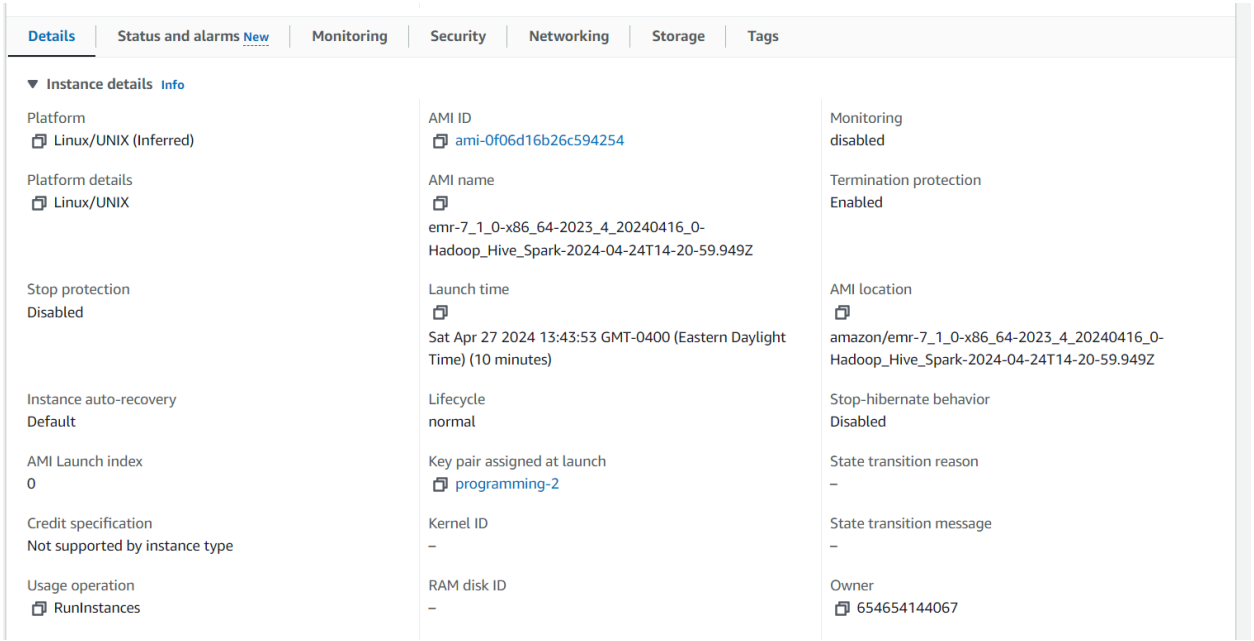
17. Navigate to the EC2 instance page. As seen below, there are four EC2 instances launched, one of which is a Master node and the other three are Slave nodes.



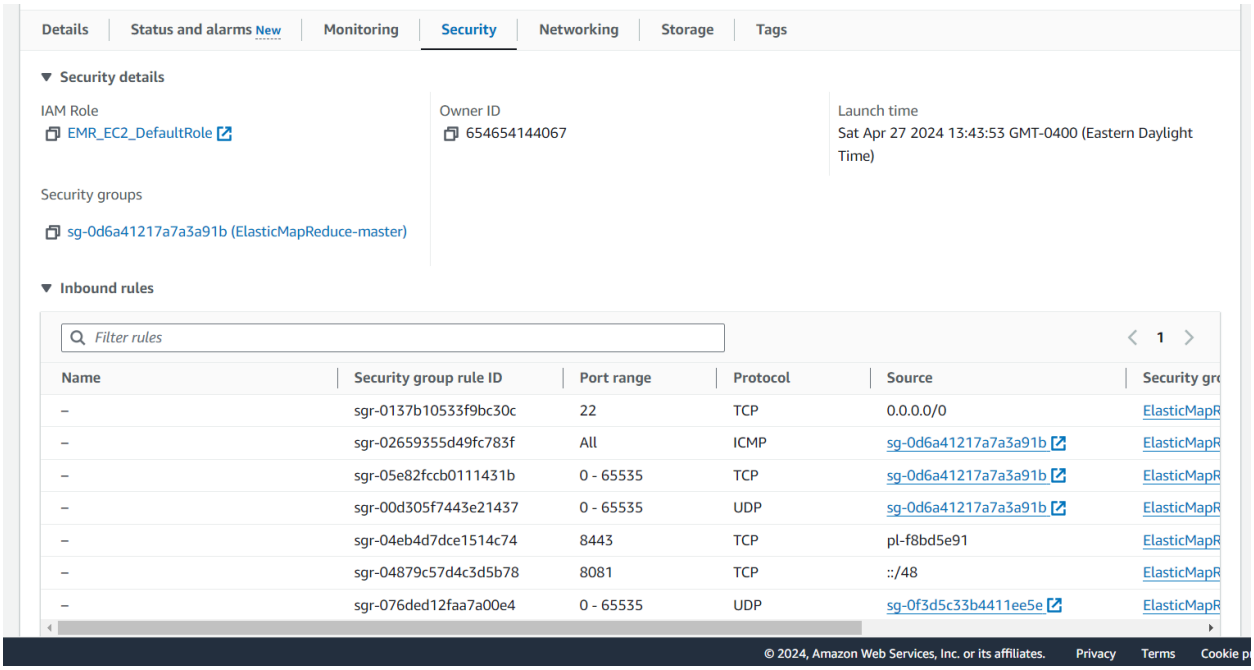


18. To access the "ElasticMapReduce-Master" security group, navigate to the EC2 service and select the appropriate Security ID.

