NJIT

New Jersey Institute
of Technology

# Module 07 Assignment 01: Programming Assignment 2

**Name: Advaithbarath Raghuraman Bhuvaneswari**

**UCID: ar2728**

**GitHub: rbadvaith/Wine_Cluster**

**Docker: advaith123/bucketcluster Tags | Docker Hub**

## Short Summary

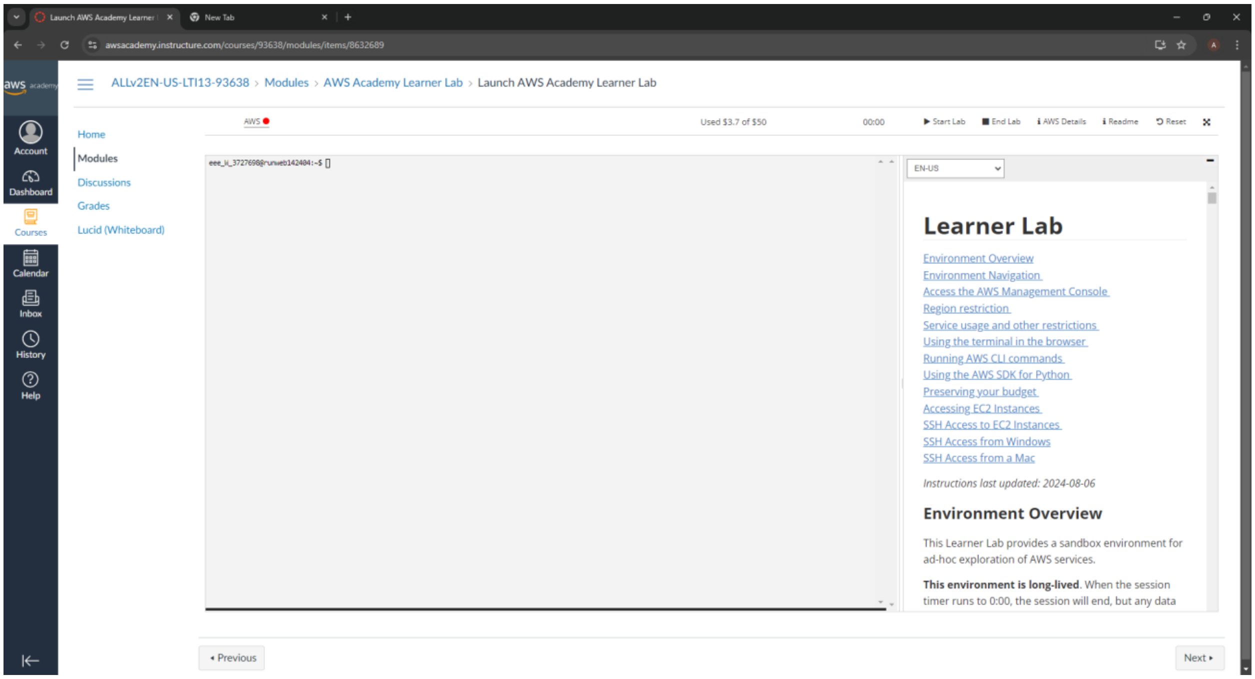
This assignment involves building a machine learning application for predicting wine quality using Apache Spark on the Amazon AWS cloud platform. The goal is to train a model in parallel across multiple EC2 instances and deploy it effectively. Using provided datasets, the application will first train a model on Spark's MLlib in parallel on four EC2 instances, validate its performance using a second dataset, and optimize parameters for the best accuracy. The trained model will then be used on a single EC2 instance for predictions, with its performance measured by the F1 score. Additionally, the prediction application must be containerized using Docker to enable efficient deployment in diverse environments. The implementation is to be completed in Java on an Ubuntu Linux setup, integrating skills in parallel computing, ML model development, and containerization.

## Initial steps for AWS Credentials setup

The credentials created here will be used by our application for connecting to Elasticmapreduce service and S3.

<u>Via Aws Learner(If you are a new user)</u>

1. Create an AWS Account with NJIT email.
2. Go to *Academy Learner lab > Modules>Launch Learner Academy Lab*
3. Learner lab is being displayed here



```
eee_W_3727698@runweb142404:~$ 
```

### Learner Lab

Environment Overview
Environment Navigation
Access the AWS Management Console
Region restriction
Service usage and other restrictions
Using the terminal in the browser
Running AWS CLI commands
Using the AWS SDK for Python
Preserving your budget
Accessing EC2 Instances
SSH Access to EC2 Instances
SSH Access from Windows
SSH Access from a Mac

*Instructions last updated: 2024-08-06*

#### Environment Overview

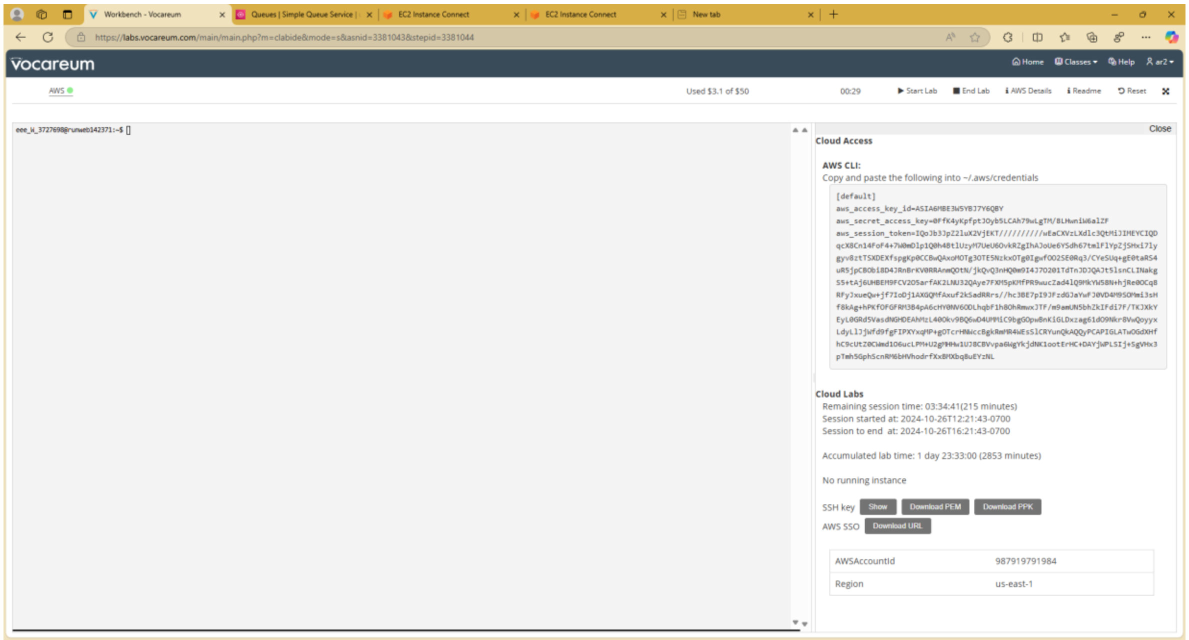
This Learner Lab provides a sandbox environment for ad-hoc exploration of AWS services.

**This environment is long-lived**. When the session timer runs to 0:00, the session will end, but any data

Fig 1. Learner Lab via AWS Academy

Via Vocareum(If you are an existing user | Use Microsoft Edge)

1. If you are an existing user, log in via NJIT email.
2. Select role as student, the search result will display a lab link. Click that to open .
3. Learner lab is being displayed here



```
eee_li_3727698@runweb142371:~$ []
```

**Cloud Access**

**AWS CLI:**

Copy and paste the following into ~/.aws/credentials

```
[default]
aws_access_key_id=ASIA6MBE3W5YBJ7Y6QBY
aws_secret_access_key=0FfK4yKpfptJOyb5LCAh79wLgTM/BLHwniW6alZF
aws_session_token=IQoJb3JpZ2luX2VjEKT//////////wEaCXVzLXdlc3QtMiJIMEYCIQD
qcX8Cn14FoF4+7W0mDlp1Q0h48tlUzyM7UeU6OvkRZgIhAJoUe6YSdh67tmlFlYpZjSHxi7ly
gyv8ztTSXDEXfspgKp0CCBwQAxoMOTg3OTE5NzkxOTg0IgwfOO2SE0Rq3/CYeSUq+gE0taRS4
uR5jpCBObi8D4JRnBrKV0RRAnmQOtN/jkQvQ3nHQ0m9I4J7O201TdTnJDJQAJt5lsnCLINakg
S5+tAj6UHBEM9FCV2O5arfAK2LNU32QAye7FXM5pKMfPR9wucZad4lQ9MkYW58N+hjRe0OCq8
RFyJxueQw+jf7IoDj1AXGQMfAxuf2kSadRRrs//hc3BE7pI9JFzdGJaYwFJ0VD4M9SOMmi3sH
f8kAg+hPKfOFGFRM3B4pA6cHY0NV6ODLhqbF1h8OhRmwxJTF/m9amUN5bhZkIFdi7F/TKJXkY
EyL0GRd5VasdNGHDEAhMzL40Okv9BQ6wD4UMMiC9bgGOpwBnKiGLDxzag61dO9Nkr8VwQoyyx
LdyLlJjWfd9fgFIPXYxqMP+gOTcrHNWccBgkRmMR4WEsSlCRYunQkAQQyPCAPIGLATwOGdXHf
hC9cUtZ0CWmd1O6ucLPM+U2gMHHw1UJ8CBVvpa6WgYkjdNK1ootErHC+DAYjWPLSIj+SgVHx3
pTmh5GphScnRM6bHVhodrfXxBMXbq8uEYzNL
```

**Cloud Labs**

Remaining session time: 03:34:41(215 minutes)
Session started at: 2024-10-26T12:21:43-0700
Session to end at: 2024-10-26T16:21:43-0700

Accumulated lab time: 1 day 23:33:00 (2853 minutes)

No running instance

SSH key

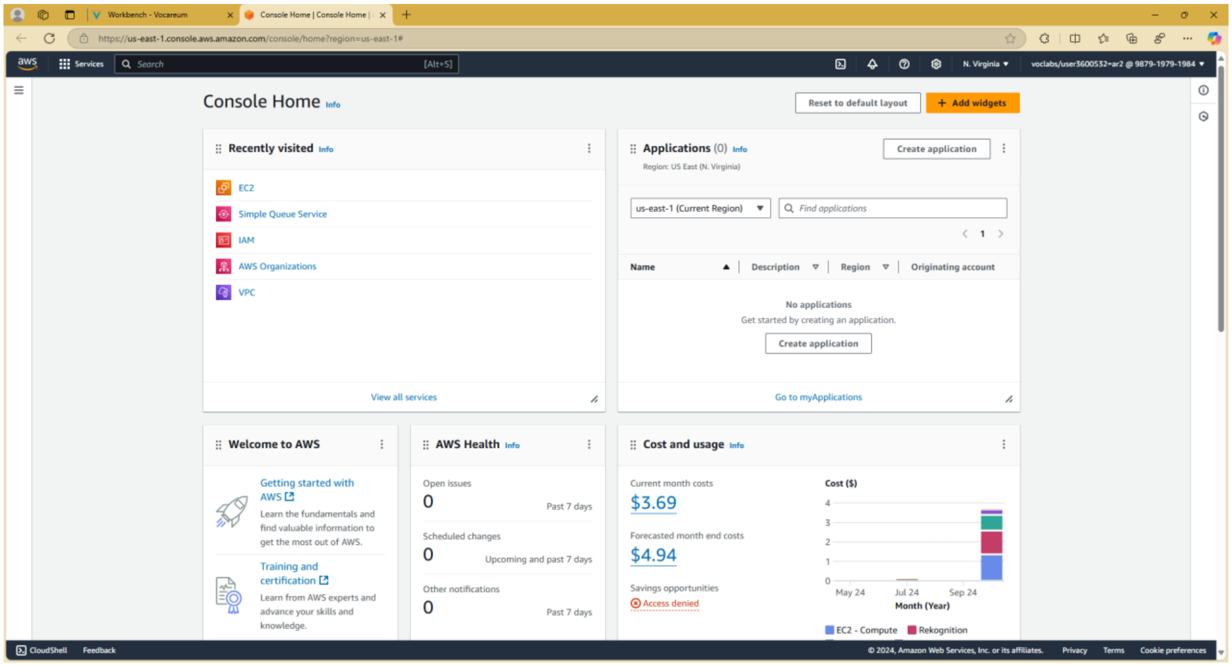
AWS SSO

| | |
|---|---|
| AWSAccountId | 987919791984 |
| Region | us-east-1 |

Fig 2. Learner Lab via AWS Academy

## Launching AWS Management Console

1. Click Start Lab on the right-hand corner
2. Wait for one minute to see the AWS symbol turned from red to green
3. Click the AWS symbol once it turned green
4. AWS Management Console is being displayed. Please note that the session will last for 4 hrs. only



Workbench - Vocareum | Console Home | Console Home |

https://us-east-1.console.aws.amazon.com/console/home?region=us-east-1#

aws Services Search [Alt+S] N. Virginia ▾ voclabs/user3600532=ar2 @ 9879-1979-1984 ▾

Console Home Info

Reset to default layout | + Add widgets

Recently visited Info

- EC2
- Simple Queue Service
- IAM
- AWS Organizations
- VPC

View all services

Applications (0) Info

Region: US East (N. Virginia)

Create application

us-east-1 (Current Region) ▾ | Find applications

| Name ▲ | Description ▽ | Region ▽ | Originating account |
|---|---|---|---|

No applications

Get started by creating an application.

Create application

Go to myApplications

Welcome to AWS

Getting started with AWS

Learn the fundamentals and find valuable information to get the most out of AWS.

Training and certification

Learn from AWS experts and advance your skills and knowledge.

