# CS 643 Programming Assignment-2

mb2332@njit.edu

Manas Bhut

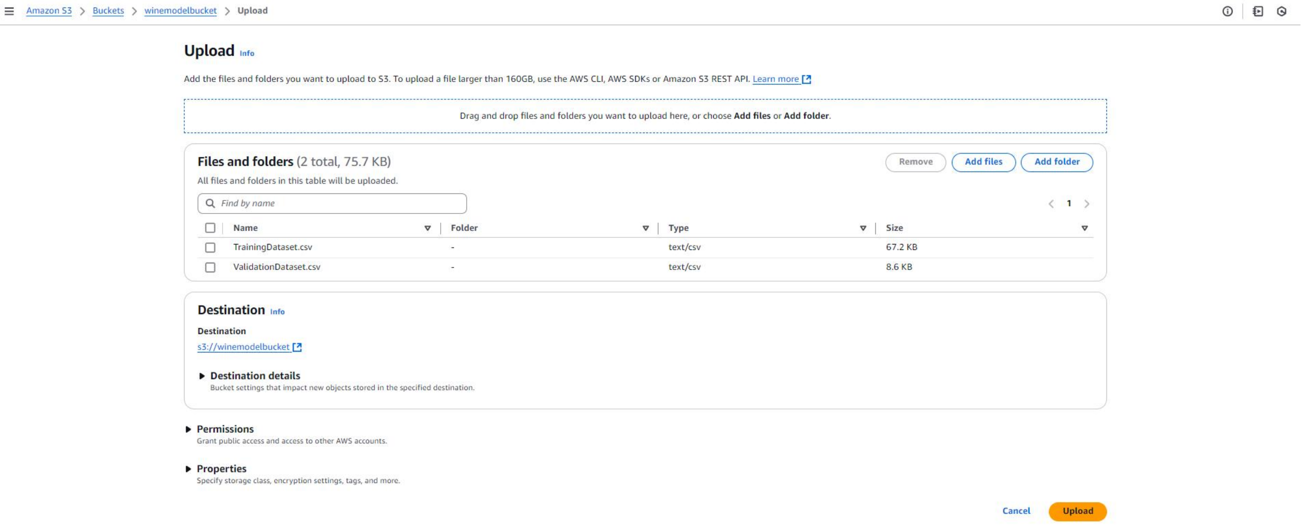
Github Link
https://github.com/Manas1227/cs643-853-pa2-mb2332/tree/master

Docker Link
https://hub.docker.com/repository/docker/manasbhut/cs643-pa2-aws-spark/general

## SECTION 1: AWS Cloud setup for running the training ML application - training.py

### Step-1: Create s3 bucket and upload required files

After creating new bucket, click on Add Files and upload both the datasets files to newly created S3 bucket



Amazon S3 > Buckets > winemodelbucket > Upload

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDKs or Amazon S3 REST API. Learn more

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

**Files and folders** (2 total, 75.7 KB)

All files and folders in this table will be uploaded.

| □ | Name | Folder | Type | Size |
|---|---|---|---|---|
| □ | TrainingDataset.csv | - | text/csv | 67.2 KB |
| □ | ValidationDataset.csv | - | text/csv | 8.6 KB |