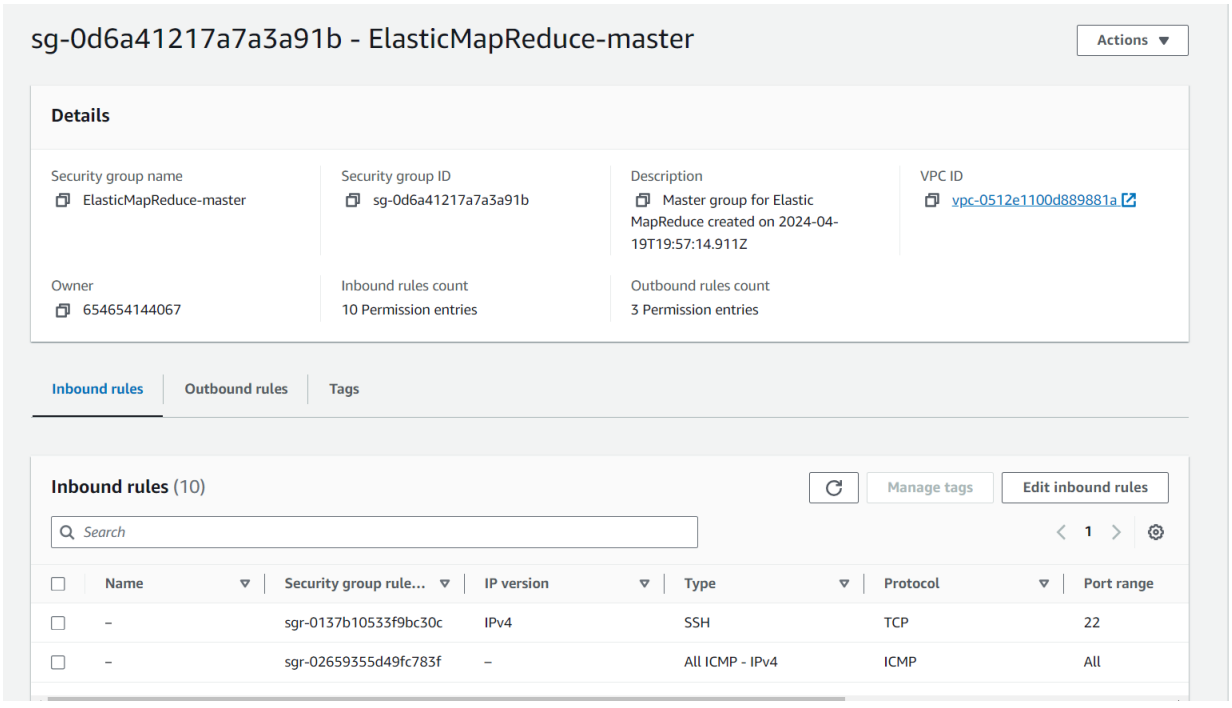


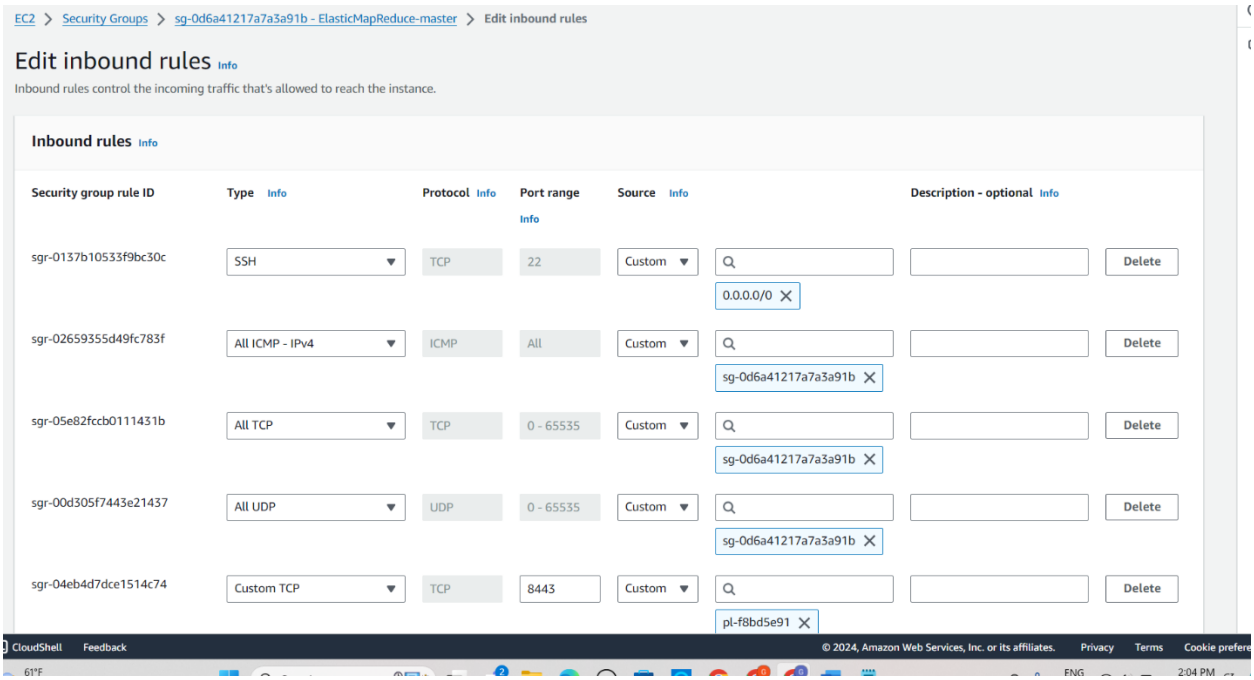


19.Click on the security group to edit its inbound rules

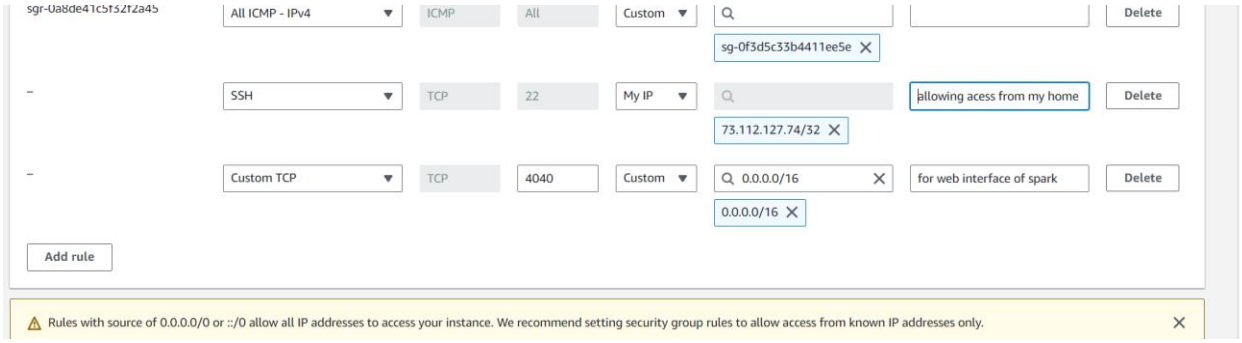


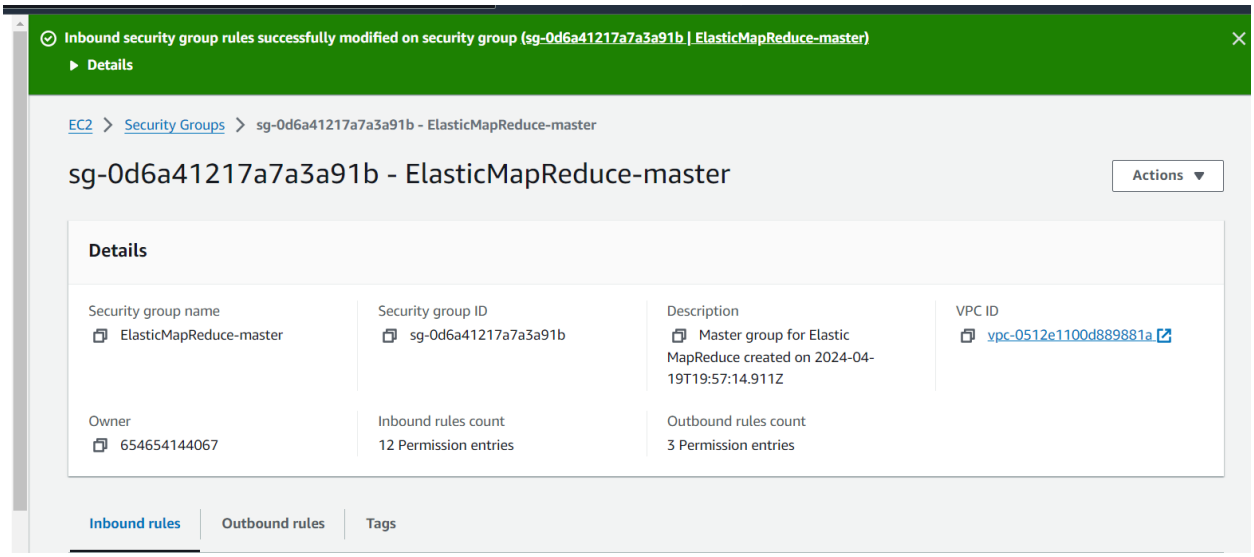
20. In the Inbound Rules section, select Edit Inbound Rules.





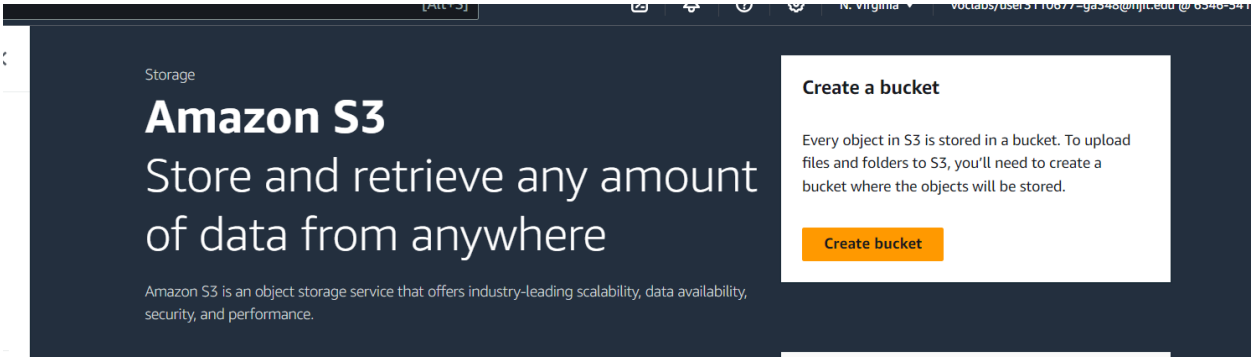
21. Enter the port numbers 22 and 4040 with the settings indicated below, then click the Add rule button and save the rules.