AWS Health Info

Open issues

0 Past 7 days

Scheduled changes

0 Upcoming and past 7 days

Other notifications

0 Past 7 days

Cost and usage Info

Current month costs

$3.69

Forecasted month end costs

$4.94

Savings opportunities

Access denied

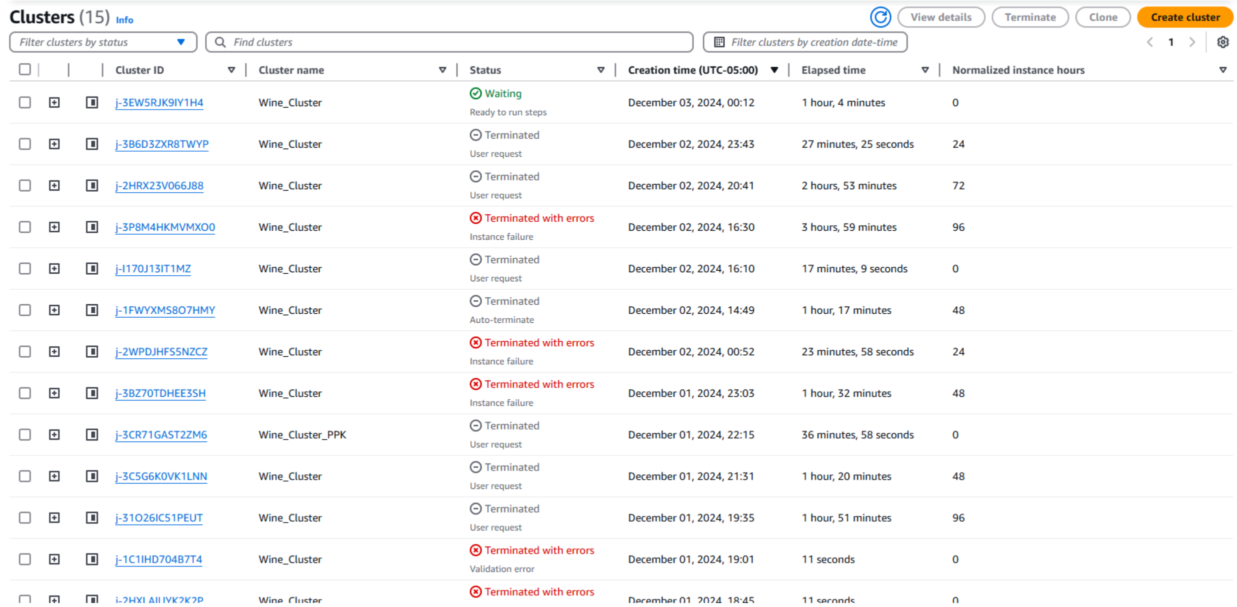
CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Fig 3. AWS Management Console

## Creating EMR Cluster

EMR Cluster 1 core and 4 slaves will be created for the Elasticmapreduce service. These service will only run until the session end. After session finished, you cant able to access the cluster. But don't worry, you can clone the old terminated cluster and re-run it again.

1. In the search bar, type "*EMR*" and then select the first result.
2. Select "*Create cluster*" from the navigation menu on the right. If you want to clone the existing cluster, select your preferred cluster and select "*clone*"



Clusters (15) Info

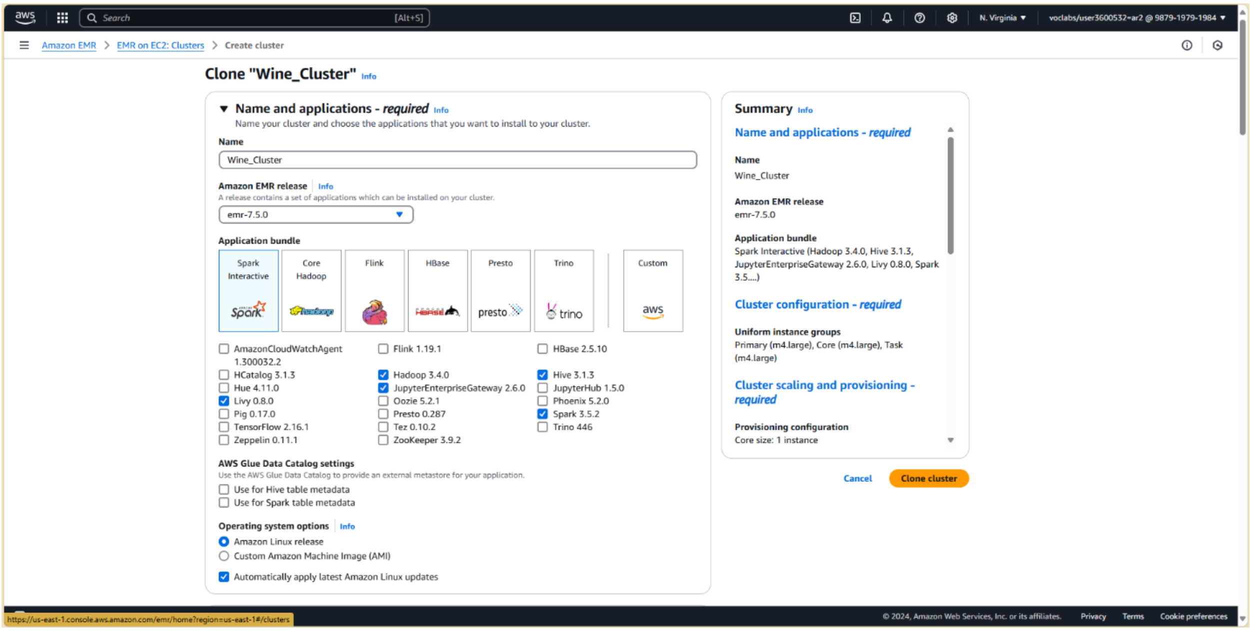
View details | Terminate | Clone | Create cluster

Filter clusters by status ▾ | Find clusters | Filter clusters by creation date-time

| □ | | | Cluster ID | Cluster name | Status | Creation time (UTC-05:00) | Elapsed time | Normalized instance hours |
|---|---|---|---|---|---|---|---|---|
| □ | ⊞ | ◫ | j-3EW5RJK9IY1H4 | Wine_Cluster | Waiting<br>Ready to run steps | December 03, 2024, 00:12 | 1 hour, 4 minutes | 0 |
| □ | ⊞ | ◫ | j-3B6D3ZXR8TWYP | Wine_Cluster | Terminated<br>User request | December 02, 2024, 23:43 | 27 minutes, 25 seconds | 24 |
| □ | ⊞ | ◫ | j-2HRX23V066J88 | Wine_Cluster | Terminated<br>User request | December 02, 2024, 20:41 | 2 hours, 53 minutes | 72 |
| □ | ⊞ | ◫ | j-3P8M4HKMVMXO0 | Wine_Cluster | Terminated with errors<br>Instance failure | December 02, 2024, 16:30 | 3 hours, 59 minutes | 96 |
| □ | ⊞ | ◫ | j-I170J13IT1MZ | Wine_Cluster | Terminated<br>User request | December 02, 2024, 16:10 | 17 minutes, 9 seconds | 0 |
| □ | ⊞ | ◫ | j-1FWYXMS8O7HMY | Wine_Cluster | Terminated<br>Auto-terminate | December 02, 2024, 14:49 | 1 hour, 17 minutes | 48 |
| □ | ⊞ | ◫ | j-2WPDJHFS5NZCZ | Wine_Cluster | Terminated with errors<br>Instance failure | December 02, 2024, 00:52 | 23 minutes, 58 seconds | 24 |
| □ | ⊞ | ◫ | j-3BZ70TDHEE3SH | Wine_Cluster | Terminated with errors<br>Instance failure | December 01, 2024, 23:03 | 1 hour, 32 minutes | 48 |
| □ | ⊞ | ◫ | j-3CR71GAST2ZM6 | Wine_Cluster_PPK | Terminated<br>User request | December 01, 2024, 22:15 | 36 minutes, 58 seconds | 0 |
| □ | ⊞ | ◫ | j-3C5G6K0VK1LNN | Wine_Cluster | Terminated<br>User request | December 01, 2024, 21:31 | 1 hour, 20 minutes | 48 |
| □ | ⊞ | ◫ | j-31O26IC51PEUT | Wine_Cluster | Terminated<br>User request | December 01, 2024, 19:35 | 1 hour, 51 minutes | 96 |
| □ | ⊞ | ◫ | j-1C1IHD704B7T4 | Wine_Cluster | Terminated with errors<br>Validation error | December 01, 2024, 19:01 | 11 seconds | 0 |
| □ | ⊞ | ◫ | j-2HXLAIUYK2K2P | Wine_Cluster | Terminated with errors | December 01, 2024, 18:45 | 11 seconds | 0 |

Fig 4. Create Cluster

3. In *Name*: Type Wine_Cluster
4. Ensure that the application bundle should be the same as in Fig.5


aws
Search [Alt+S]
N. Virginia ▾
voclabs/user3600532=ar2 @ 9879-1979-1984 ▾
Amazon EMR > EMR on EC2: Clusters > Create cluster
Info
Cancel
Clone cluster
© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences
https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusters

**Clone "Wine_Cluster"**

▼ **Name and applications - *required***

Name your cluster and choose the applications that you want to install to your cluster.

**Name**

Wine_Cluster

**Amazon EMR release**

A release contains a set of applications which can be installed on your cluster.

emr-7.5.0

**Application bundle**

Spark Interactive | Core Hadoop | Flink | HBase | Presto | Trino | Custom

- ☐ AmazonCloudWatchAgent 1.300032.2
- ☐ HCatalog 3.1.3
- ☐ Hue 4.11.0
- ☑ Livy 0.8.0
- ☐ Pig 0.17.0
- ☐ TensorFlow 2.16.1
- ☐ Zeppelin 0.11.1
- ☐ Flink 1.19.1
- ☑ Hadoop 3.4.0
- ☑ JupyterEnterpriseGateway 2.6.0
- ☐ Oozie 5.2.1
- ☐ Presto 0.287
- ☐ Tez 0.10.2
- ☐ ZooKeeper 3.9.2
- ☐ HBase 2.5.10
- ☑ Hive 3.1.3
- ☐ JupyterHub 1.5.0
- ☐ Phoenix 5.2.0
- ☑ Spark 3.5.2
- ☐ Trino 446

**AWS Glue Data Catalog settings**

Use the AWS Glue Data Catalog to provide an external metastore for your application.

- ☐ Use for Hive table metadata
- ☐ Use for Spark table metadata

**Operating system options**

- ● Amazon Linux release
- ○ Custom Amazon Machine Image (AMI)

☑ Automatically apply latest Amazon Linux updates

**Summary**

**Name and applications - *required***

**Name**
Wine_Cluster

**Amazon EMR release**
emr-7.5.0

**Application bundle**
Spark Interactive (Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)

**Cluster configuration - *required***

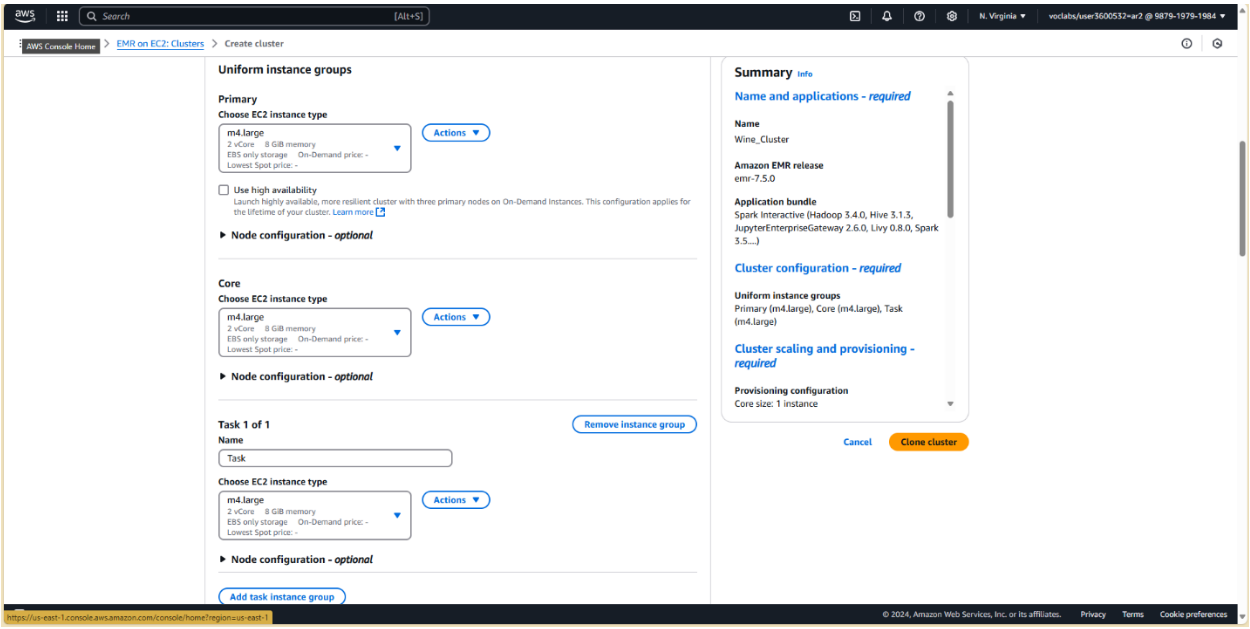
**Uniform instance groups**
Primary (m4.large), Core (m4.large), Task (m4.large)

**Cluster scaling and provisioning - *required***

**Provisioning configuration**
Core size: 1 instance

Fig 5. Clone or Create Cluster

5. Select instance type as m4large for all


aws
Search [Alt+S]
N. Virginia ▾
voclabs/user3600532=ar2 @ 9879-1979-1984 ▾
AWS Console Home > EMR on EC2: Clusters > Create cluster
Actions ▼
Remove instance group
Add task instance group
Cancel
Clone cluster
© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences
https://us-east-1.console.aws.amazon.com/console/home?region=us-east-1

**Uniform instance groups**

**Primary**

**Choose EC2 instance type**

m4.large
2 vCore 8 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

☐ Use high availability
Launch highly available, more resilient cluster with three primary nodes on On-Demand Instances. This configuration applies for the lifetime of your cluster. Learn more