**Destination** Info

**Destination**
s3://winemodelbucket

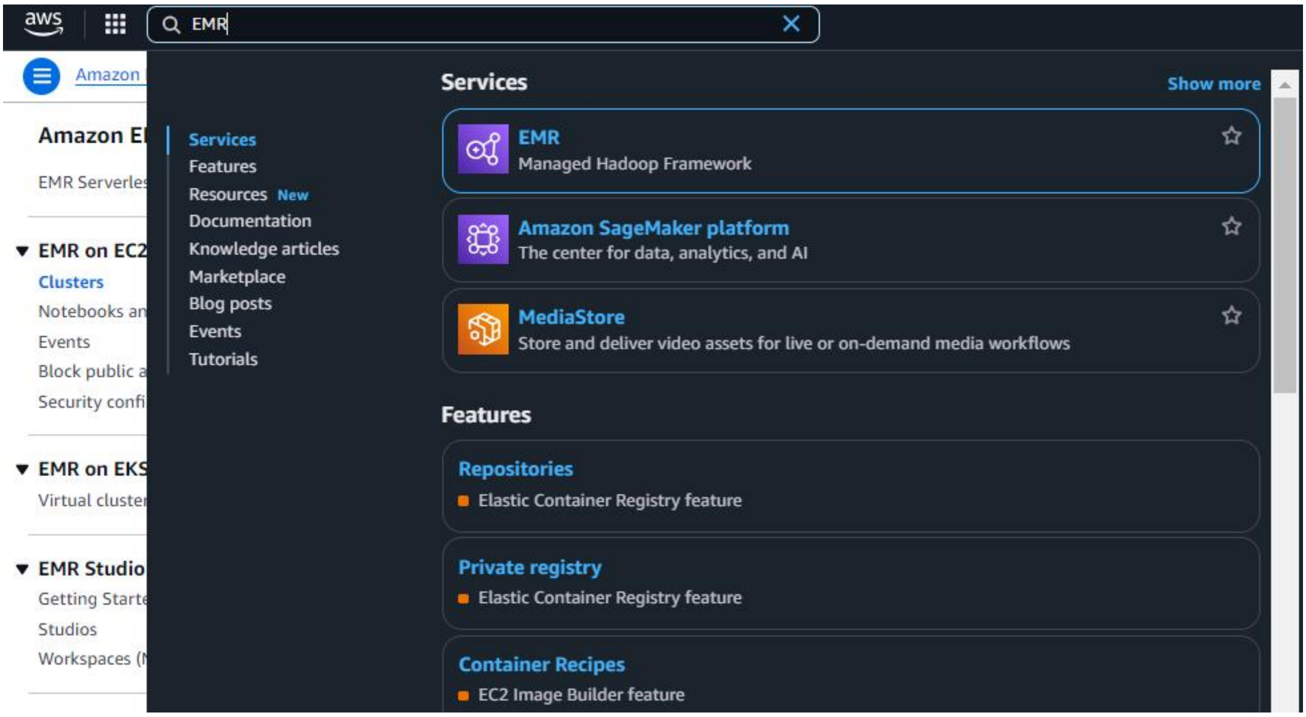
▶ **Destination details**
Bucket settings that impact new objects stored in the specified destination.

▶ **Permissions**
Grant public access and access to other AWS accounts.