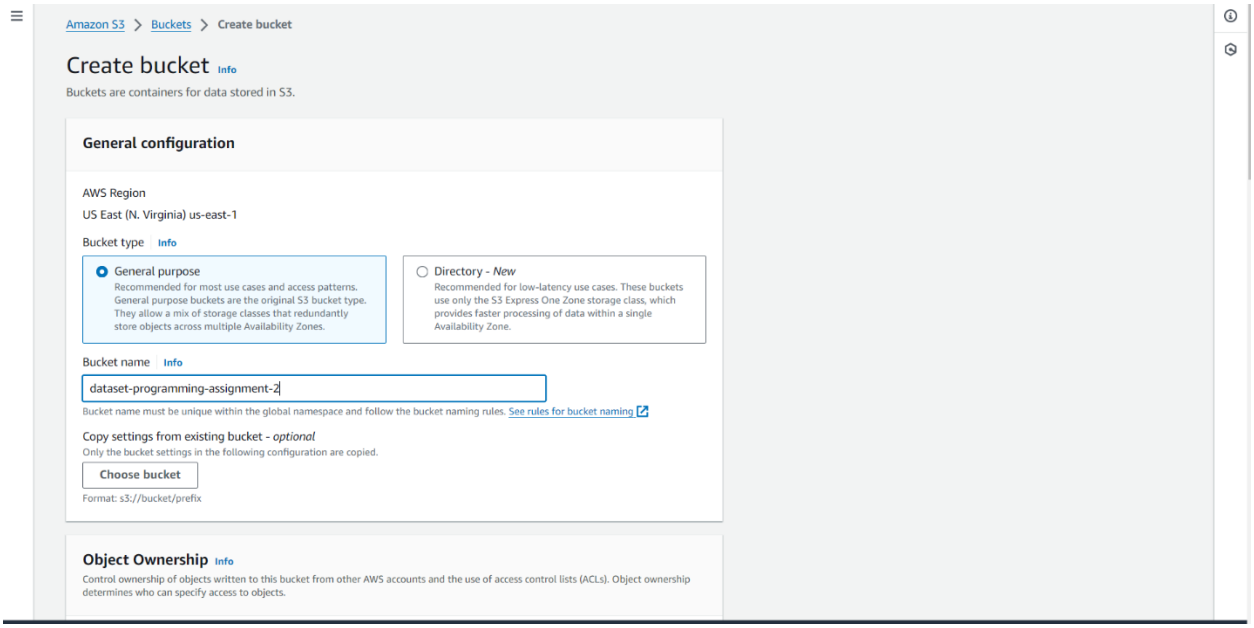


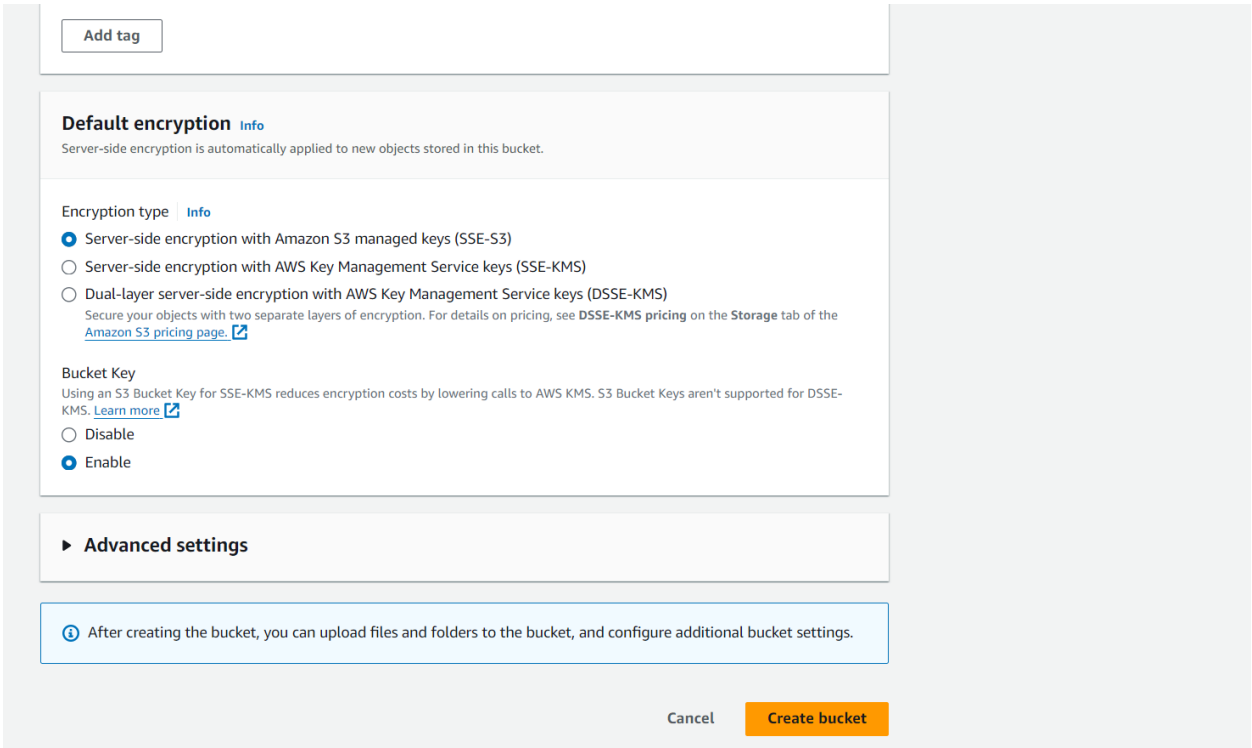
22. Create an S3 bucket in AWS services to store the dataset.

23.Choose "Create Bucket".

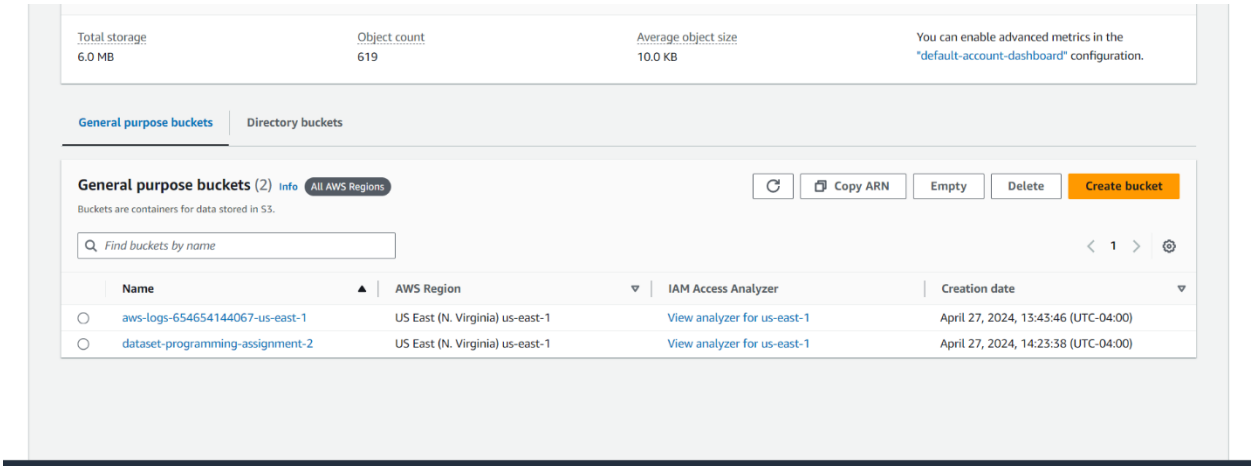


24.Label your bucket as "dataset-programming-assignment-2". Scroll down and click on the "Create bucket" button.