▶ **Node configuration - *optional***

**Core**

**Choose EC2 instance type**

m4.large
2 vCore 8 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

▶ **Node configuration - *optional***

**Task 1 of 1**

**Name**

Task

**Choose EC2 instance type**

m4.large
2 vCore 8 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

▶ **Node configuration - *optional***

**Summary**

**Name and applications - *required***

**Name**
Wine_Cluster

**Amazon EMR release**
emr-7.5.0

**Application bundle**
Spark Interactive (Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)

**Cluster configuration - *required***

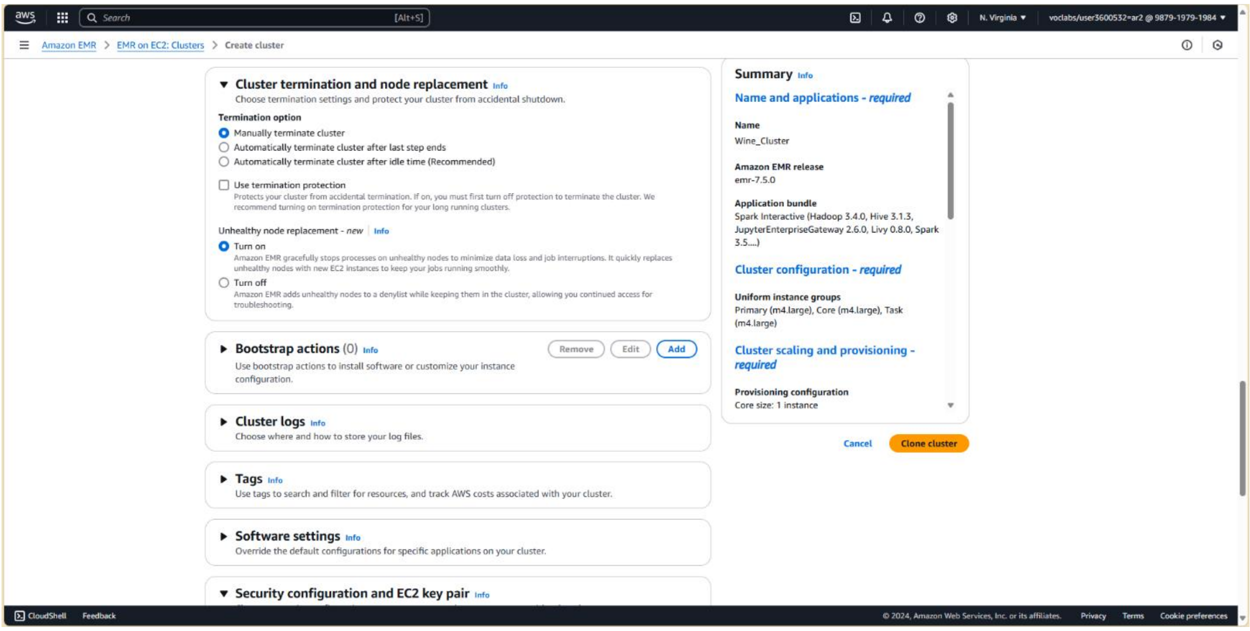
**Uniform instance groups**
Primary (m4.large), Core (m4.large), Task (m4.large)

**Cluster scaling and provisioning - *required***

**Provisioning configuration**
Core size: 1 instance

Fig 6. Instance Type

6. In Cluster Termination, set termination to manual from automatic


aws
Search [Alt+S]
N. Virginia
voclabs/user3600532=ar2 @ 9879-1979-1984
Amazon EMR > EMR on EC2: Clusters > Create cluster
Info
Remove
Edit
Add
Cancel
Clone cluster
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

▼ **Cluster termination and node replacement**

Choose termination settings and protect your cluster from accidental shutdown.

**Termination option**

- ● Manually terminate cluster
- ○ Automatically terminate cluster after last step ends
- ○ Automatically terminate cluster after idle time (Recommended)

□ Use termination protection
Protects your cluster from accidental termination. If on, you must first turn off protection to terminate the cluster. We recommend turning on termination protection for your long running clusters.

Unhealthy node replacement - *new*

- ● Turn on
  Amazon EMR gracefully stops processes on unhealthy nodes to minimize data loss and job interruptions. It quickly replaces unhealthy nodes with new EC2 instances to keep your jobs running smoothly.
- ○ Turn off
  Amazon EMR adds unhealthy nodes to a denylist while keeping them in the cluster, allowing you continued access for troubleshooting.

▶ **Bootstrap actions** (0)
Use bootstrap actions to install software or customize your instance configuration.

▶ **Cluster logs**
Choose where and how to store your log files.

▶ **Tags**
Use tags to search and filter for resources, and track AWS costs associated with your cluster.

▶ **Software settings**
Override the default configurations for specific applications on your cluster.

▼ **Security configuration and EC2 key pair**

**Summary**

Name and applications - *required*

**Name**
Wine_Cluster

**Amazon EMR release**
emr-7.5.0

**Application bundle**
Spark Interactive (Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)

Cluster configuration - *required*

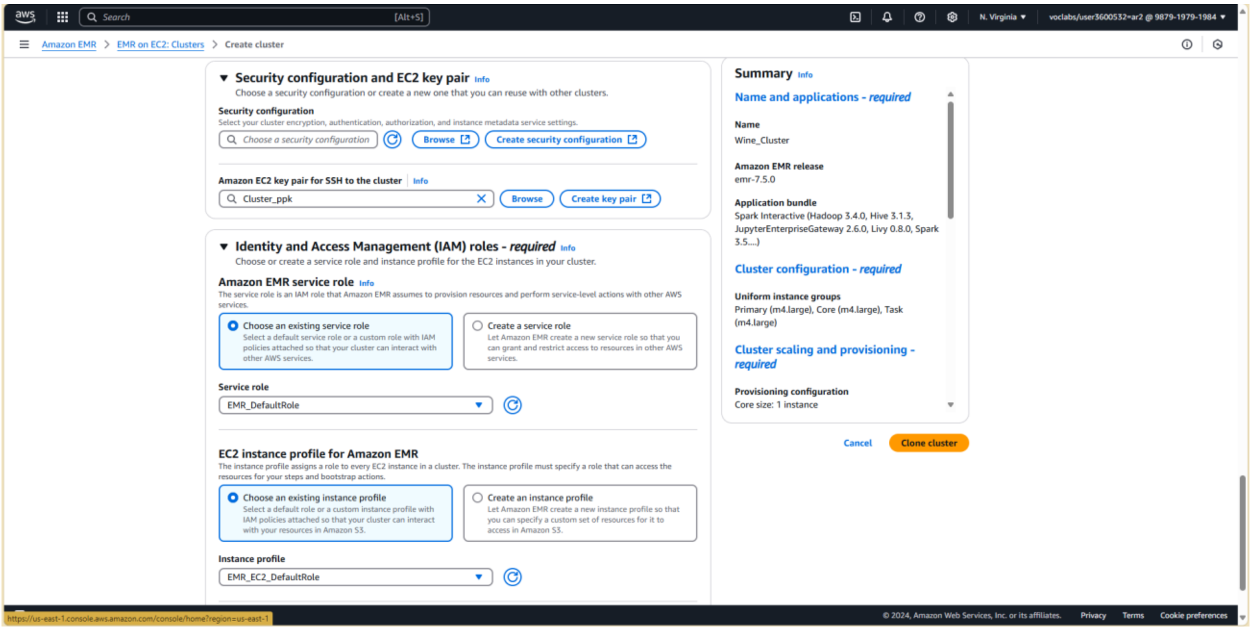
**Uniform instance groups**
Primary (m4.large), Core (m4.large), Task (m4.large)

Cluster scaling and provisioning - *required*

**Provisioning configuration**
Core size: 1 instance

Fig 7. Cluster Termination

7. Set EC2 Private key and name it as cluster.ppk


aws
Search [Alt+S]
N. Virginia
voclabs/user3600532=ar2 @ 9879-1979-1984
Amazon EMR > EMR on EC2: Clusters > Create cluster
Info
Browse
Create security configuration
Create key pair
Cancel
Clone cluster
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
https://us-east-1.console.aws.amazon.com/console/home?region=us-east-1

▼ **Security configuration and EC2 key pair**

Choose a security configuration or create a new one that you can reuse with other clusters.

**Security configuration**
Select your cluster encryption, authentication, authorization, and instance metadata service settings.
Choose a security configuration

**Amazon EC2 key pair for SSH to the cluster**
Cluster_ppk

▼ **Identity and Access Management (IAM) roles - *required***

Choose or create a service role and instance profile for the EC2 instances in your cluster.

**Amazon EMR service role**
The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

- ● Choose an existing service role
  Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.
- ○ Create a service role
  Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

**Service role**
EMR_DefaultRole

**EC2 instance profile for Amazon EMR**
The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

- ● Choose an existing instance profile
  Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.
- ○ Create an instance profile
  Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

**Instance profile**
EMR_EC2_DefaultRole

**Summary**

Name and applications - *required*

**Name**
Wine_Cluster

**Amazon EMR release**
emr-7.5.0

**Application bundle**
Spark Interactive (Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)

Cluster configuration - *required*

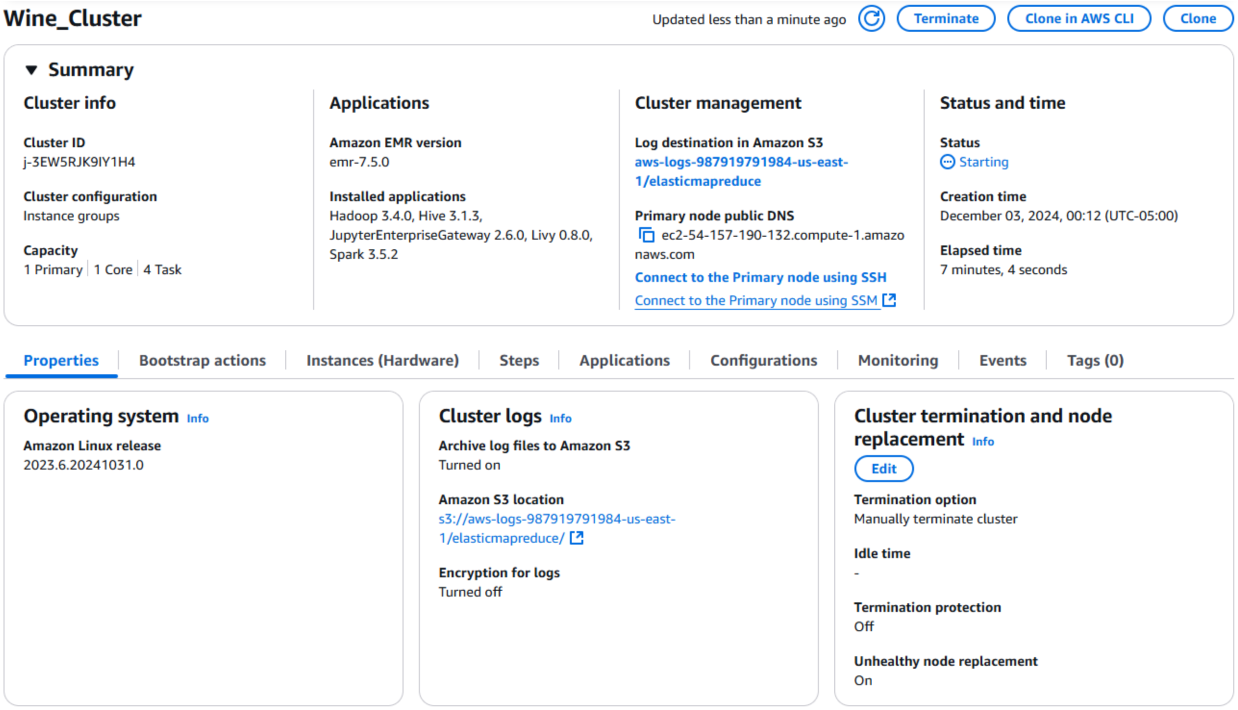
**Uniform instance groups**
Primary (m4.large), Core (m4.large), Task (m4.large)

Cluster scaling and provisioning - *required*

**Provisioning configuration**
Core size: 1 instance

Fig 8. EC2 Key Pair

8. Select Create Cluster or Clone Cluster and wait for 5 minutes to load



## Wine_Cluster

Updated less than a minute ago | Terminate | Clone in AWS CLI | Clone

▼ Summary

**Cluster info**

**Cluster ID**
j-3EW5RJK9IY1H4

**Cluster configuration**
Instance groups

**Capacity**
1 Primary | 1 Core | 4 Task

**Applications**

**Amazon EMR version**
emr-7.5.0

**Installed applications**
Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.2

**Cluster management**

**Log destination in Amazon S3**
aws-logs-987919791984-us-east-1/elasticmapreduce

**Primary node public DNS**
ec2-54-157-190-132.compute-1.amazonaws.com

Connect to the Primary node using SSH
Connect to the Primary node using SSM

**Status and time**

**Status**
Starting

**Creation time**
December 03, 2024, 00:12 (UTC-05:00)

**Elapsed time**
7 minutes, 4 seconds

**Operating system** Info

**Amazon Linux release**
2023.6.20241031.0

**Cluster logs** Info

**Archive log files to Amazon S3**
Turned on

**Amazon S3 location**
s3://aws-logs-987919791984-us-east-1/elasticmapreduce/

**Encryption for logs**
Turned off

**Cluster termination and node replacement** Info

Edit

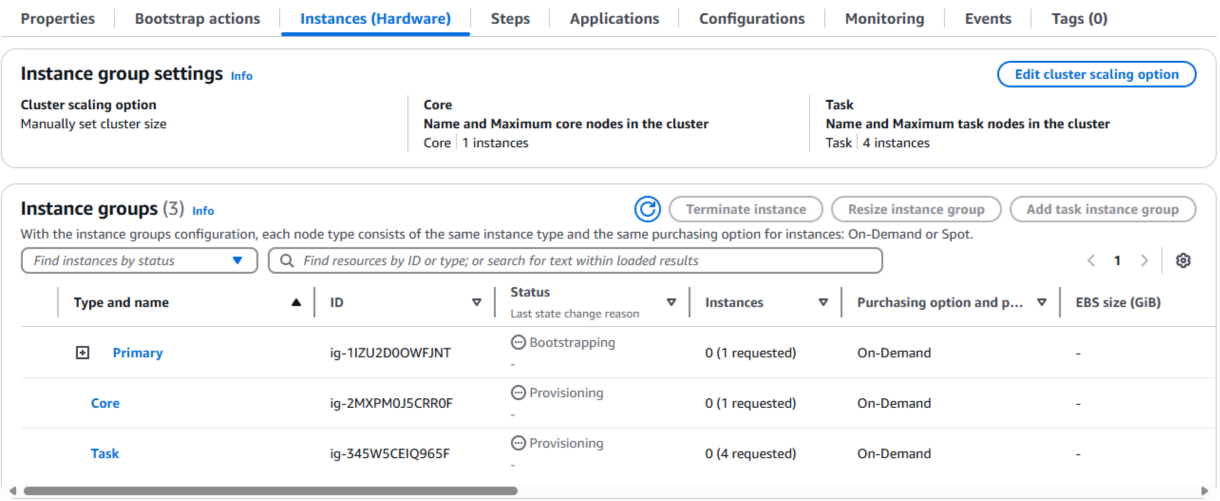
**Termination option**
Manually terminate cluster

**Idle time**
-

**Termination protection**
Off

**Unhealthy node replacement**
On

Fig 9. Wine_Cluster



**Instance group settings** Info

Edit cluster scaling option

**Cluster scaling option**
Manually set cluster size

**Core**
**Name and Maximum core nodes in the cluster**
Core | 1 instances

**Task**
**Name and Maximum task nodes in the cluster**
Task | 4 instances

**Instance groups** (3) Info

With the instance groups configuration, each node type consists of the same instance type and the same purchasing option for instances: On-Demand or Spot.