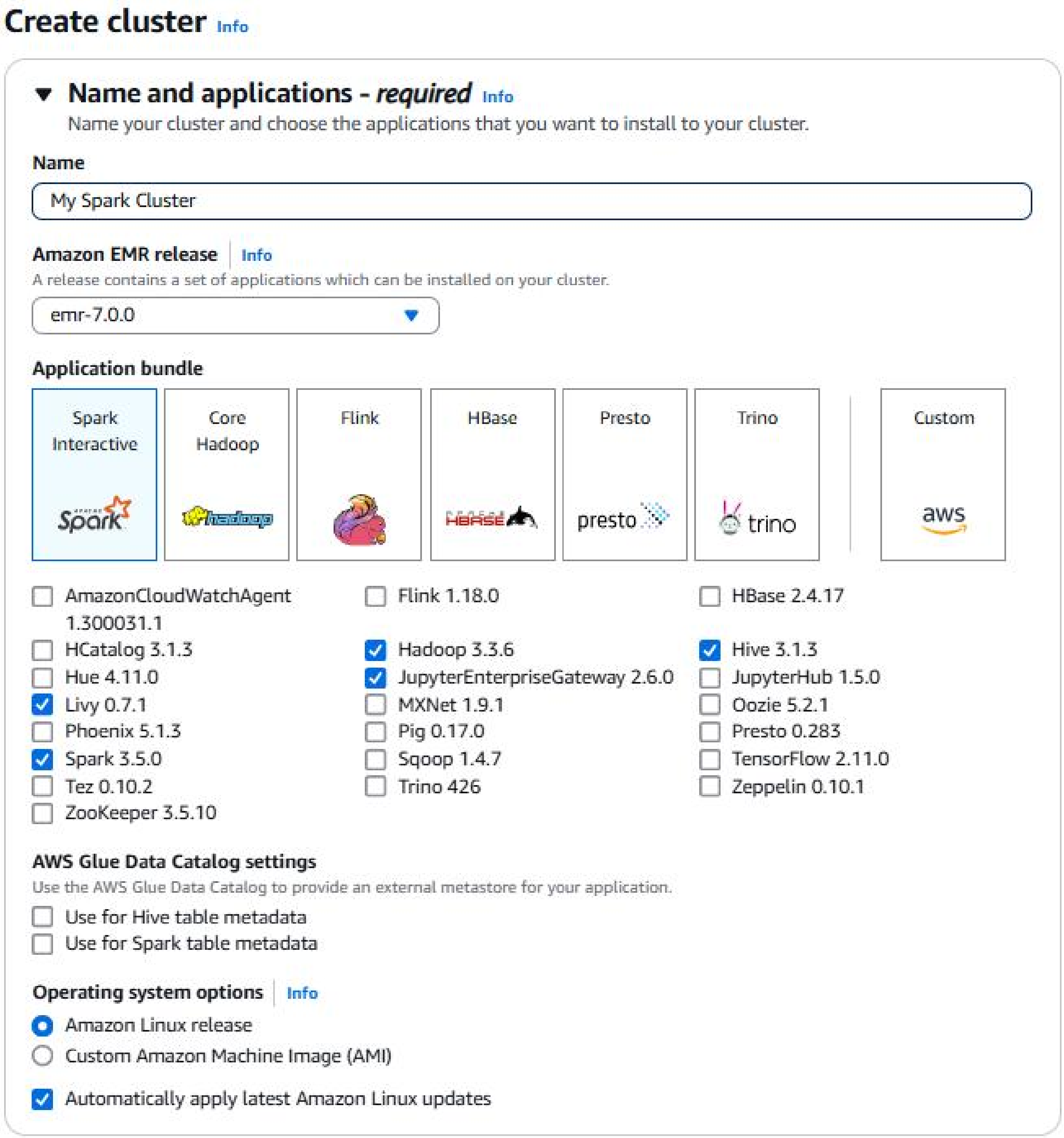
▶ **Properties**
Specify storage class, encryption settings, tags, and more.

## Step-2 Create an EMR cluster

Login to AWS console
Search for “EMR”
Click on “Create CLuster” option



Follow the below screen shots to sucessfully create the SPARK Cluster



## Create cluster Info

### ▼ Name and applications - *required* Info

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

**Name**

My Spark Cluster

**Amazon EMR release** Info

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

emr-7.0.0 ▼

**Application bundle**

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

- □ AmazonCloudWatchAgent 1.300031.1
- □ HCatalog 3.1.3
- □ Hue 4.11.0
- ☑ Livy 0.7.1
- □ Phoenix 5.1.3
- ☑ Spark 3.5.0
- □ Tez 0.10.2
- □ ZooKeeper 3.5.10
- □ Flink 1.18.0
- ☑ Hadoop 3.3.6
- ☑ JupyterEnterpriseGateway 2.6.0
- □ MXNet 1.9.1
- □ Pig 0.17.0
- □ Sqoop 1.4.7
- □ Trino 426
- □ HBase 2.4.17
- ☑ Hive 3.1.3
- □ JupyterHub 1.5.0
- □ Oozie 5.2.1
- □ Presto 0.283
- □ TensorFlow 2.11.0
- □ Zeppelin 0.10.1