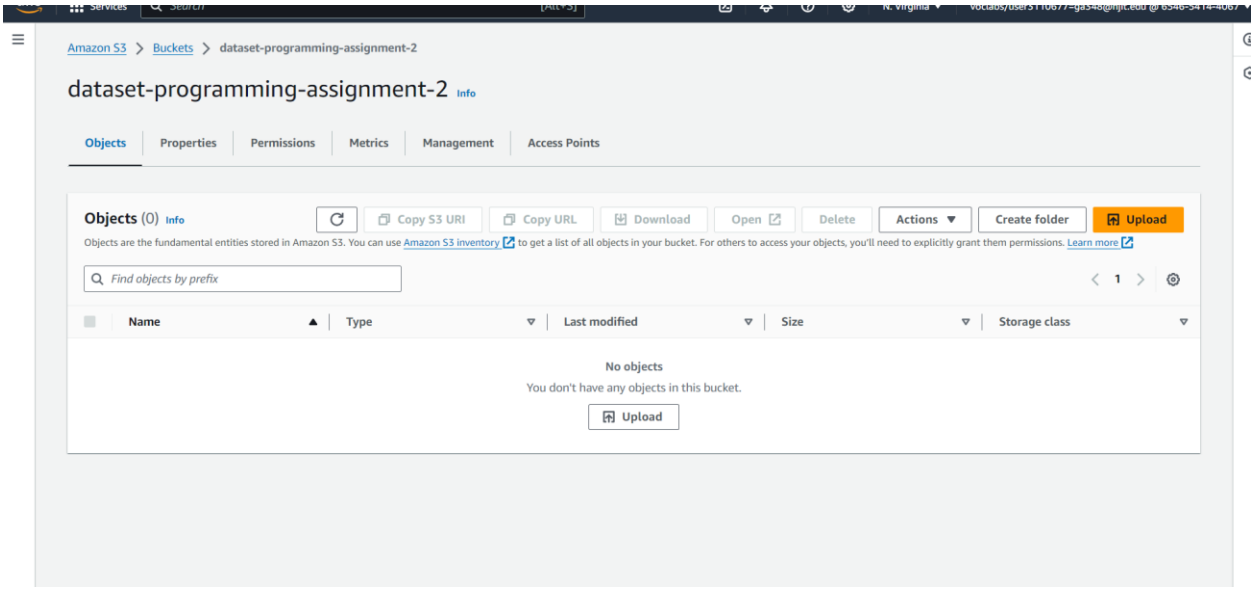




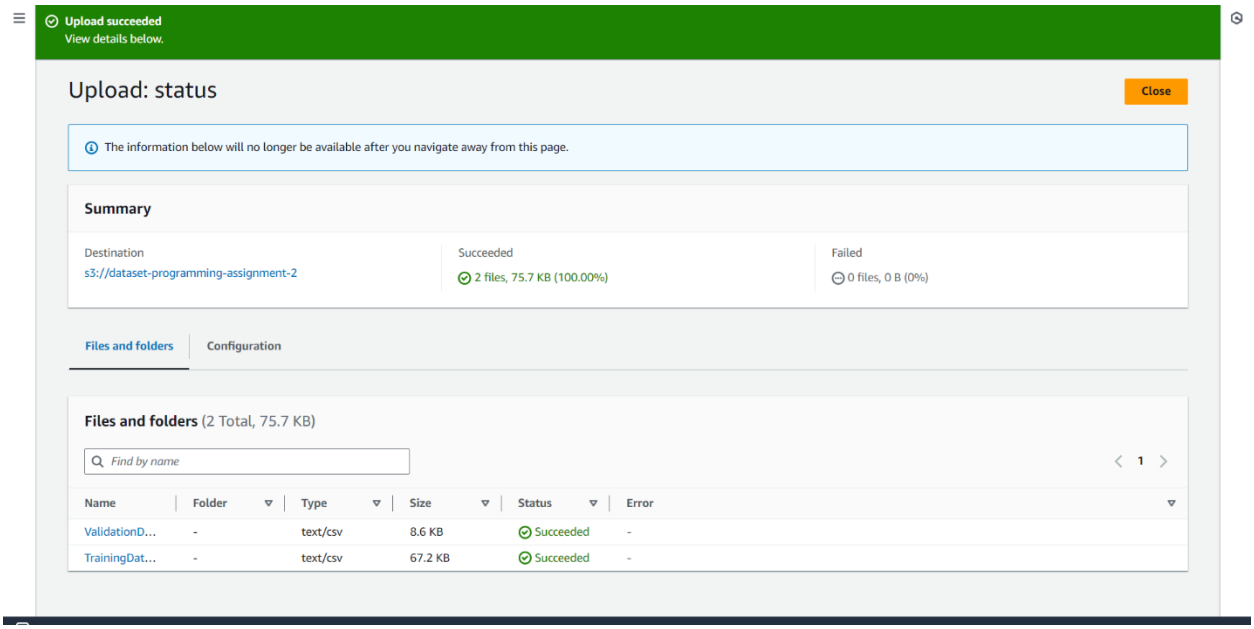
25. Access the buckets page to view your newly created bucket.
26. Click on the bucket name you created.



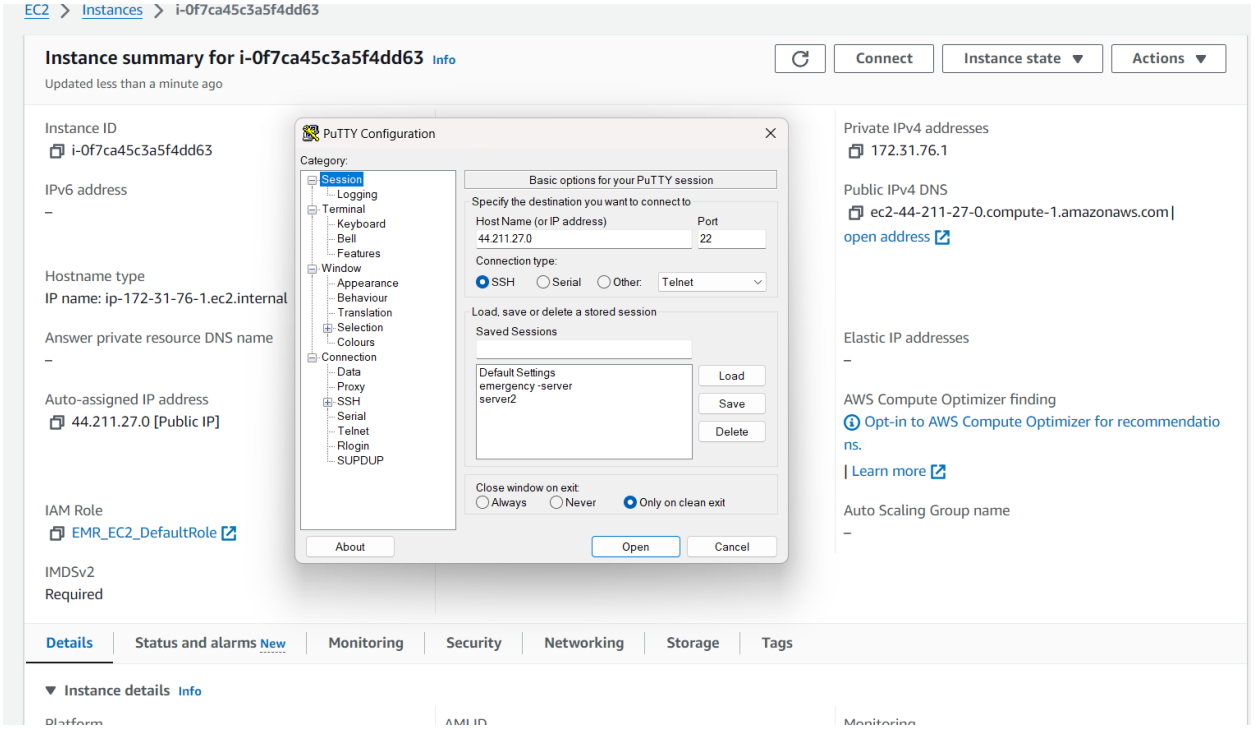
27. Click the upload button.



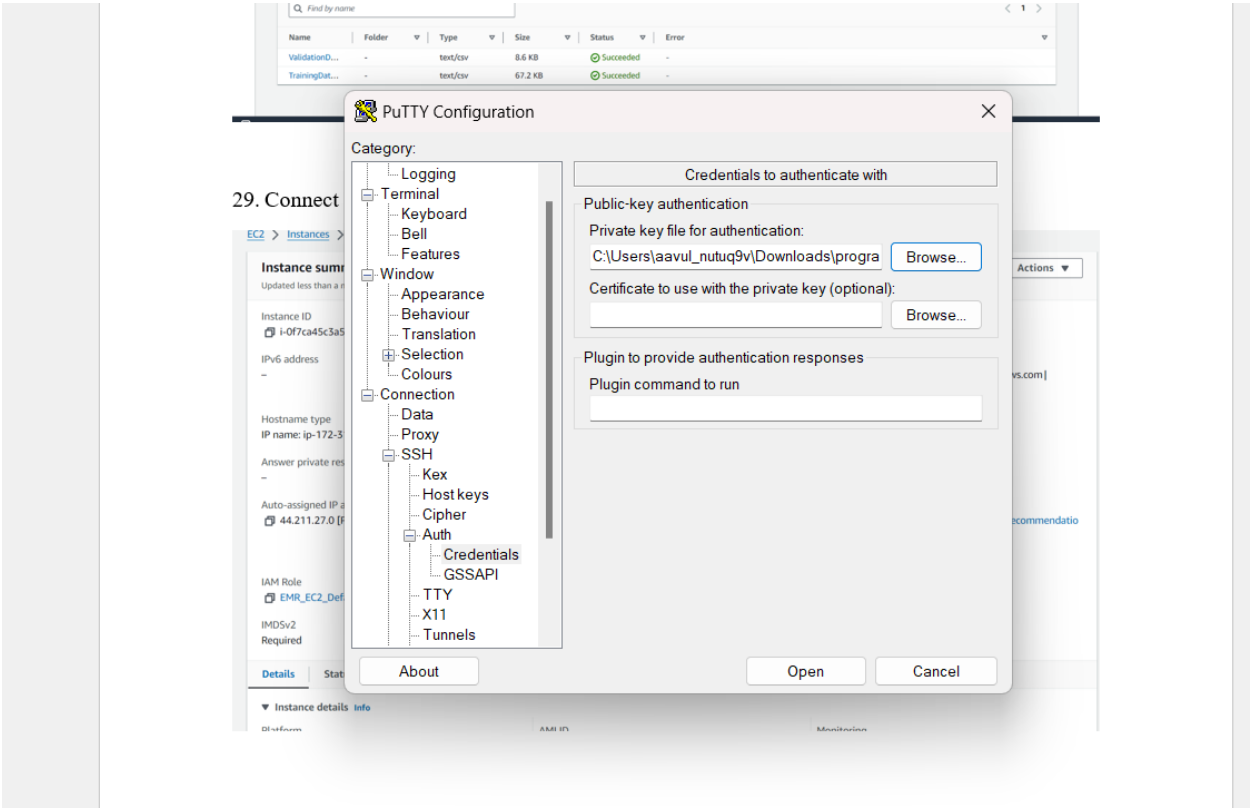
28. Select the.csv files from your local system by clicking on "Add Files." Now, click the "Upload button" to place the dataset in the S3 bucket. You now have two.csv files in your S3 bucket: ValidationDataset.csv and TrainingDataset.csv.