Find instances by status
Find resources by ID or type; or search for text within loaded results
1

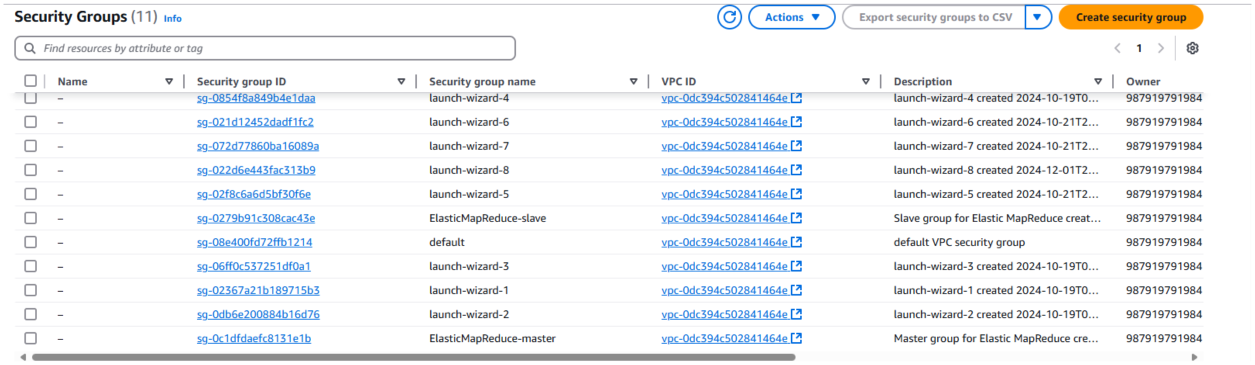
| Type and name | ID | Status<br>Last state change reason | Instances | Purchasing option and p... | EBS size (GiB) |
|---|---|---|---|---|---|
| ⊞ Primary | ig-1IZU2D0OWFJNT | Bootstrapping<br>- | 0 (1 requested) | On-Demand | - |
| Core | ig-2MXPM0J5CRR0F | Provisioning<br>- | 0 (1 requested) | On-Demand | - |
| Task | ig-345W5CEIQ965F | Provisioning<br>- | 0 (4 requested) | On-Demand | - |

Fig 10. Number of Instances

## Setting SSH for master(core) EC2 in the cluster[Only required to do this for first time]

We need to provide ssh access for the master instance which is not provided when initially created.

1. Search EC2 in the search bar and click enter
2. In the Dashboard select security group
3. Over there Select the group id that have group name elasticmapreduce master in it

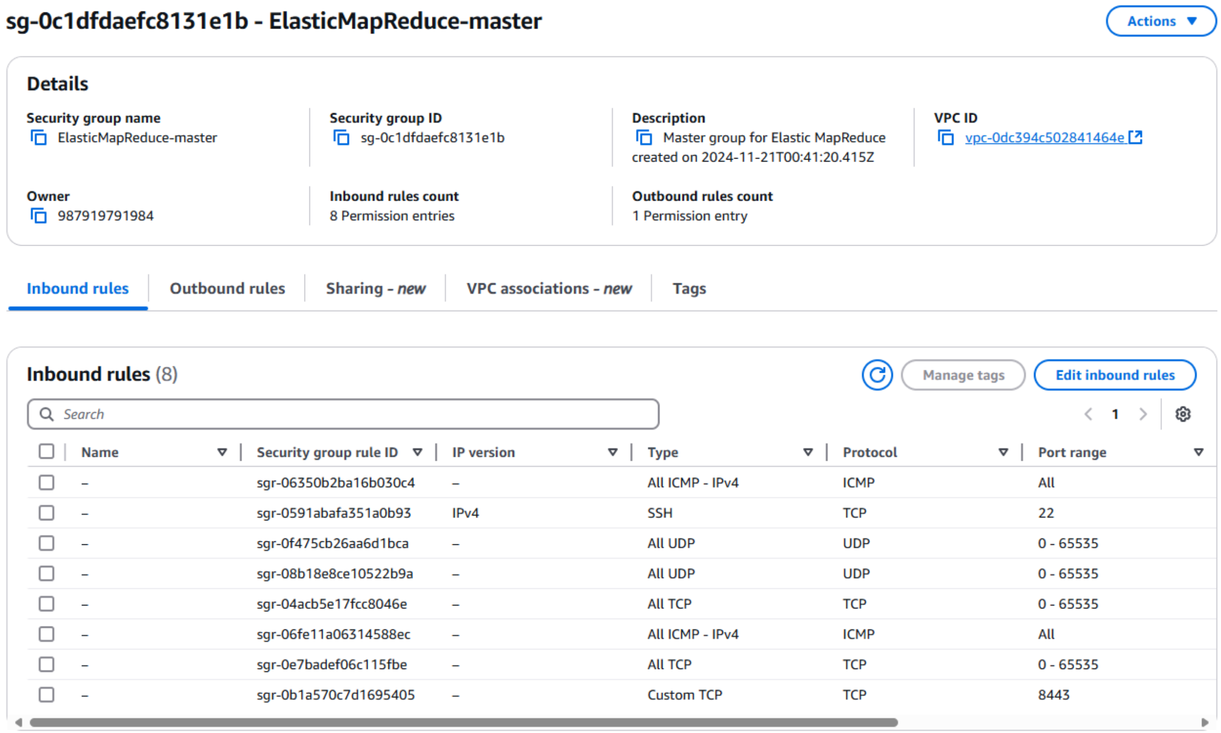


Security Groups (11) Info
Actions ▼
Export security groups to CSV ▼
Create security group
Find resources by attribute or tag
< 1 >

| □ | Name | Security group ID | Security group name | VPC ID | Description | Owner |
|---|---|---|---|---|---|---|
| □ | – | sg-0854f8a849b4e1daa | launch-wizard-4 | vpc-0dc394c502841464e | launch-wizard-4 created 2024-10-19T0… | 987919791984 |
| □ | – | sg-021d12452dadf1fc2 | launch-wizard-6 | vpc-0dc394c502841464e | launch-wizard-6 created 2024-10-21T2… | 987919791984 |
| □ | – | sg-072d77860ba16089a | launch-wizard-7 | vpc-0dc394c502841464e | launch-wizard-7 created 2024-10-21T2… | 987919791984 |
| □ | – | sg-022d6e443fac313b9 | launch-wizard-8 | vpc-0dc394c502841464e | launch-wizard-8 created 2024-12-01T2… | 987919791984 |
| □ | – | sg-02f8c6a6d5bf30f6e | launch-wizard-5 | vpc-0dc394c502841464e | launch-wizard-5 created 2024-10-21T2… | 987919791984 |
| □ | – | sg-0279b91c308cac43e | ElasticMapReduce-slave | vpc-0dc394c502841464e | Slave group for Elastic MapReduce creat… | 987919791984 |
| □ | – | sg-08e400fd72ffb1214 | default | vpc-0dc394c502841464e | default VPC security group | 987919791984 |
| □ | – | sg-06ff0c537251df0a1 | launch-wizard-3 | vpc-0dc394c502841464e | launch-wizard-3 created 2024-10-19T0… | 987919791984 |
| □ | – | sg-02367a21b189715b3 | launch-wizard-1 | vpc-0dc394c502841464e | launch-wizard-1 created 2024-10-19T0… | 987919791984 |
| □ | – | sg-0db6e200884b16d76 | launch-wizard-2 | vpc-0dc394c502841464e | launch-wizard-2 created 2024-10-19T0… | 987919791984 |
| □ | – | sg-0c1dfdaefc8131e1b | ElasticMapReduce-master | vpc-0dc394c502841464e | Master group for Elastic MapReduce cre… | 987919791984 |

Fig 11. Security Group Name

4. Select edit inbound rules in the middle right edge



Details

Security group name
ElasticMapReduce-master

Security group ID
sg-0c1dfdaefc8131e1b

Description
Master group for Elastic MapReduce created on 2024-11-21T00:41:20.415Z

VPC ID
vpc-0dc394c502841464e

Owner
987919791984

Inbound rules count
8 Permission entries

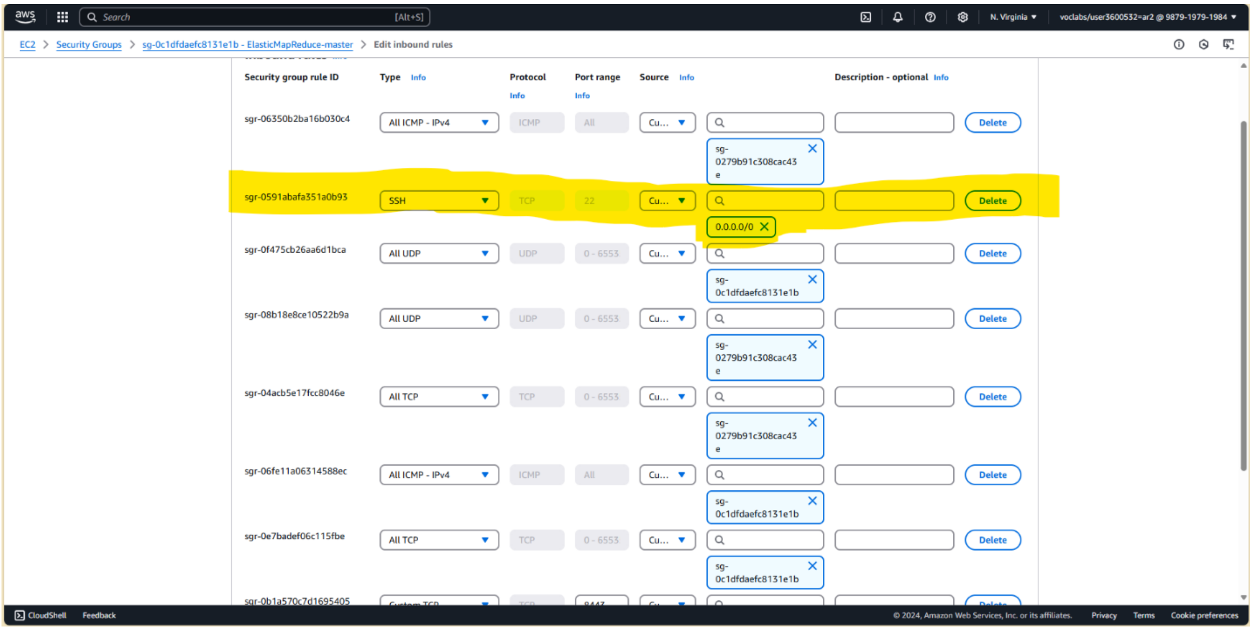
Outbound rules count
1 Permission entry

Inbound rules (8)

| □ | Name | Security group rule ID | IP version | Type | Protocol | Port range |
|---|---|---|---|---|---|---|
| □ | – | sgr-06350b2ba16b030c4 | – | All ICMP - IPv4 | ICMP | All |
| □ | – | sgr-0591abafa351a0b93 | IPv4 | SSH | TCP | 22 |
| □ | – | sgr-0f475cb26aa6d1bca | – | All UDP | UDP | 0 - 65535 |
| □ | – | sgr-08b18e8ce10522b9a | – | All UDP | UDP | 0 - 65535 |
| □ | – | sgr-04acb5e17fcc8046e | – | All TCP | TCP | 0 - 65535 |
| □ | – | sgr-06fe11a06314588ec | – | All ICMP - IPv4 | ICMP | All |
| □ | – | sgr-0e7badef06c115fbe | – | All TCP | TCP | 0 - 65535 |
| □ | – | sgr-0b1a570c7d1695405 | – | Custom TCP | TCP | 8443 |