**AWS Glue Data Catalog settings**

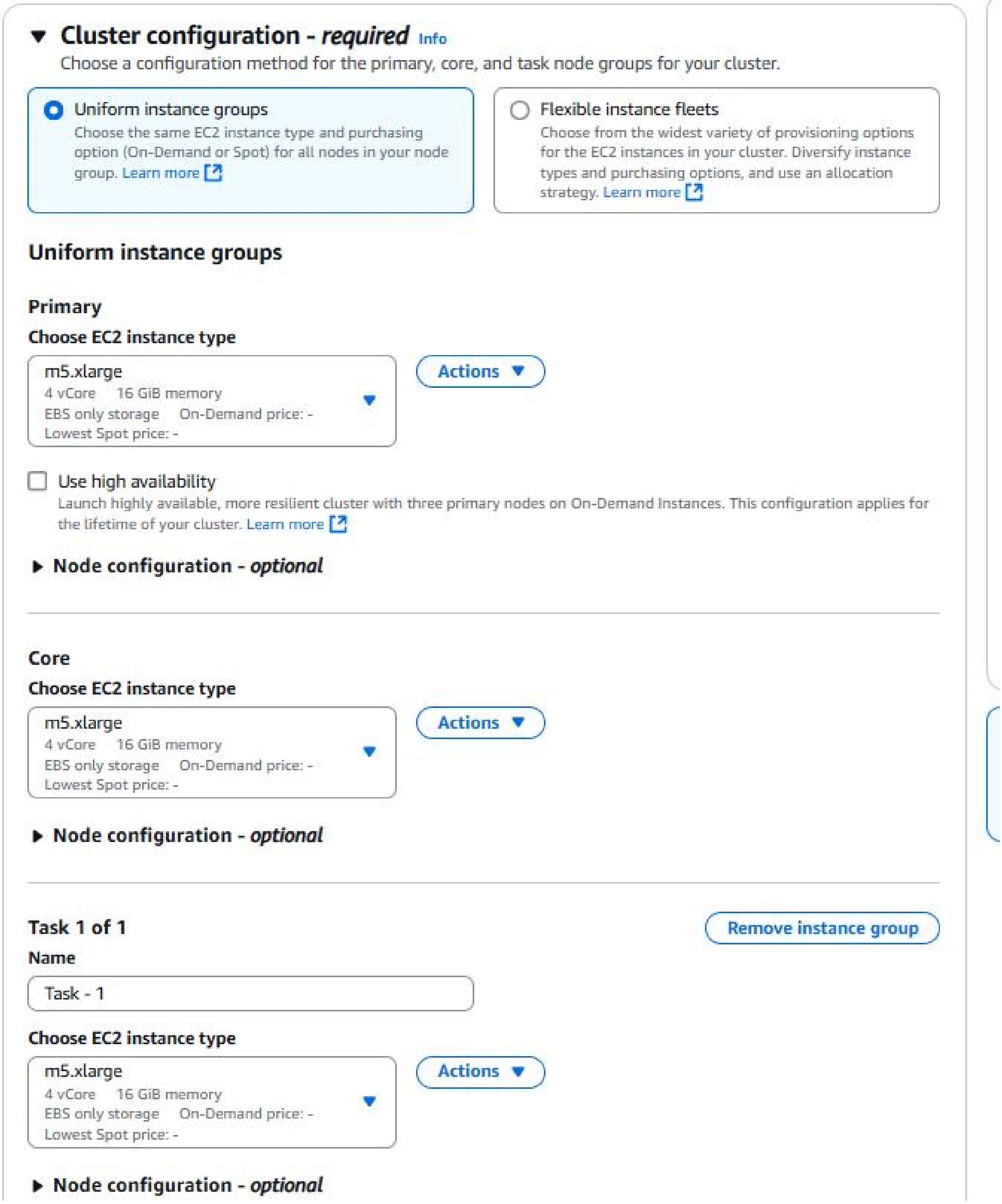
Use the AWS Glue Data Catalog to provide an external metastore for your application.

- □ Use for Hive table metadata
- □ Use for Spark table metadata

**Operating system options** Info

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

☑ Automatically apply latest Amazon Linux updates



▼ Cluster configuration - *required* Info

Choose a configuration method for the primary, core, and task node groups for your cluster.