29. Connect to the server using PuTTY by specifying the PPK file for authentication.

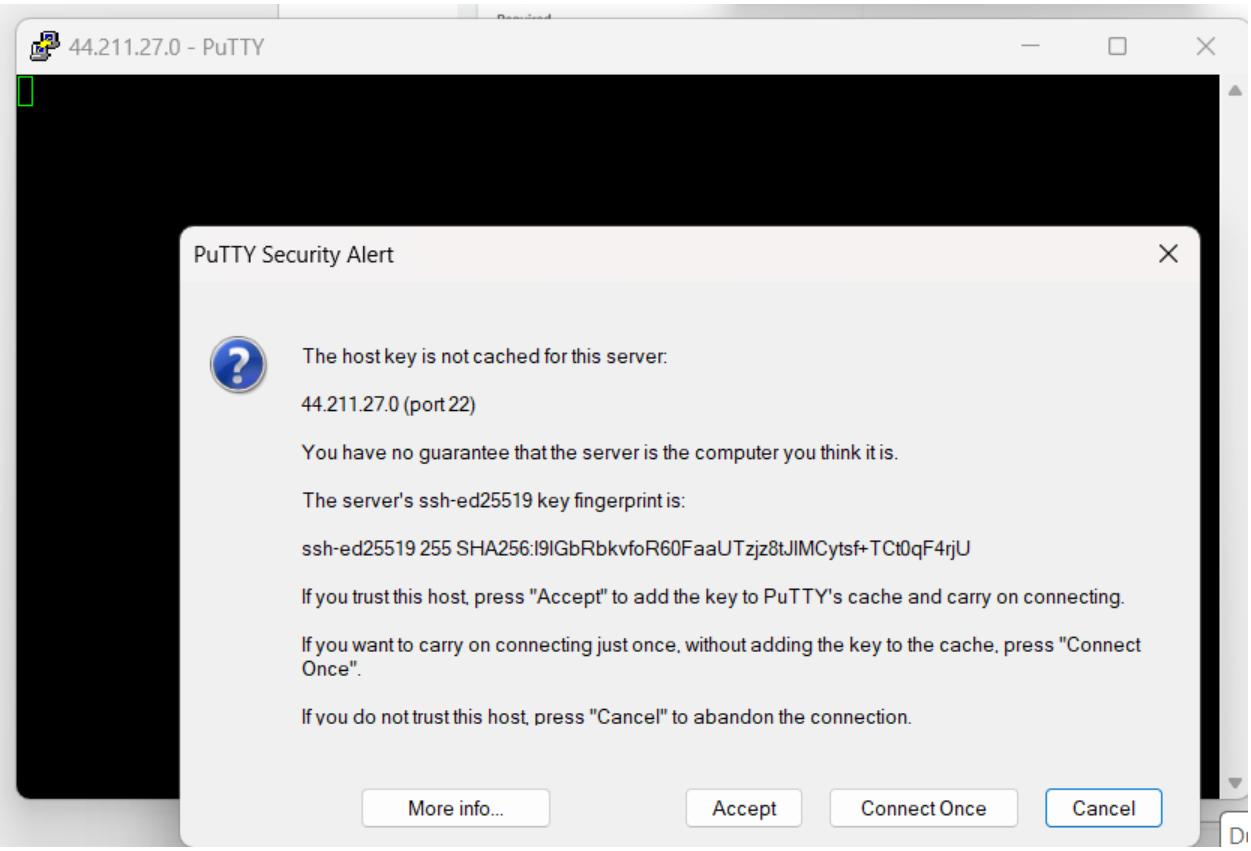


30. Click on "SSH" under "Auth", select "Credentials", provide the path to the location of the PPK file, and then click "Open".

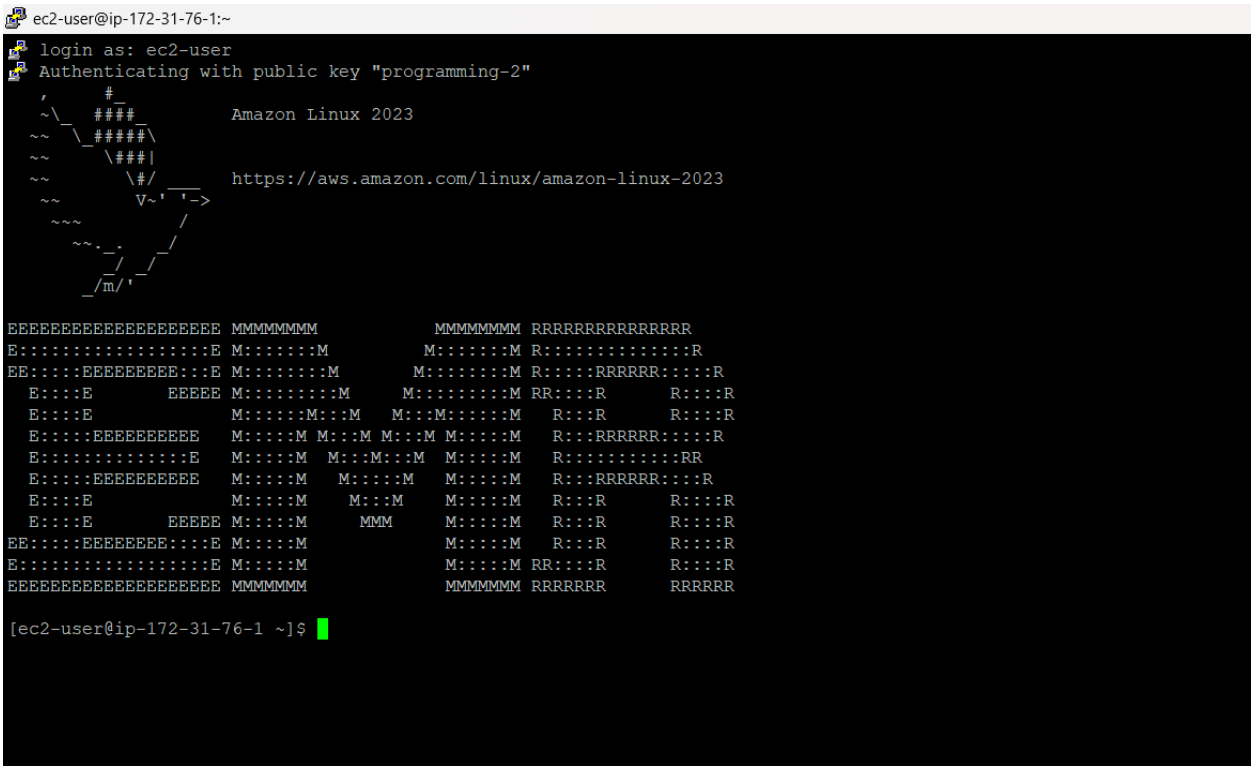


29. Connect

31.click on accept



32. provide ec2-user



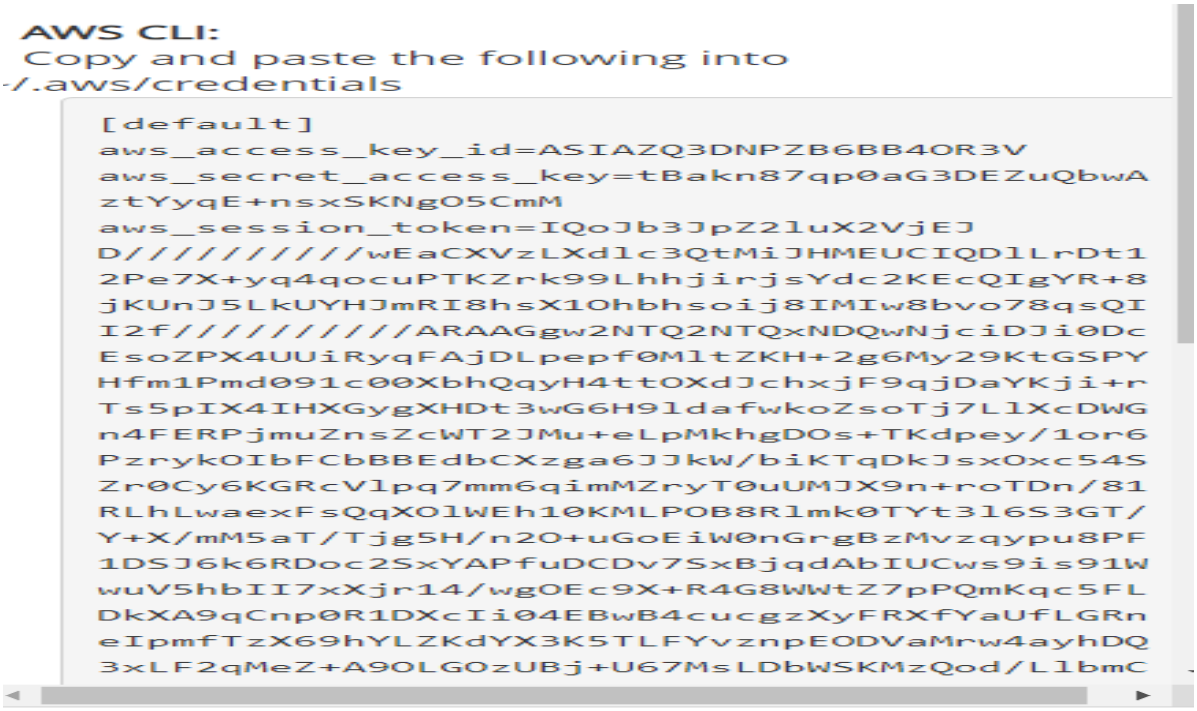
33.To set up credentials, configure them in the Master node EC2 instance.
To configure credentials for the master node, type the following instructions in your terminal:

mkdir .aws

touch .aws/credentials

vi .aws/credentials

34. Copy and paste the credentials from the AWS Academy page and AWS information.



35.To execute our Spark application, we must first install the required packages.
Run the commands below:
sudo yum update.
sudo yum install git.
pip install pyspark findspark boto3 numpy pandas scikit-learn datetime

36.To clone the GitHub repository, enter the following command:
#git clone https://github.com/gayatriaavula/winepred_quality.git