Fig 12. ElasticMapReduce-master

5. Verify if ssh is present there else, create a new one



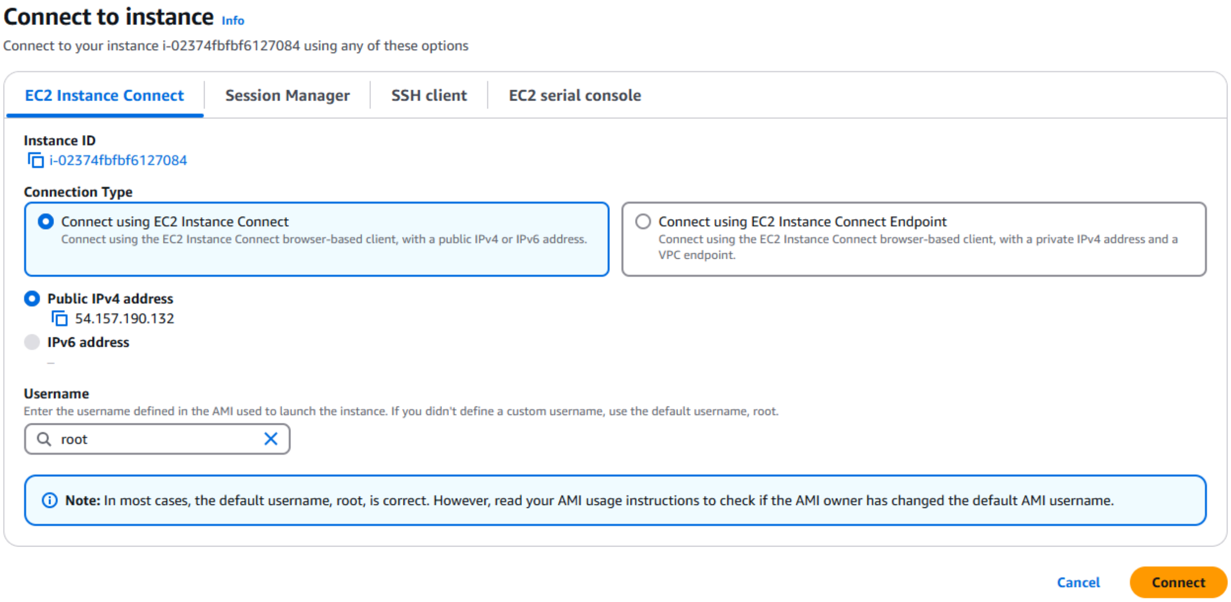
EC2 > Security Groups > sg-0c1dfdaefc8131e1b - ElasticMapReduce-master > Edit inbound rules

| Security group rule ID | Type | Protocol | Port range | Source | | Description - optional |
|---|---|---|---|---|---|---|
| sgr-06350b2ba16b030c4 | All ICMP - IPv4 | ICMP | All | Cu... | sg-0279b91c308cac43e | |
| sgr-0591abafa351a0b93 | SSH | TCP | 22 | Cu... | 0.0.0.0/0 | |
| sgr-0f475cb26aa6d1bca | All UDP | UDP | 0 - 6553 | Cu... | sg-0c1dfdaefc8131e1b | |
| sgr-08b18e8ce10522b9a | All UDP | UDP | 0 - 6553 | Cu... | sg-0279b91c308cac43e | |
| sgr-04acb5e17fcc8046e | All TCP | TCP | 0 - 6553 | Cu... | sg-0279b91c308cac43e | |
| sgr-06fe11a06314588ec | All ICMP - IPv4 | ICMP | All | Cu... | sg-0c1dfdaefc8131e1b | |
| sgr-0e7badef06c115fbe | All TCP | TCP | 0 - 6553 | Cu... | sg-0c1dfdaefc8131e1b | |
| sgr-0b1a570c7d1695405 | Custom TCP | TCP | 8443 | Cu... | | |

aws
Search [Alt+S]
N. Virginia
voclabs/user3600532=ar2 @ 9879-1979-1984
Info
Delete
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Fig 13. SSH in Inbound rules

6. Now you could able to access the master instance, now to EMR menu, select connect via ssm, go to EC2 connect tab and click connect



## Connect to instance

Connect to your instance i-02374fbfbf6127084 using any of these options

**Instance ID**
i-02374fbfbf6127084

**Connection Type**

● Connect using EC2 Instance Connect
Connect using the EC2 Instance Connect browser-based client, with a public IPv4 or IPv6 address.

○ Connect using EC2 Instance Connect Endpoint
Connect using the EC2 Instance Connect browser-based client, with a private IPv4 address and a VPC endpoint.

● Public IPv4 address
54.157.190.132

○ IPv6 address
–

**Username**
Enter the username defined in the AMI used to launch the instance. If you didn't define a custom username, use the default username, root.

root

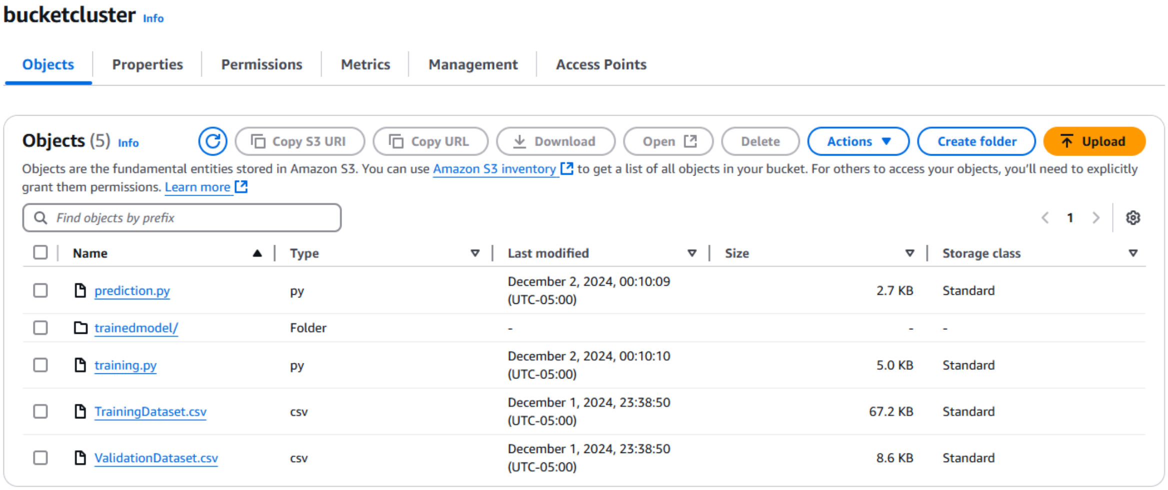
**Note:** In most cases, the default username, root, is correct. However, read your AMI usage instructions to check if the AMI owner has changed the default AMI username.

Fig 14. Connect to EC2

## S3 Bucket

We are going to store the code and csv file in S3 bucket to sync them and use it in EMR Cluster

1. In the search bar, type S3 and select create bucket.
2. Give name *buckercluster* to it
3. After creating it, *upload training.py, TrainingDataset.csv, prediction.py* and *ValidationDataset.csv* to this bucket
4. Check whether is these file present as given in below figure. Pls note that *trainedmodel* folder is also present in the bucket which is created only after training.py have finished execution in the EMR Cluster



bucketcluster Info

Objects (5) Info

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 objects, you'll need to explicitly grant them permissions. Learn more

Find objects by prefix

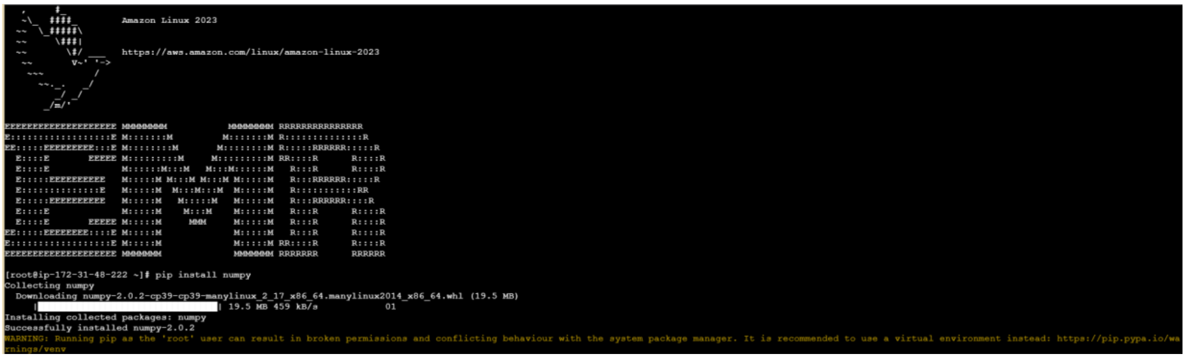
| □ | Name ▲ | Type | Last modified | Size | Storage class |
|---|---|---|---|---|---|
| □ | prediction.py | py | December 2, 2024, 00:10:09 (UTC-05:00) | 2.7 KB | Standard |
| □ | trainedmodel/ | Folder | - | - | - |
| □ | training.py | py | December 2, 2024, 00:10:10 (UTC-05:00) | 5.0 KB | Standard |
| □ | TrainingDataset.csv | csv | December 1, 2024, 23:38:50 (UTC-05:00) | 67.2 KB | Standard |
| □ | ValidationDataset.csv | csv | December 1, 2024, 23:38:50 (UTC-05:00) | 8.6 KB | Standard |

Fig 15. Bucketcluster

## Execution without Docker

We are going to execute both training.py and prediction.py without docker

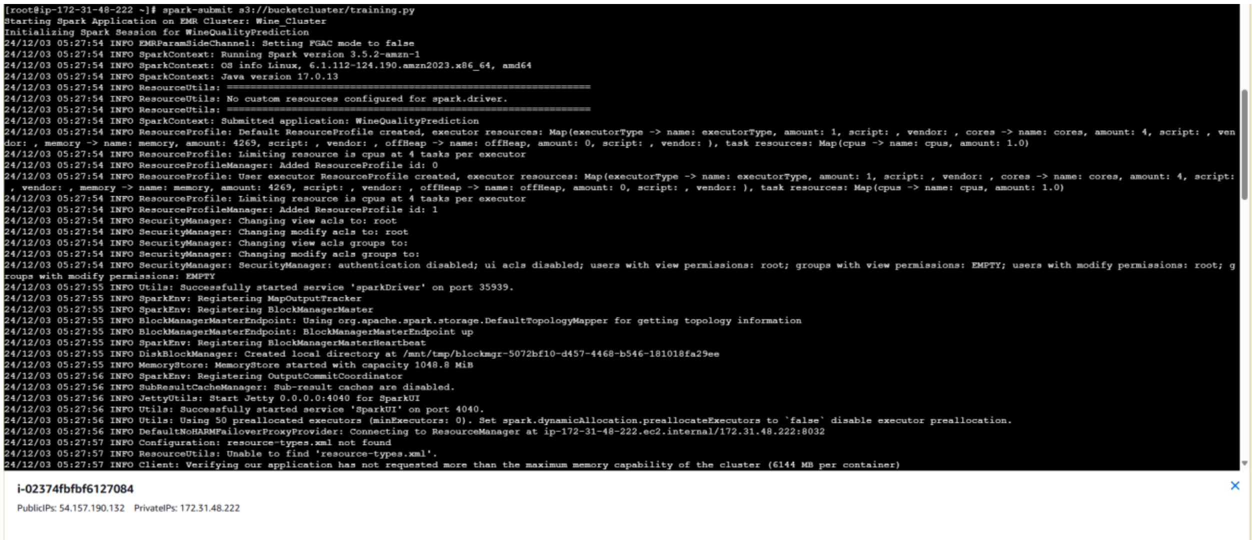
1. Go to EC2 Instance
2. Install numpy via cmd "***pip install numpy***"



```
Amazon Linux 2023

https://aws.amazon.com/linux/amazon-linux-2023

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

[root@ip-172-31-48-222 ~]# pip install numpy
Collecting numpy
  Downloading numpy-2.0.2-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (19.5 MB)
     |████████████████████████████████| 19.5 MB 459 kB/s            01
Installing collected packages: numpy
Successfully installed numpy-2.0.2
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/wa
rnings/venv
```

Fig 16. Install numpy

3. Now Run training.py via command “***spark-submit s3://bucketcluster/training.py***” and wait until execution get finished

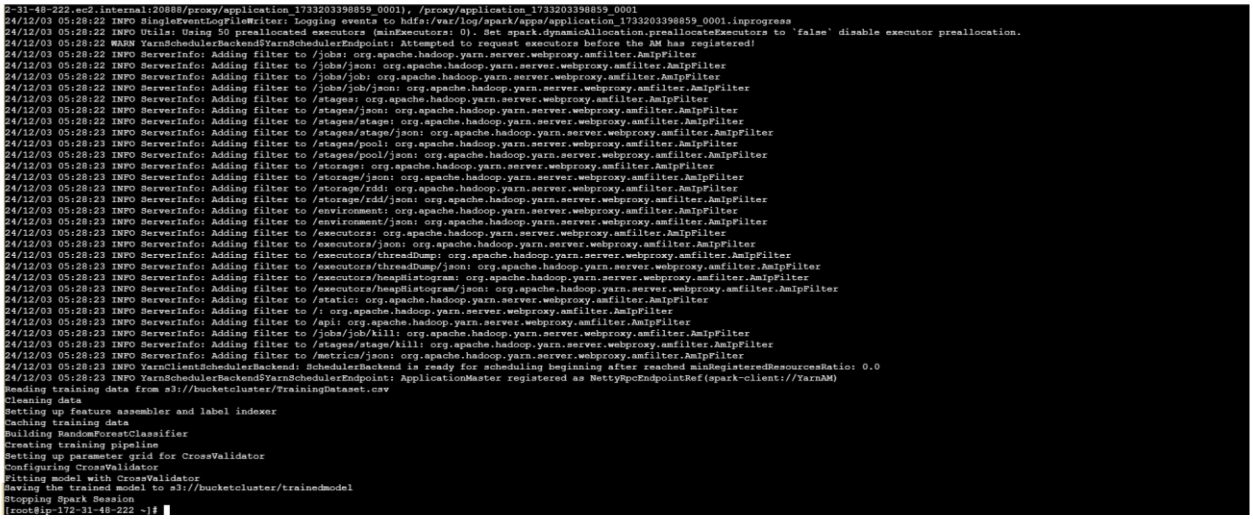


```
[root@ip-172-31-48-222 ~]# spark-submit s3://bucketcluster/training.py
Starting Spark Application on EMR Cluster: Wine_Cluster
Initializing Spark Session for WineQualityPrediction
24/12/03 05:27:54 INFO EMRParamSideChannel: Setting FGAC mode to false
24/12/03 05:27:54 INFO SparkContext: Running Spark version 3.5.2-amzn-1
24/12/03 05:27:54 INFO SparkContext: OS info Linux, 6.1.112-124.190.amzn2023.x86_64, amd64
24/12/03 05:27:54 INFO SparkContext: Java version 17.0.13
24/12/03 05:27:54 INFO ResourceUtils: ==============================================================
24/12/03 05:27:54 INFO ResourceUtils: No custom resources configured for spark.driver.
24/12/03 05:27:54 INFO ResourceUtils: ==============================================================
24/12/03 05:27:54 INFO SparkContext: Submitted application: WineQualityPrediction
24/12/03 05:27:54 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(executorType -> name: executorType, amount: 1, script: , vendor: , cores -> name: cores, amount: 4, script: , ven
dor: , memory -> name: memory, amount: 4269, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/12/03 05:27:54 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/12/03 05:27:54 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/12/03 05:27:54 INFO ResourceProfile: User executor ResourceProfile created, executor resources: Map(executorType -> name: executorType, amount: 1, script: , vendor: , cores -> name: cores, amount: 4, script:
 , vendor: , memory -> name: memory, amount: 4269, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/12/03 05:27:54 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/12/03 05:27:54 INFO ResourceProfileManager: Added ResourceProfile id: 1
24/12/03 05:27:54 INFO SecurityManager: Changing view acls to: root
24/12/03 05:27:54 INFO SecurityManager: Changing modify acls to: root
24/12/03 05:27:54 INFO SecurityManager: Changing view acls groups to:
24/12/03 05:27:54 INFO SecurityManager: Changing modify acls groups to:
24/12/03 05:27:54 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: root; groups with view permissions: EMPTY; users with modify permissions: root; g
roups with modify permissions: EMPTY
24/12/03 05:27:55 INFO Utils: Successfully started service 'sparkDriver' on port 35939.
24/12/03 05:27:55 INFO SparkEnv: Registering MapOutputTracker
24/12/03 05:27:55 INFO SparkEnv: Registering BlockManagerMaster
24/12/03 05:27:55 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/12/03 05:27:55 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/12/03 05:27:55 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/12/03 05:27:55 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-5072bf10-d457-4468-b546-181018fa29ee
24/12/03 05:27:55 INFO MemoryStore: MemoryStore started with capacity 1048.8 MiB
24/12/03 05:27:56 INFO SparkEnv: Registering OutputCommitCoordinator
24/12/03 05:27:56 INFO SubResultCacheManager: Sub-result caches are disabled.
24/12/03 05:27:56 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/12/03 05:27:56 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/12/03 05:27:56 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
24/12/03 05:27:56 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-48-222.ec2.internal/172.31.48.222:8032
24/12/03 05:27:57 INFO Configuration: resource-types.xml not found
24/12/03 05:27:57 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/12/03 05:27:57 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (6144 MB per container)
```