● Uniform instance groups
Choose the same EC2 instance type and purchasing option (On-Demand or Spot) for all nodes in your node group. Learn more

○ Flexible instance fleets
Choose from the widest variety of provisioning options for the EC2 instances in your cluster. Diversify instance types and purchasing options, and use an allocation strategy. Learn more

Uniform instance groups

Primary

Choose EC2 instance type

m5.xlarge
4 vCore 16 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

m5.xlarge
4 vCore 16 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

▶ Node configuration - *optional*

Task 1 of 1

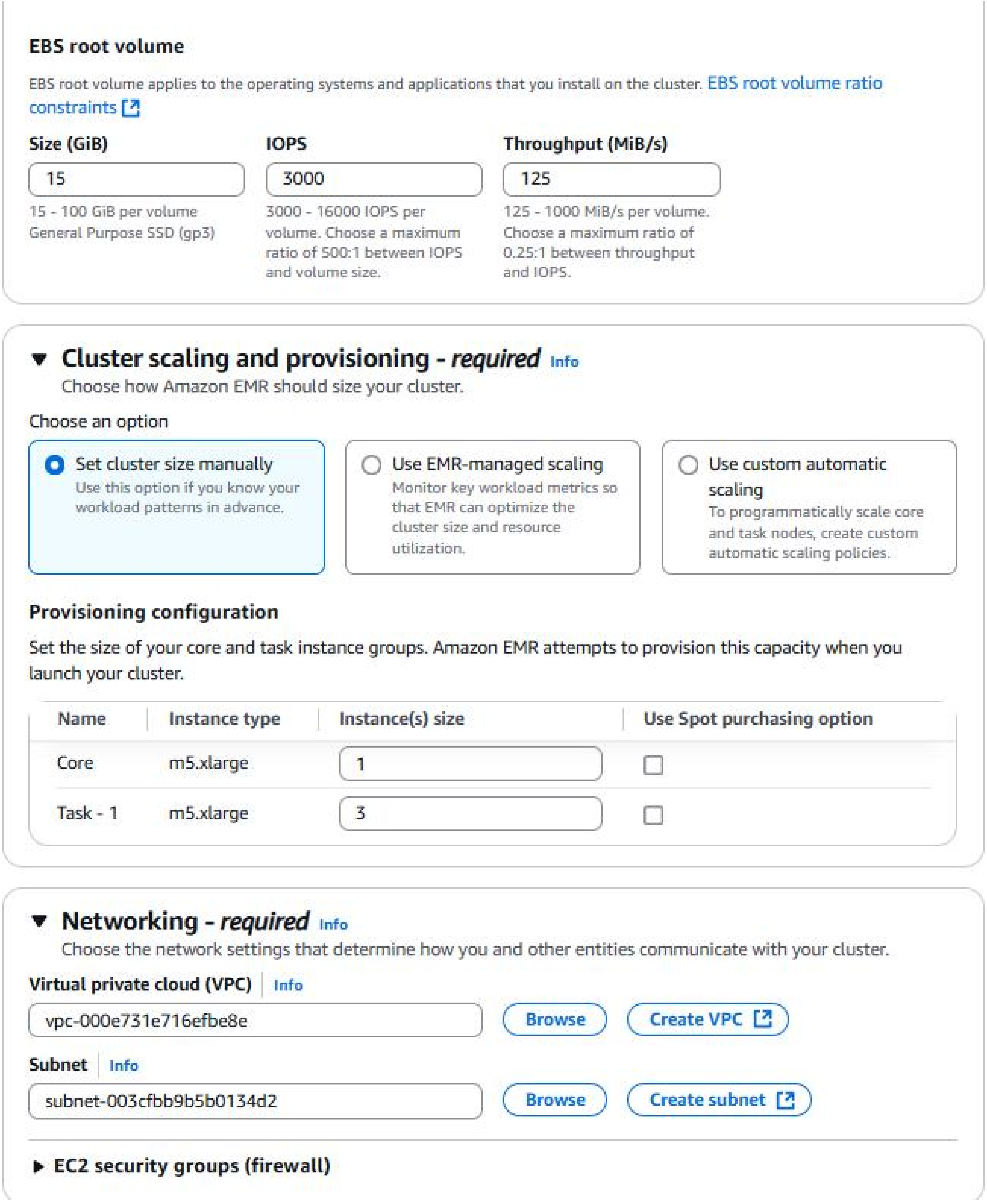
Name

Task - 1

Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

▶ Node configuration - *optional*



### EBS root volume

EBS root volume applies to the operating systems and applications that you install on the cluster. EBS root volume ratio constraints