37.To get started with the Spark Application, perform the following commands:

```
# spark-submit --master yarn CS643_Programming_assignment_2/WineTraining.py
# spark-submit --master yarn CS643_Programming_assignment_2/WineTesting.py > output.txt
```

```
[hadoop@ip-172-31-69-133 ~]$ spark-submit --master yarn CS643_Programming_assignment_2/WineTraining.py
Apr 26, 2024 8:56:28 PM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location

24/04/26 20:56:32 INFO SparkContext: Running Spark version 3.5.0-amzn-1
24/04/26 20:56:32 INFO SparkContext: OS info Linux, 6.1.84-99.169.amzn2023.x86_64, amd64
24/04/26 20:56:32 INFO SparkContext: Java version 17.0.10
24/04/26 20:56:32 INFO ResourceUtils: ==============================================================
24/04/26 20:56:32 INFO ResourceUtils: No custom resources configured for spark.driver.
24/04/26 20:56:32 INFO ResourceUtils: ==============================================================
24/04/26 20:56:32 INFO SparkContext: Submitted application: WineQuality Training
24/04/26 20:56:32 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 4, script: , vendor: , memory -> name: memory, amount: 9486, s
cript: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/04/26 20:56:32 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/04/26 20:56:32 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/04/26 20:56:32 INFO SecurityManager: Changing view acls to: hadoop
24/04/26 20:56:32 INFO SecurityManager: Changing modify acls to: hadoop
24/04/26 20:56:32 INFO SecurityManager: Changing view acls groups to:
24/04/26 20:56:32 INFO SecurityManager: Changing modify acls groups to:
24/04/26 20:56:32 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modi
fy permissions: hadoop; groups with modify permissions: EMPTY
24/04/26 20:56:32 INFO Utils: Successfully started service 'sparkDriver' on port 40123.
24/04/26 20:56:32 INFO SparkEnv: Registering MapOutputTracker
24/04/26 20:56:32 INFO SparkEnv: Registering BlockManagerMaster
24/04/26 20:56:32 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/04/26 20:56:32 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/04/26 20:56:32 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/26 20:56:32 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-19a6d5b6-a51f-4d1b-b115-eb9b8f2239af
24/04/26 20:56:32 INFO MemoryStore: MemoryStore started with capacity 1048.8 MiB
24/04/26 20:56:32 INFO SparkEnv: Registering OutputCommitCoordinator
24/04/26 20:56:32 INFO SubResultCacheManager: Sub-result caches are disabled.
24/04/26 20:56:32 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/04/26 20:56:33 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/26 20:56:33 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
24/04/26 20:56:33 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-69-133.ec2.internal/172.31.69.133:8032
24/04/26 20:56:33 INFO Configuration: resource-types.xml not found
24/04/26 20:56:33 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/04/26 20:56:33 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
24/04/26 20:56:33 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/04/26 20:56:33 INFO Client: Setting up container launch context for our AM
24/04/26 20:56:33 INFO Client: Setting up the launch environment for our AM container
24/04/26 20:56:33 INFO Client: Preparing resources for our AM container
24/04/26 20:56:33 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
24/04/26 20:56:34 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f66a-0ddf-4955-bde0-c52a65fe4c83/__spark_libs__1278197152423226073.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
/user/hadoop/.sparkStaging/application_1714161677101_0003/__spark_libs__1278197152423226073.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_00
03/hive-site.xml
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171416167710
1_0003/hudi-defaults.conf
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101
_0003/pyspark.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171
```

```
24/04/26 20:56:33 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/26 20:56:33 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
24/04/26 20:56:33 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-69-133.ec2.internal/172.31.69.133:8032
24/04/26 20:56:33 INFO Configuration: resource-types.xml not found
24/04/26 20:56:33 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/04/26 20:56:33 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
24/04/26 20:56:33 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/04/26 20:56:33 INFO Client: Setting up container launch context for our AM
24/04/26 20:56:33 INFO Client: Setting up the launch environment for our AM container
24/04/26 20:56:33 INFO Client: Preparing resources for our AM container
24/04/26 20:56:33 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
24/04/26 20:56:34 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f66a-0ddf-4955-bde0-c52a65fe4c83/__spark_libs__1278197152423226073.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
/user/hadoop/.sparkStaging/application_1714161677101_0003/__spark_libs__1278197152423226073.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_00
03/hive-site.xml
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171416167710
1_0003/hudi-defaults.conf
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101
_0003/pyspark.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171
4161677101_0003/py4j-0.10.9.7-src.zip
24/04/26 20:56:36 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f66a-0ddf-4955-bde0-c52a65fe4c83/__spark_conf__1049270680638660577.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
/user/hadoop/.sparkStaging/application_1714161677101_0003/__spark_conf__.zip
24/04/26 20:56:36 INFO SecurityManager: Changing view acls to: hadoop
24/04/26 20:56:36 INFO SecurityManager: Changing modify acls to: hadoop
24/04/26 20:56:36 INFO SecurityManager: Changing view acls groups to:
24/04/26 20:56:36 INFO SecurityManager: Changing modify acls groups to:
24/04/26 20:56:36 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modi
fy permissions: hadoop; groups with modify permissions: EMPTY
24/04/26 20:56:36 INFO Client: Submitting application application_1714161677101_0003 to ResourceManager
24/04/26 20:56:36 INFO YarnClientImpl: Submitted application application_1714161677101_0003
24/04/26 20:56:37 INFO Client: Application report for application_1714161677101_0003 (state: ACCEPTED)
24/04/26 20:56:37 INFO Client:
        client token: N/A
        diagnostics: AM container is launched, waiting for AM container to Register with RM
        ApplicationMaster host: N/A
        ApplicationMaster RPC port: -1
        queue: default
        start time: 1714164996129
        final status: UNDEFINED
        tracking URL: http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003/
        user: hadoop
24/04/26 20:56:40 INFO YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY_HOSTS -> ip-172-31-69-133.ec2.internal, PROXY_URI_
BASES -> http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003), /proxy/application_1714161677101_0003
24/04/26 20:56:41 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
24/04/26 20:56:41 INFO Client: Application report for application_1714161677101_0003 (state: RUNNING)
24/04/26 20:56:41 INFO Client:
        client token: N/A
        diagnostics: N/A
        ApplicationMaster host: 172.31.79.179
        ApplicationMaster RPC port: -1
        queue: default
        start time: 1714164996129
        final status: UNDEFINED
        tracking URL: http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003/
```

```
        user: hadoop
24/04/26 20:56:41 INFO YarnClientSchedulerBackend: Application application_1714161677101_0003 has started running.
24/04/26 20:56:41 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 40265.
24/04/26 20:56:41 INFO NettyBlockTransferService: Server created on ip-172-31-69-133.ec2.internal:40265
24/04/26 20:56:41 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/26 20:56:41 INFO BlockManager: external shuffle service port = 7337
24/04/26 20:56:41 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-69-133.ec2.internal:40265 with 1048.8 MiB RAM, BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 4
0265, None)
24/04/26 20:56:41 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO SingleEventLogFileWriter: Logging events to hdfs:/var/log/spark/apps/application_1714161677101_0003.inprogress
24/04/26 20:56:41 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/job: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/job/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/pool: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/pool/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/rdd: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/rdd/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /environment: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
Importing: s3a://dataset-programming-assignment-2/TrainingDataset.csv
>>>> Model Path set: s3a://dataset-programming-assignment-2/models
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
root
 |-- """"fixed acidity""""": string (nullable = true)
 |-- """"volatile acidity""""": string (nullable = true)
 |-- """"citric acid""""": string (nullable = true)
 |-- """"residual sugar""""": string (nullable = true)
 |-- """"chlorides""""": string (nullable = true)
 |-- """"free sulfur dioxide""""": string (nullable = true)
 |-- """"total sulfur dioxide""""": string (nullable = true)
```

```
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
root
 |-- """"fixed acidity""""": string (nullable = true)
 |-- """"volatile acidity""""": string (nullable = true)
 |-- """"citric acid""""": string (nullable = true)
 |-- """"residual sugar""""": string (nullable = true)
 |-- """"chlorides""""": string (nullable = true)
 |-- """"free sulfur dioxide""""": string (nullable = true)
 |-- """"total sulfur dioxide""""": string (nullable = true)
 |-- """"density""""": string (nullable = true)
 |-- """"pH""""": string (nullable = true)
 |-- """"sulphates""""": string (nullable = true)
 |-- """"alcohol""""": string (nullable = true)
 |-- """"quality""""": string (nullable = true)

+--------------------+-----------------------+-------------------+---------------------+-----------------+--------------------------+--------------------------+--------------+------
----+-----------------+---------------+---------------+
|""""fixed acidity""""|""""volatile acidity""""|""""citric acid""""|""""residual sugar""""|""""chlorides""""|""""free sulfur dioxide""""|""""total sulfur dioxide""""|""""density""""|""""pH
""""|""""sulphates""""|""""alcohol""""|""""quality""""|
+--------------------+-----------------------+-------------------+---------------------+-----------------+--------------------------+--------------------------+--------------+------
----+-----------------+---------------+---------------+
|                8.9|                   0.22|               0.48|                  1.8|            0.077|                      29|                      60|        0.9968|
3.39|             0.53|            9.4|              6|
|                7.6|                   0.39|               0.31|                  2.3|            0.082|                      23|                      71|        0.9982|
3.52|             0.65|            9.7|              5|
|                7.9|                   0.43|               0.21|                  1.6|            0.106|                      10|                      37|        0.9966|
3.17|             0.91|            9.5|              5|
|                8.5|                   0.49|               0.11|                  2.3|            0.084|                       9|                      67|        0.9968|
3.17|             0.53|            9.4|              5|
|                6.9|                    0.4|               0.14|                  2.4|            0.085|                      21|                      40|        0.9968|
3.43|             0.63|            9.7|              6|
|                6.3|                   0.39|               0.16|                  1.4|             0.08|                      11|                      23|        0.9955|
3.34|             0.56|            9.3|              5|
|                7.6|                   0.41|               0.24|                  1.8|             0.08|                       4|                      11|        0.9962|
3.28|             0.59|            9.5|              5|
|                7.9|                   0.43|               0.21|                  1.6|            0.106|                      10|                      37|        0.9966|
3.17|             0.91|            9.5|              5|
|                7.1|                   0.71|                0.0|                  1.9|             0.08|                      14|                      35|        0.9972|
3.47|             0.55|            9.4|              5|
|                7.8|                  0.645|                0.0|                  2.0|            0.082|                       8|                      16|        0.9964|
3.38|             0.59|            9.8|              6|
|                6.7|                  0.675|               0.07|                  2.4|            0.089|                      17|                      82|        0.9958|
3.35|             0.54|           10.1|              5|
|                6.9|                  0.685|                0.0|                  2.5|            0.105|                      22|                      37|        0.9966|
3.46|             0.57|           10.6|              6|
|                8.3|                  0.655|               0.12|                  2.3|            0.083|                      15|                     113|        0.9966|
3.17|             0.66|            9.8|              5|
|                6.9|                  0.605|               0.12|                 10.7|            0.073|                      40|                      83|        0.9993|
3.45|             0.52|            9.4|              6|
|                5.2|                   0.32|               0.25|                  1.8|            0.103|                      13|                      50|        0.9957|
3.38|             0.55|            9.2|              5|
|                7.8|                  0.645|                0.0|                  5.5|            0.086|                       5|                      18|        0.9986|
 3.4|             0.55|            9.6|              6|
|                7.8|                    0.6|               0.14|                  2.4|            0.086|                       3|                      15|        0.9975|
```