i-02374fbfbf6127084

PublicIPs: 54.157.190.132 PrivateIPs: 172.31.48.222

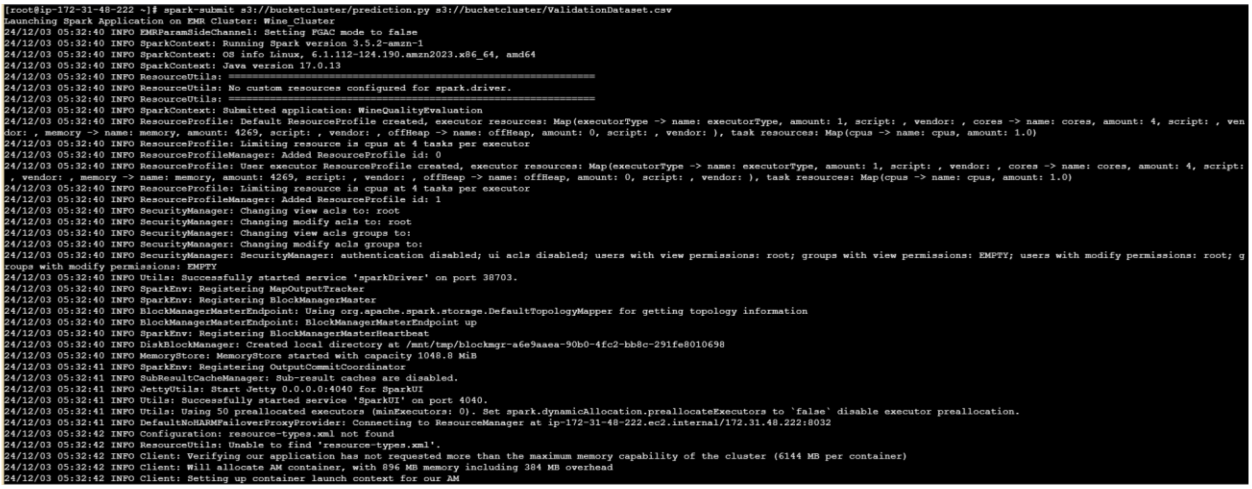
Fig 17. Running training.py



```
2-31-48-222.ec2.internal:20888/proxy/application_1733203398859_0001), /proxy/application_1733203398859_0001
24/12/03 05:28:22 INFO SingleEventLogFileWriter: Logging events to hdfs:/var/log/spark/apps/application_1733203398859_0001.inprogress
24/12/03 05:28:22 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
24/12/03 05:28:22 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
24/12/03 05:28:22 INFO ServerInfo: Adding filter to /jobs: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:22 INFO ServerInfo: Adding filter to /jobs/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:22 INFO ServerInfo: Adding filter to /jobs/job: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:22 INFO ServerInfo: Adding filter to /jobs/job/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:22 INFO ServerInfo: Adding filter to /stages: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:22 INFO ServerInfo: Adding filter to /stages/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:22 INFO ServerInfo: Adding filter to /stages/stage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /stages/stage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /stages/pool: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /stages/pool/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /storage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /storage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /storage/rdd: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /storage/rdd/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /environment: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:28:23 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
24/12/03 05:28:23 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
Reading training data from s3://bucketcluster/TrainingDataset.csv
Cleaning data
Setting up feature assembler and label indexer
Caching training data
Building RandomForestClassifier
Creating training pipeline
Setting up parameter grid for CrossValidator
Configuring CrossValidator
Fitting model with CrossValidator
Saving the trained model to s3://bucketcluster/trainedmodel
Stopping Spark Session
[root@ip-172-31-48-222 ~]#
```

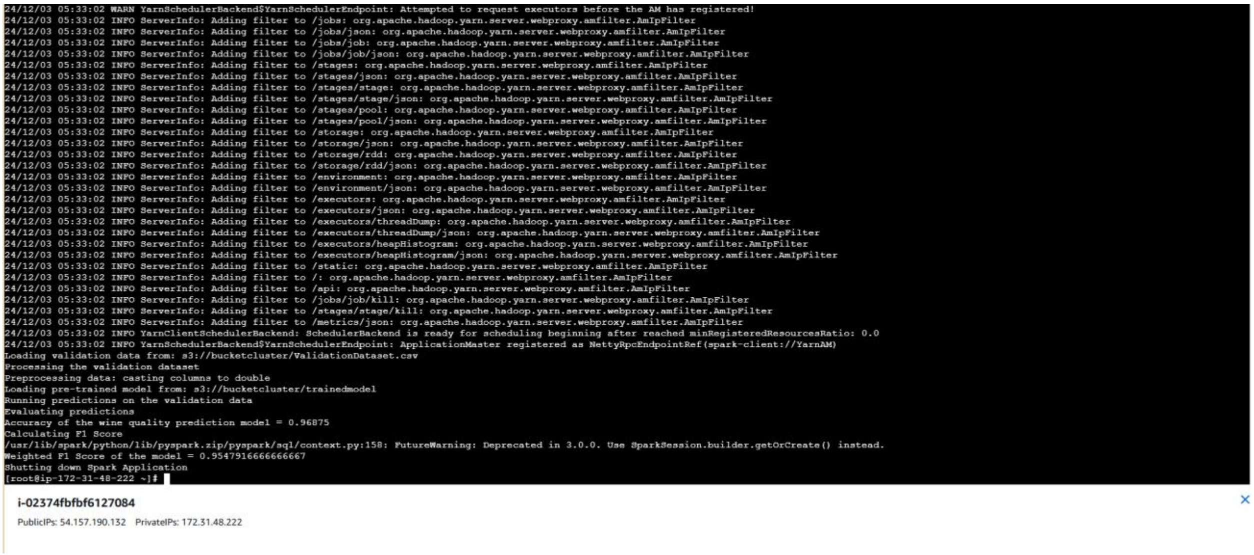
Fig 18. Completed executing training.py

4. Now Run prediction.py via command “***spark-submit s3://bucketcluster/prediction.py s3://bucketcluster/ValidationDataset.csv***” and wait until execution get finished



```
[root@ip-172-31-48-222 ~]# spark-submit s3://bucketcluster/prediction.py s3://bucketcluster/ValidationDataset.csv
Launching Spark Application on EMR Cluster: Wine_Cluster
24/12/03 05:32:40 INFO EMRParamSideChannel: Setting FGAC mode to false
24/12/03 05:32:40 INFO SparkContext: Running Spark version 3.5.2-amzn-1
24/12/03 05:32:40 INFO SparkContext: OS info Linux, 6.1.112-124.190.amzn2023.x86_64, amd64
24/12/03 05:32:40 INFO SparkContext: Java version 17.0.13
24/12/03 05:32:40 INFO ResourceUtils: ==============================================================
24/12/03 05:32:40 INFO ResourceUtils: No custom resources configured for spark.driver.
24/12/03 05:32:40 INFO ResourceUtils: ==============================================================
24/12/03 05:32:40 INFO SparkContext: Submitted application: WineQualityEvaluation
24/12/03 05:32:40 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(executorType -> name: executorType, amount: 1, script: , vendor: , cores -> name: cores, amount: 4, script: , ven
dor: , memory -> name: memory, amount: 4269, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/12/03 05:32:40 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/12/03 05:32:40 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/12/03 05:32:40 INFO ResourceProfile: User executor ResourceProfile created, executor resources: Map(executorType -> name: executorType, amount: 1, script: , vendor: , cores -> name: cores, amount: 4, script:
 , vendor: , memory -> name: memory, amount: 4269, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/12/03 05:32:40 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/12/03 05:32:40 INFO ResourceProfileManager: Added ResourceProfile id: 1
24/12/03 05:32:40 INFO SecurityManager: Changing view acls to: root
24/12/03 05:32:40 INFO SecurityManager: Changing modify acls to: root
24/12/03 05:32:40 INFO SecurityManager: Changing view acls groups to:
24/12/03 05:32:40 INFO SecurityManager: Changing modify acls groups to:
24/12/03 05:32:40 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: root; groups with view permissions: EMPTY; users with modify permissions: root; g
roups with modify permissions: EMPTY
24/12/03 05:32:40 INFO Utils: Successfully started service 'sparkDriver' on port 38703.
24/12/03 05:32:40 INFO SparkEnv: Registering MapOutputTracker
24/12/03 05:32:40 INFO SparkEnv: Registering BlockManagerMaster
24/12/03 05:32:40 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/12/03 05:32:40 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/12/03 05:32:40 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/12/03 05:32:40 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-a6e9aaea-90b0-4fc2-bb8c-291fe8010698
24/12/03 05:32:40 INFO MemoryStore: MemoryStore started with capacity 1048.8 MiB
24/12/03 05:32:41 INFO SparkEnv: Registering OutputCommitCoordinator
24/12/03 05:32:41 INFO SubResultCacheManager: Sub-result caches are disabled.
24/12/03 05:32:41 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/12/03 05:32:41 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/12/03 05:32:41 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
24/12/03 05:32:41 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-48-222.ec2.internal/172.31.48.222:8032
24/12/03 05:32:42 INFO Configuration: resource-types.xml not found
24/12/03 05:32:42 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/12/03 05:32:42 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (6144 MB per container)
24/12/03 05:32:42 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/12/03 05:32:42 INFO Client: Setting up container launch context for our AM
```

Fig 19. Running prediction.py



```
24/12/03 05:33:02 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /jobs: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /jobs/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /jobs/job: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /jobs/job/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /stages: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /stages/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /stages/stage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /stages/stage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /stages/pool: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /stages/pool/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /storage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /storage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /storage/rdd: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /storage/rdd/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /environment: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/12/03 05:33:02 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
24/12/03 05:33:02 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
Loading validation data from: s3://bucketcluster/ValidationDataset.csv
Processing the validation dataset
Preprocessing data: casting columns to double
Loading pre-trained model from: s3://bucketcluster/trainedmodel
Running predictions on the validation data
Evaluating predictions
Accuracy of the wine quality prediction model = 0.96875
Calculating F1 Score
/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/context.py:158: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
Weighted F1 Score of the model = 0.9547916666666667
Shutting down Spark Application
[root@ip-172-31-48-222 ~]#
```

i-02374fbfbf6127084

PublicIPs: 54.157.190.132 PrivateIPs: 172.31.48.222

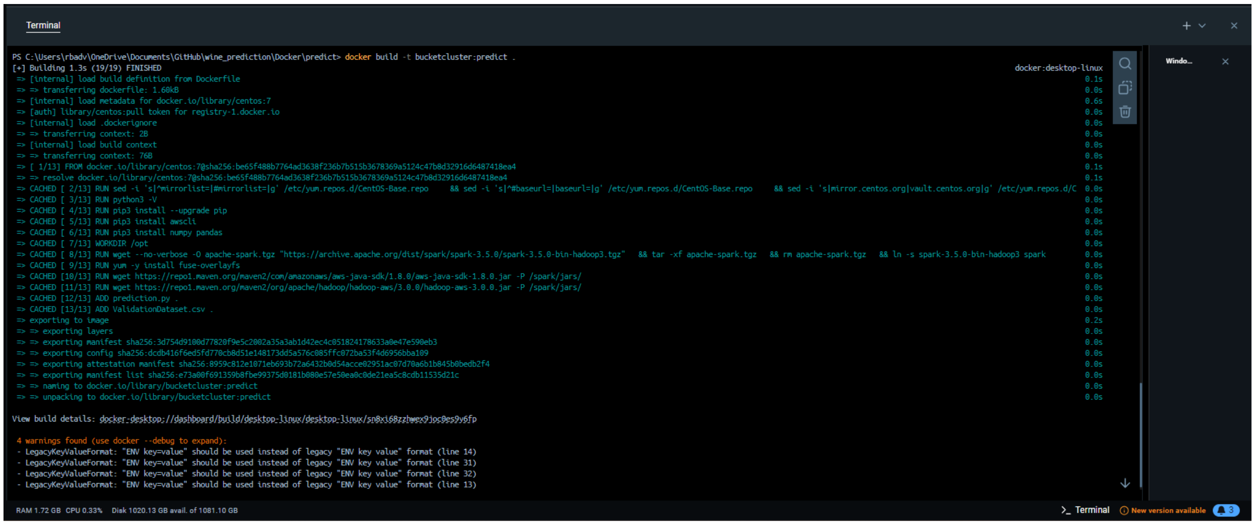
Fig 20. Completed executing prediction.py