The results will be displayed here, along with the Accuracy and F1 scores of the Machine Learning methods used.

```
Training DataSet Metrics
Accuracy: 0.6114151681000782
F-measure: 0.5904050519731796
+-------------+----------------+-----------+-------------+---------+--------------------+--------------------+-------+----+---------+-------+-----+
|fixed_acidity|volatile_acidity|citric_acid|residual_sugar|chlorides|free_sulfur_dioxide|total_sulfur_dioxide|density|  pH|sulphates|alcohol|label|
+-------------+----------------+-----------+-------------+---------+--------------------+--------------------+-------+----+---------+-------+-----+
|          7.4|             0.7|        0.0|          1.9|    0.076|                11|                34| 0.9978|3.51|     0.56|    9.4|  5.0|
|          7.8|            0.88|        0.0|          2.6|    0.098|                25|                67| 0.9968| 3.2|     0.68|    9.8|  5.0|
|          7.8|            0.76|       0.04|          2.3|    0.092|                15|                54|  0.997|3.26|     0.65|    9.8|  5.0|
|         11.2|            0.28|       0.56|          1.9|    0.075|                17|                60|  0.998|3.16|     0.58|    9.8|  6.0|
|          7.4|             0.7|        0.0|          1.9|    0.076|                11|                34| 0.9978|3.51|     0.56|    9.4|  5.0|
+-------------+----------------+-----------+-------------+---------+--------------------+--------------------+-------+----+---------+-------+-----+
only showing top 5 rows


 Validation Training Set Metrics
+--------------------------------------------------+-----+----------+
|features                                          |label|prediction|
+--------------------------------------------------+-----+----------+
|[9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5.0  |5.0       |
|[9.8,0.68,3.2,0.9968,25.0,67.0,0.098,2.6,0.0,0.88,7.8]  |5.0  |5.0       |
|[9.8,0.65,3.26,0.997,15.0,54.0,0.092,2.3,0.04,0.76,7.8] |5.0  |5.0       |
|[9.8,0.58,3.16,0.998,17.0,60.0,0.075,1.9,0.56,0.28,11.2]|6.0  |5.0       |
|[9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]  |5.0  |5.0       |
+--------------------------------------------------+-----+----------+
only showing top 5 rows


The accuracy of the model is 0.575
F1: 0.561940707133917 3
```

DOCKER IMPLEMENTATION –

1.Update System Packages:

sudo yum update -y

2.Install Docker:

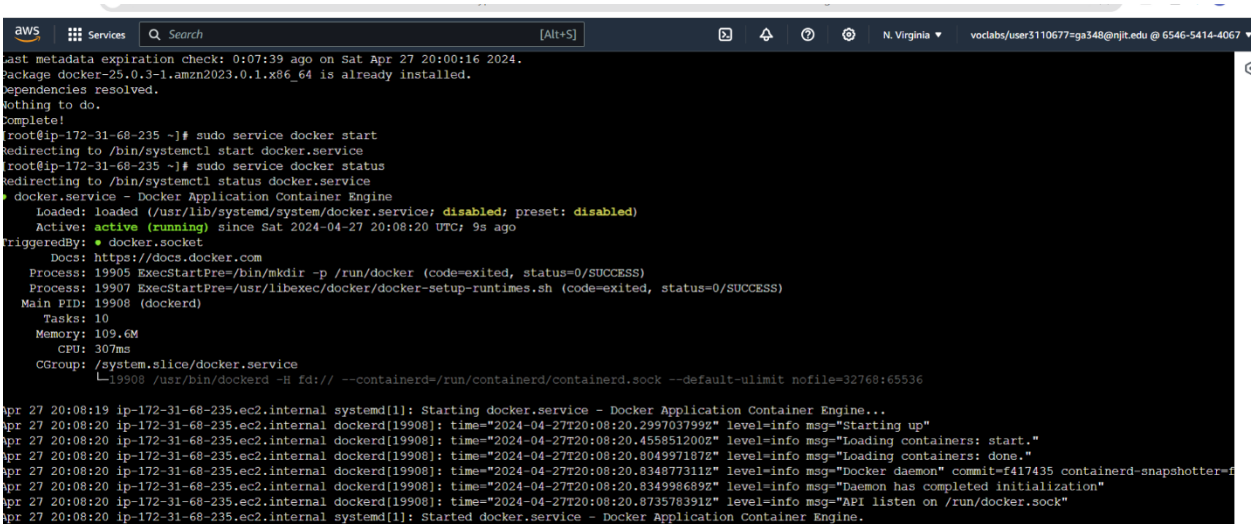
sudo yum install -y docker

3.Start Docker Service:

sudo service docker start

4.Check Docker Service Status:

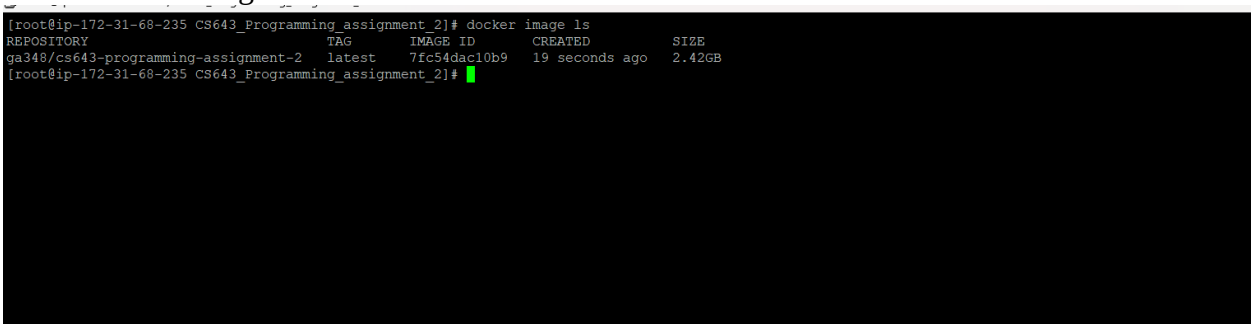
sudo service docker status



5. Create a Dockerfile and build an image with the docker build command.

```
sudo docker build -t ga348/cs643-programming-assignment-2 .
```

6.To check if a Docker image was built, use the following command:
sudo docker image ls



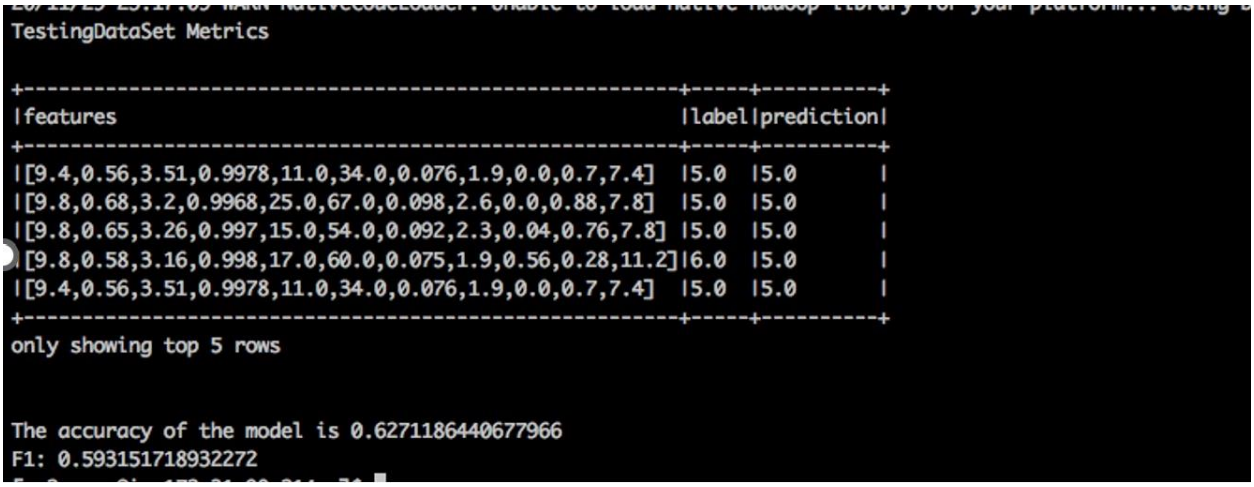
You can see here that your docker image has been built.

7. To run the docker image, use the following command:

sudo docker run -it ga348/cs643-programming-assignment-2

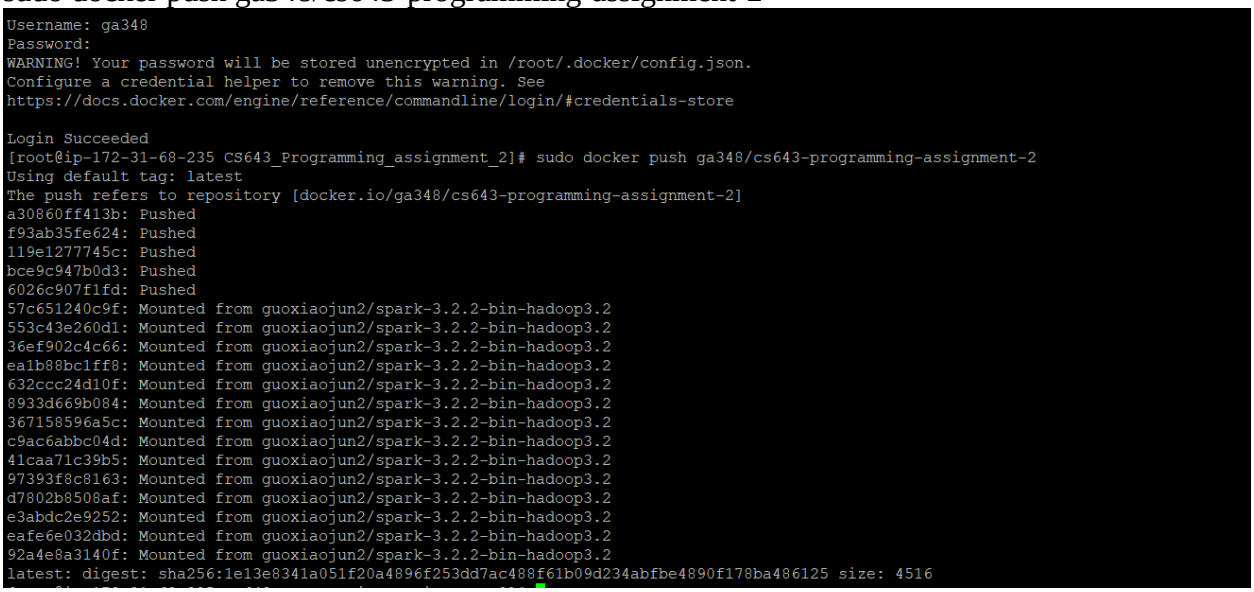
Instead of using the image name, you can use the image ID: # sudo docker run -it

8.This results in the same return for Accuracy and F1 scores.



9. Run the following command to submit the produced Docker image to the DockerHub repository:

sudo docker push ga348/cs643-programming-assignment-2



10.Download and execute the Docker image from the DockerHub repository, following the instructions provided on the website.

Git Bash:

the steps for moving code from my local system to a GitHub repository using Git Bash.

1.git init

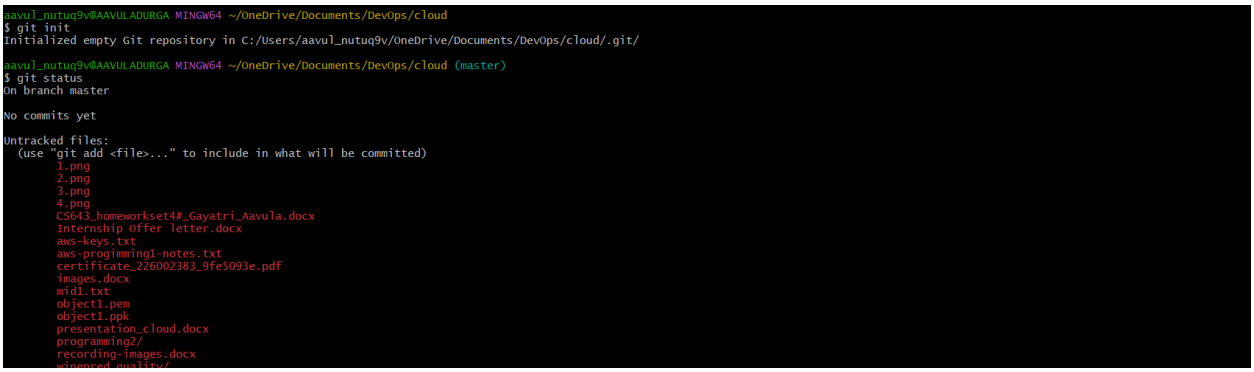
2.git add.

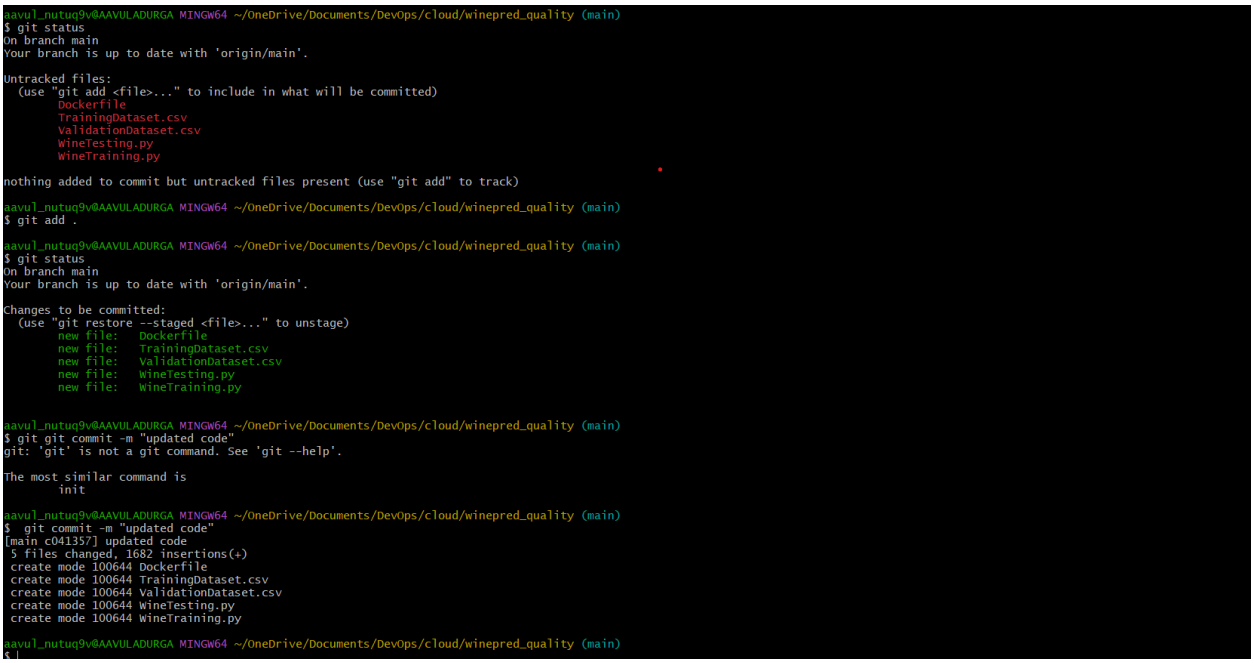
3.git status

4.git commit -m “ updated code”

5.git remote add origin https://github.com/gayatriaavula/winepred_quality.git

6. git push -u origin main





```
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git remote add origin https://github.com/gayatriaavula/winepred_quality.git
error: remote origin already exists.

aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git push -u origin main
Enumerating objects: 8, done.
Counting objects: 100% (8/8), done.
Delta compression using up to 12 threads
Compressing objects: 100% (7/7), done.
Writing objects: 100% (7/7), 24.54 KiB | 2.73 MiB/s, done.
Total 7 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/gayatriaavula/winepred_quality.git
   f8dd3c3..c041357  main -> main
branch 'main' set up to track 'origin/main'.

aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$
```