5. We could able to conclude that after running prediction.py, we could able to find accuracy score is 96% and F1 score is 0.954791

## Execution with Docker

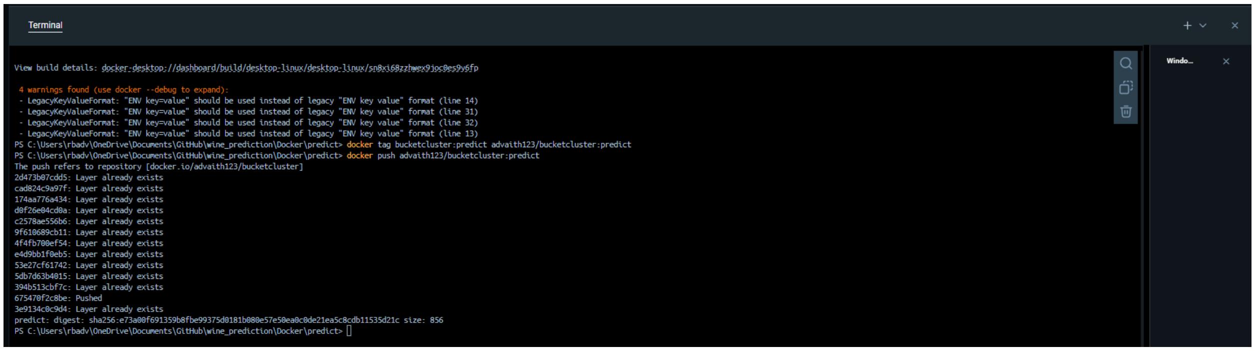
We are going to execute prediction.py with docker

1. Go to docker desktop, login there
2. Build, tag and push the compiled prediction.py images into docker so that it could be accessed by EC2. We need to do this step only if there any changes in the code or if you are pushing it for the first time.



```
PS C:\Users\rbadv\OneDrive\Documents\GitHub\wine_prediction\Docker\predict> docker build -t bucketcluster:predict .
[+] Building 1.3s (19/19) FINISHED                                                                                                   docker:desktop-linux
 => [internal] load build definition from Dockerfile                                                                                                 0.1s
 => => transferring dockerfile: 1.60kB                                                                                                               0.0s
 => [internal] load metadata for docker.io/library/centos:7                                                                                          0.6s
 => [auth] library/centos:pull token for registry-1.docker.io                                                                                        0.0s
 => [internal] load .dockerignore                                                                                                                    0.0s
 => => transferring context: 2B                                                                                                                      0.0s
 => [internal] load build context                                                                                                                    0.0s
 => => transferring context: 76B                                                                                                                     0.0s
 => [ 1/13] FROM docker.io/library/centos:7@sha256:be65f488b7764ad3638f236b7b515b3678369a5124c47b8d32916d6487418ea4                                   0.1s
 => => resolve docker.io/library/centos:7@sha256:be65f488b7764ad3638f236b7b515b3678369a5124c47b8d32916d6487418ea4                                     0.1s
 => CACHED [ 2/13] RUN sed -i 's|^mirrorlist=|#mirrorlist=|g' /etc/yum.repos.d/CentOS-Base.repo     && sed -i 's|^#baseurl=|baseurl=|g' /etc/yum.repos.d/CentOS-Base.repo     && sed -i 's|mirror.centos.org|vault.centos.org|g' /etc/yum.repos.d/C  0.0s
 => CACHED [ 3/13] RUN python3 -V                                                                                                                    0.0s
 => CACHED [ 4/13] RUN pip3 install --upgrade pip                                                                                                    0.0s
 => CACHED [ 5/13] RUN pip3 install awscli                                                                                                           0.0s
 => CACHED [ 6/13] RUN pip3 install numpy pandas                                                                                                     0.0s
 => CACHED [ 7/13] WORKDIR /opt                                                                                                                      0.0s
 => CACHED [ 8/13] RUN wget --no-verbose -O apache-spark.tgz "https://archive.apache.org/dist/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz"   && tar -xf apache-spark.tgz   && rm apache-spark.tgz   && ln -s spark-3.5.0-bin-hadoop3 spark      0.0s
 => CACHED [ 9/13] RUN yum -y install fuse-overlayfs                                                                                                 0.0s
 => CACHED [10/13] RUN wget https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk/1.8.0/aws-java-sdk-1.8.0.jar -P /spark/jars/                   0.0s
 => CACHED [11/13] RUN wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.0.0/hadoop-aws-3.0.0.jar -P /spark/jars/                   0.0s
 => CACHED [12/13] ADD prediction.py .                                                                                                               0.0s
 => CACHED [13/13] ADD ValidationDataset.csv .                                                                                                       0.0s
 => exporting to image                                                                                                                               0.2s
 => => exporting layers                                                                                                                              0.0s
 => => exporting manifest sha256:3d754d9100d77820f9e5c2002a35a3ab1d42ec4c051824178633a0e47e590eb3                                                    0.0s
 => => exporting config sha256:dcdb416f6ed5fd770cb8d51e148173dd5a576c085ffc072ba53f4d6956bba109                                                      0.0s
 => => exporting attestation manifest sha256:8959c812e1071eb693b72a6432b0d54acce02951ac07d70a6b1b845b0bedb2f4                                        0.0s
 => => exporting manifest list sha256:e73a00f691359b8fbe99375d0181b080e57e50ea0c0de21ea5c8cdb11535d21c                                               0.0s
 => => naming to docker.io/library/bucketcluster:predict                                                                                             0.0s
 => => unpacking to docker.io/library/bucketcluster:predict                                                                                          0.0s

View build details: docker-desktop://dashboard/build/desktop-linux/desktop-linux/sn8xi68zzhwex9joc0es9y6fp

 4 warnings found (use docker --debug to expand):
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 14)
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 31)
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 32)
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 13)
```

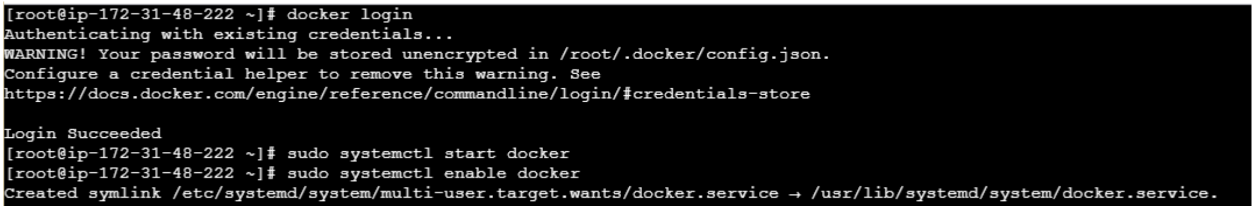
Fig 21. Building Image of prediction.py in docker



```
View build details: docker-desktop://dashboard/build/desktop-linux/desktop-linux/sn8xi68zzhwex9joc0es9y6fp

 4 warnings found (use docker --debug to expand):
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 14)
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 31)
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 32)
 - LegacyKeyValueFormat: "ENV key=value" should be used instead of legacy "ENV key value" format (line 13)
PS C:\Users\rbadv\OneDrive\Documents\GitHub\wine_prediction\Docker\predict> docker tag bucketcluster:predict advaith123/bucketcluster:predict
PS C:\Users\rbadv\OneDrive\Documents\GitHub\wine_prediction\Docker\predict> docker push advaith123/bucketcluster:predict
The push refers to repository [docker.io/advaith123/bucketcluster]
2d473b07cdd5: Layer already exists
cad824c9a97f: Layer already exists
174aa776a434: Layer already exists
d0f26e04cd0a: Layer already exists
c2578ae556b6: Layer already exists
9f610689cb11: Layer already exists
4f4fb700ef54: Layer already exists
e4d9bb1f0eb5: Layer already exists
53e27cf61742: Layer already exists
5db7d63b4015: Layer already exists
394b513cbf7c: Layer already exists
675470f2c8be: Pushed
3e9134c0c9d4: Layer already exists
predict: digest: sha256:e73a00f691359b8fbe99375d0181b080e57e50ea0c0de21ea5c8cdb11535d21c size: 856
PS C:\Users\rbadv\OneDrive\Documents\GitHub\wine_prediction\Docker\predict> 
```

Fig 22. Tagging and Pushing Image of prediction.py in docker

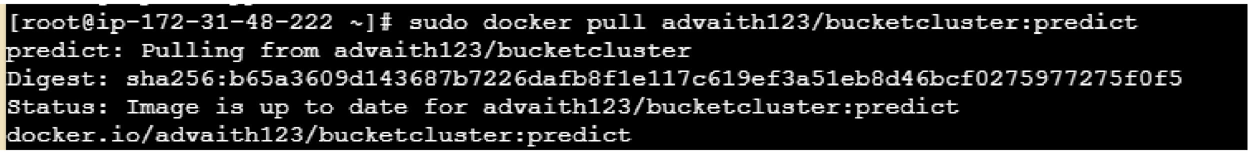
3. Now go to EC2 instance and login into docker there. Also type cmd “***sudo systemctl start docker*** “ and “ ***sudo systemctl enable docker***” as well



```
[root@ip-172-31-48-222 ~]# docker login
Authenticating with existing credentials...
WARNING! Your password will be stored unencrypted in /root/.docker/config.json.
Configure a credential helper to remove this warning. See
https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
[root@ip-172-31-48-222 ~]# sudo systemctl start docker
[root@ip-172-31-48-222 ~]# sudo systemctl enable docker
Created symlink /etc/systemd/system/multi-user.target.wants/docker.service → /usr/lib/systemd/system/docker.service.
```

Fig 23. Starting Docker in EC2

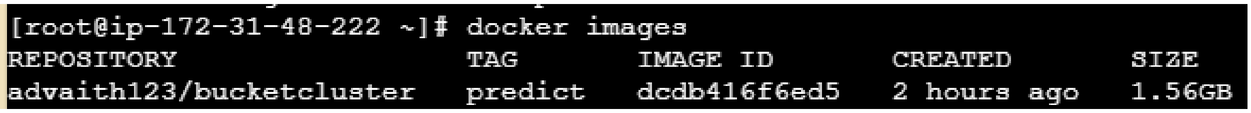
4. Pull prediction.py image from docker using command “***sudo docker pull advaith123/bucketcluster:predict*** “. The below Figure states that the images have already been pulled



```
[root@ip-172-31-48-222 ~]# sudo docker pull advaith123/bucketcluster:predict
predict: Pulling from advaith123/bucketcluster
Digest: sha256:b65a3609d143687b7226dafb8f1e117c619ef3a51eb8d46bcf0275977275f0f5
Status: Image is up to date for advaith123/bucketcluster:predict
docker.io/advaith123/bucketcluster:predict
```

Fig 24. Pulling images to EC2

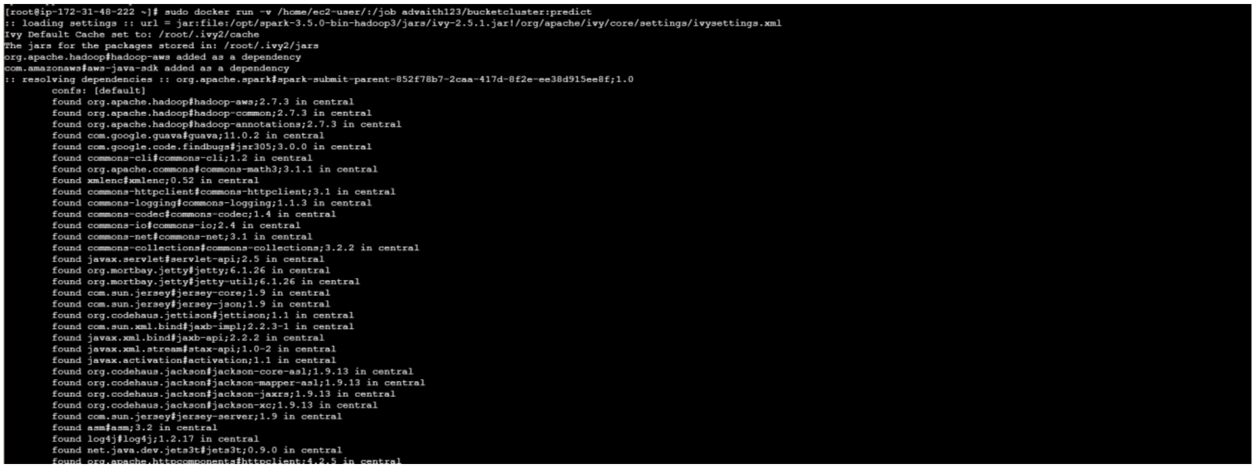
5. Use command “***docker images***” to check whether is image uploaded or not



```
[root@ip-172-31-48-222 ~]# docker images
REPOSITORY                 TAG       IMAGE ID       CREATED        SIZE
advaith123/bucketcluster   predict   dcdb416f6ed5   2 hours ago    1.56GB
```

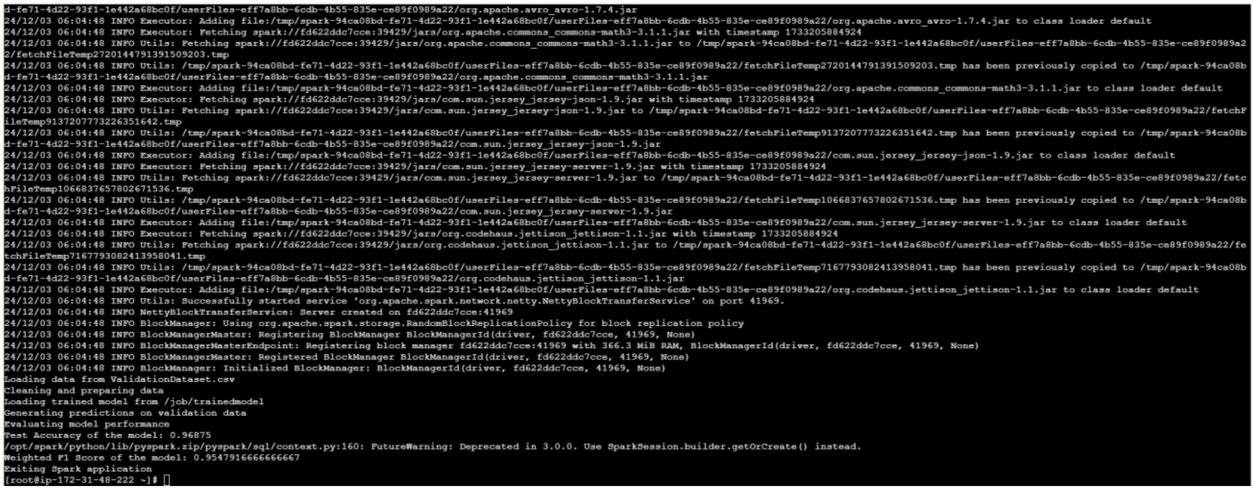
Fig 25. Checking whether if images exists in EC2

6. Now Run prediction.py via command “***sudo docker run -v /home/ec2-user/:/job advaith123/bucketcluster:predict***” and wait until execution get finished



```
[root@ip-172-31-48-222 ~]# sudo docker run -v /home/ec2-user/:/job advaith123/bucketcluster:predict
:: loading settings :: url = jar:file:/opt/spark-3.5.0-bin-hadoop3/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
org.apache.hadoop#hadoop-aws added as a dependency
com.amazonaws#aws-java-sdk added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-852f78b7-2caa-417d-8f2e-ee38d915ee8f;1.0
        confs: [default]
        found org.apache.hadoop#hadoop-aws;2.7.3 in central
        found org.apache.hadoop#hadoop-common;2.7.3 in central
        found org.apache.hadoop#hadoop-annotations;2.7.3 in central
        found com.google.guava#guava;11.0.2 in central
        found com.google.code.findbugs#jsr305;3.0.0 in central
        found commons-cli#commons-cli;1.2 in central
        found org.apache.commons#commons-math3;3.1.1 in central
        found xmlenc#xmlenc;0.52 in central
        found commons-httpclient#commons-httpclient;3.1 in central
        found commons-logging#commons-logging;1.1.3 in central
        found commons-codec#commons-codec;1.4 in central
        found commons-io#commons-io;2.4 in central
        found commons-net#commons-net;3.1 in central
        found commons-collections#commons-collections;3.2.2 in central
        found javax.servlet#servlet-api;2.5 in central
        found org.mortbay.jetty#jetty;6.1.26 in central
        found org.mortbay.jetty#jetty-util;6.1.26 in central
        found com.sun.jersey#jersey-core;1.9 in central
        found com.sun.jersey#jersey-json;1.9 in central
        found org.codehaus.jettison#jettison;1.1 in central
        found com.sun.xml.bind#jaxb-impl;2.2.3-1 in central
        found javax.xml.bind#jaxb-api;2.2.2 in central
        found javax.xml.stream#stax-api;1.0-2 in central
        found javax.activation#activation;1.1 in central
        found org.codehaus.jackson#jackson-core-asl;1.9.13 in central
        found org.codehaus.jackson#jackson-mapper-asl;1.9.13 in central
        found org.codehaus.jackson#jackson-jaxrs;1.9.13 in central
        found org.codehaus.jackson#jackson-xc;1.9.13 in central
        found com.sun.jersey#jersey-server;1.9 in central
        found asm#asm;3.2 in central
        found log4j#log4j;1.2.17 in central
        found net.java.dev.jets3t#jets3t;0.9.0 in central
        found org.apache.httpcomponents#httpclient;4.2.5 in central
```

Fig 26. Running Docker Prediction Image in EC2

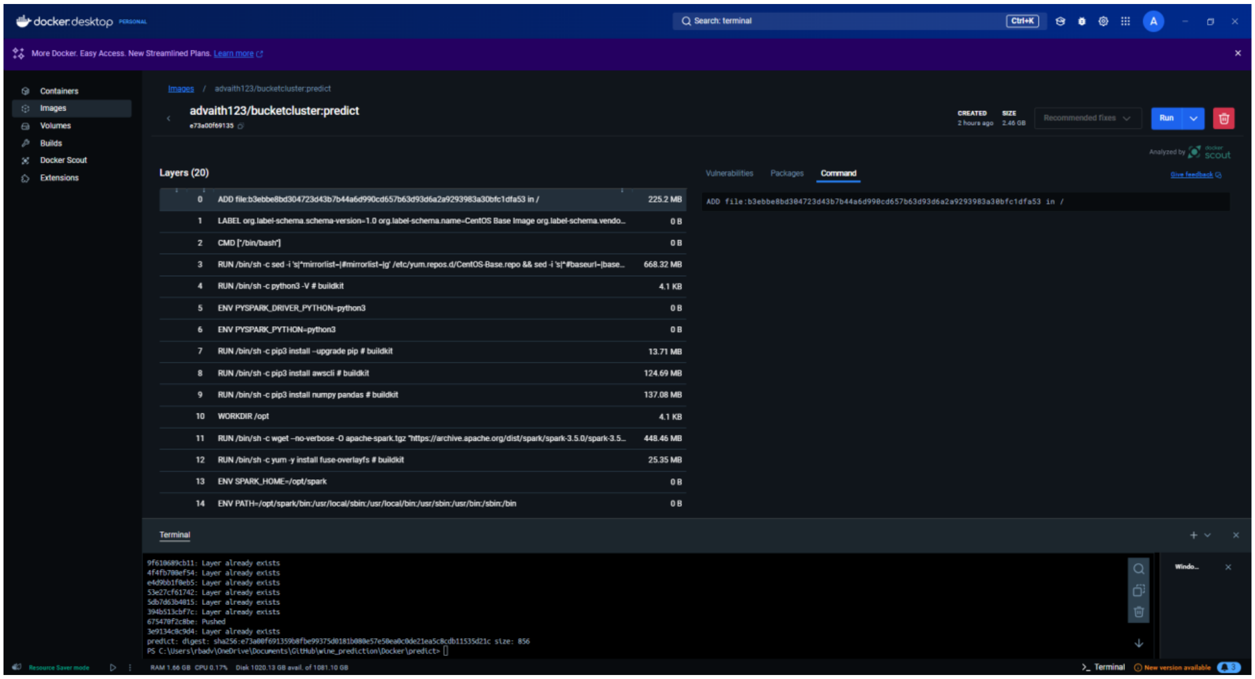


```
d-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/org.apache.avro_avro-1.7.4.jar
24/12/03 06:04:48 INFO Executor: Adding file:/tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/org.apache.avro_avro-1.7.4.jar to class loader default
24/12/03 06:04:48 INFO Executor: Fetching spark://fd622ddc7cce:39429/jars/org.apache.commons_commons-math3-3.1.1.jar with timestamp 1733205884924
24/12/03 06:04:48 INFO Utils: Fetching spark://fd622ddc7cce:39429/jars/org.apache.commons_commons-math3-3.1.1.jar to /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a2
2/fetchFileTemp2720144791391509203.tmp
24/12/03 06:04:48 INFO Utils: /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/fetchFileTemp2720144791391509203.tmp has been previously copied to /tmp/spark-94ca08b
d-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/org.apache.commons_commons-math3-3.1.1.jar
24/12/03 06:04:48 INFO Executor: Adding file:/tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/org.apache.commons_commons-math3-3.1.1.jar to class loader default
24/12/03 06:04:48 INFO Executor: Fetching spark://fd622ddc7cce:39429/jars/com.sun.jersey_jersey-json-1.9.jar with timestamp 1733205884924
24/12/03 06:04:48 INFO Utils: Fetching spark://fd622ddc7cce:39429/jars/com.sun.jersey_jersey-json-1.9.jar to /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/fetchF
ileTemp9137207773226351642.tmp
24/12/03 06:04:48 INFO Utils: /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/fetchFileTemp9137207773226351642.tmp has been previously copied to /tmp/spark-94ca08b
d-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/com.sun.jersey_jersey-json-1.9.jar
24/12/03 06:04:48 INFO Executor: Adding file:/tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/com.sun.jersey_jersey-json-1.9.jar to class loader default
24/12/03 06:04:48 INFO Executor: Fetching spark://fd622ddc7cce:39429/jars/com.sun.jersey_jersey-server-1.9.jar with timestamp 1733205884924
24/12/03 06:04:48 INFO Utils: Fetching spark://fd622ddc7cce:39429/jars/com.sun.jersey_jersey-server-1.9.jar to /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/fetc
hFileTemp1066837657802671536.tmp
24/12/03 06:04:48 INFO Utils: /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/fetchFileTemp1066837657802671536.tmp has been previously copied to /tmp/spark-94ca08b
d-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/com.sun.jersey_jersey-server-1.9.jar
24/12/03 06:04:48 INFO Executor: Adding file:/tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/com.sun.jersey_jersey-server-1.9.jar to class loader default
24/12/03 06:04:48 INFO Executor: Fetching spark://fd622ddc7cce:39429/jars/org.codehaus.jettison_jettison-1.1.jar with timestamp 1733205884924
24/12/03 06:04:48 INFO Utils: Fetching spark://fd622ddc7cce:39429/jars/org.codehaus.jettison_jettison-1.1.jar to /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/fe
tchFileTemp7167793082413958041.tmp
24/12/03 06:04:48 INFO Utils: /tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/fetchFileTemp7167793082413958041.tmp has been previously copied to /tmp/spark-94ca08b
d-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/org.codehaus.jettison_jettison-1.1.jar
24/12/03 06:04:48 INFO Executor: Adding file:/tmp/spark-94ca08bd-fe71-4d22-93f1-1e442a68bc0f/userFiles-eff7a8bb-6cdb-4b55-835e-ce89f0989a22/org.codehaus.jettison_jettison-1.1.jar to class loader default
24/12/03 06:04:48 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 41969.
24/12/03 06:04:48 INFO NettyBlockTransferService: Server created on fd622ddc7cce:41969
24/12/03 06:04:48 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/12/03 06:04:48 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, fd622ddc7cce, 41969, None)
24/12/03 06:04:48 INFO BlockManagerMasterEndpoint: Registering block manager fd622ddc7cce:41969 with 366.3 MiB RAM, BlockManagerId(driver, fd622ddc7cce, 41969, None)
24/12/03 06:04:48 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, fd622ddc7cce, 41969, None)
24/12/03 06:04:48 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, fd622ddc7cce, 41969, None)
Loading data from ValidationDataset.csv
Cleaning and preparing data
Loading trained model from /job/trainedmodel
Generating predictions on validation data
Evaluating model performance
Test Accuracy of the model: 0.96875
/opt/spark/python/lib/pyspark.zip/pyspark/sql/context.py:160: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
Weighted F1 Score of the model: 0.9547916666666667
Exiting Spark application
[root@ip-172-31-48-222 ~]# 
```

Fig 27. Completed executing Docker Prediction Image in EC2

7. We could able to conclude that after running prediction image from docker, we could able to find accuracy score is 96% and F1 score is 0.954791

## Docker Image



advaith123/bucketcluster:predict

e73a00f69135

Layers (20)

| | | |
|---|---|---|
| 0 | ADD file:b3ebbe8bd304723d43b7b44a6d990cd657b63d93d6a2a9293983a30bfc1dfa53 in / | 225.2 MB |
| 1 | LABEL org.label-schema.schema-version=1.0 org.label-schema.name=CentOS Base Image org.label-schema.vendo... | 0 B |
| 2 | CMD ["/bin/bash"] | 0 B |
| 3 | RUN /bin/sh -c sed -i 's\|^mirrorlist=\|#mirrorlist=\|g' /etc/yum.repos.d/CentOS-Base.repo && sed -i 's\|^#baseurl=\|base... | 668.32 MB |
| 4 | RUN /bin/sh -c python3 -V # buildkit | 4.1 KB |
| 5 | ENV PYSPARK_DRIVER_PYTHON=python3 | 0 B |
| 6 | ENV PYSPARK_PYTHON=python3 | 0 B |
| 7 | RUN /bin/sh -c pip3 install --upgrade pip # buildkit | 13.71 MB |
| 8 | RUN /bin/sh -c pip3 install awscli # buildkit | 124.69 MB |
| 9 | RUN /bin/sh -c pip3 install numpy pandas # buildkit | 137.08 MB |
| 10 | WORKDIR /opt | 4.1 KB |
| 11 | RUN /bin/sh -c wget --no-verbose -O apache-spark.tgz "https://archive.apache.org/dist/spark/spark-3.5.0/spark-3.5... | 448.46 MB |
| 12 | RUN /bin/sh -c yum -y install fuse-overlayfs # buildkit | 25.35 MB |
| 13 | ENV SPARK_HOME=/opt/spark | 0 B |
| 14 | ENV PATH=/opt/spark/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin | 0 B |

```
ADD file:b3ebbe8bd304723d43b7b44a6d990cd657b63d93d6a2a9293983a30bfc1dfa53 in /
```

```
9f610689cb11: Layer already exists
4f4fb700ef54: Layer already exists
e4d9bb1f0eb5: Layer already exists
53e27cf61742: Layer already exists
5db7d63b4015: Layer already exists
394b513cbf7c: Layer already exists
675470f2c8be: Pushed
3e9134c0c9d4: Layer already exists
predict: digest: sha256:e73a00f691359b8fbe99375d0181b000e57e50ea0c0de21ea5c8cdb11535d21c size: 856
PS C:\Users\rbadv\OneDrive\Documents\GitHub\wine_prediction\Docker\predict> 
```

Fig 28. Docker Image