| Size (GiB) | IOPS | Throughput (MiB/s) |
|---|---|---|
| 15 | 3000 | 125 |
| 15 - 100 GiB per volume General Purpose SSD (gp3) | 3000 - 16000 IOPS per volume. Choose a maximum ratio of 500:1 between IOPS and volume size. | 125 - 1000 MiB/s per volume. Choose a maximum ratio of 0.25:1 between throughput and IOPS. |

### ▼ Cluster scaling and provisioning - *required* Info

Choose how Amazon EMR should size your cluster.

Choose an option

● Set cluster size manually
Use this option if you know your workload patterns in advance.

○ Use EMR-managed scaling
Monitor key workload metrics so that EMR can optimize the cluster size and resource utilization.

○ Use custom automatic scaling
To programmatically scale core and task nodes, create custom automatic scaling policies.

**Provisioning configuration**

Set the size of your core and task instance groups. Amazon EMR attempts to provision this capacity when you launch your cluster.

| Name | Instance type | Instance(s) size | Use Spot purchasing option |
|---|---|---|---|
| Core | m5.xlarge | 1 | □ |
| Task - 1 | m5.xlarge | 3 | □ |

### ▼ Networking - *required* Info

Choose the network settings that determine how you and other entities communicate with your cluster.

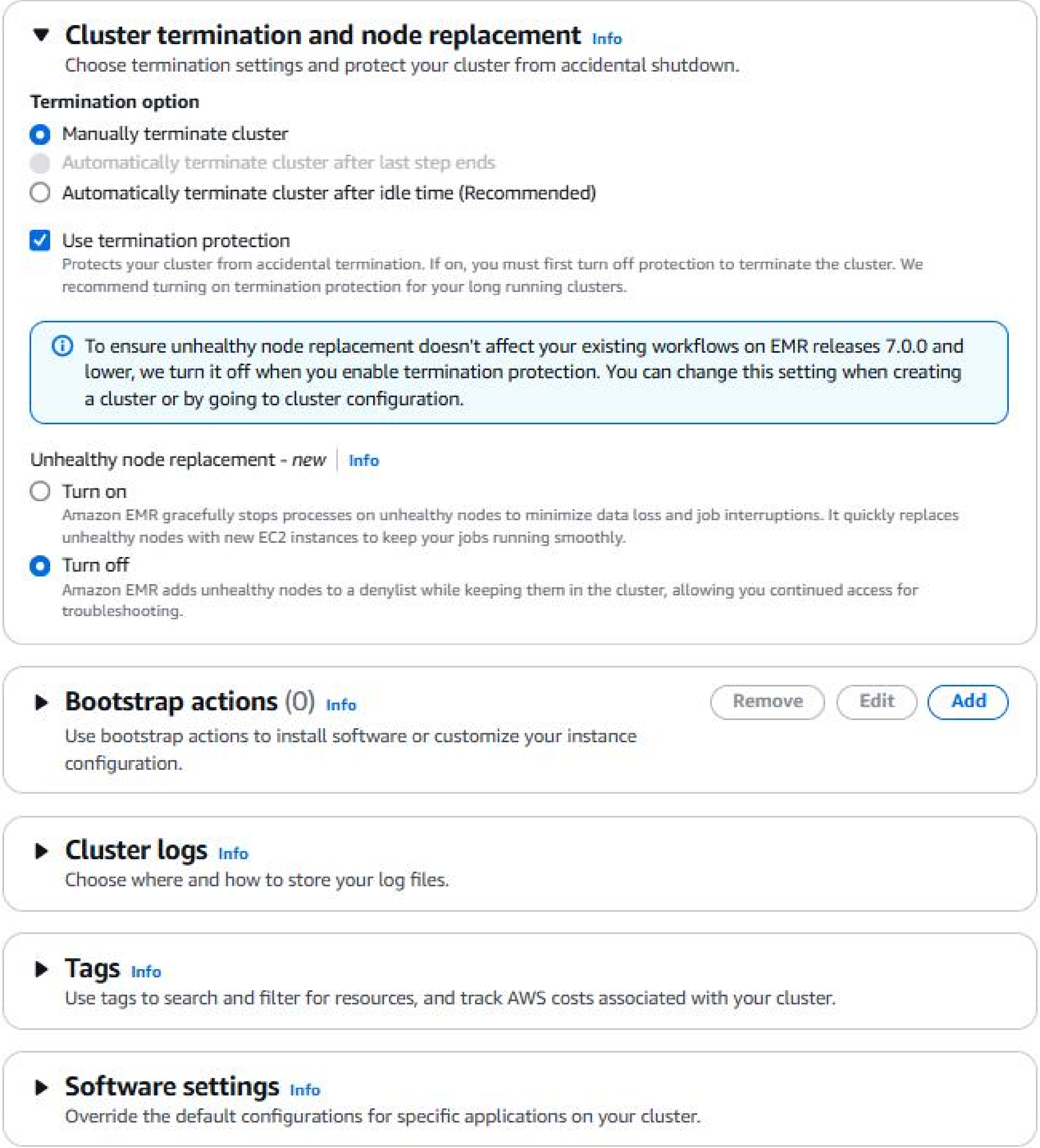
**Virtual private cloud (VPC)** Info

vpc-000e731e716efbe8e

**Subnet** Info

subnet-003cfbb9b5b0134d2

▶ **EC2 security groups (firewall)**



▼ **Cluster termination and node replacement** Info

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.

ⓘ To ensure unhealthy node replacement doesn't affect your existing workflows on EMR releases 7.0.0 and lower, we turn it off when you enable termination protection. You can change this setting when creating a cluster or by going to cluster configuration.

Unhealthy node replacement - *new* | Info

○ 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) Info

Use bootstrap actions to install software or customize your instance configuration.

▶ **Cluster logs** Info

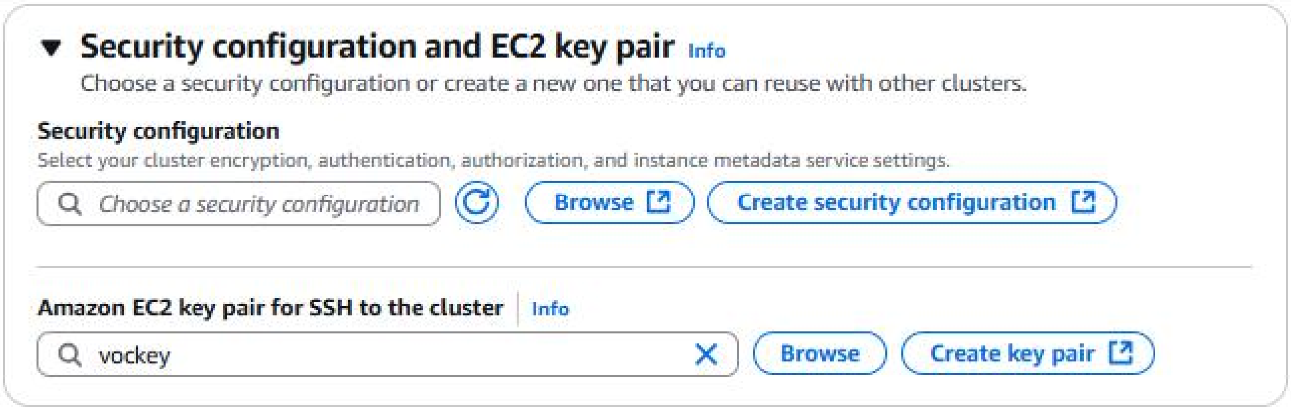
Choose where and how to store your log files.

▶ **Tags** Info

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

▶ **Software settings** Info

Override the default configurations for specific applications on your cluster.



▼ Security configuration and EC2 key pair Info

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

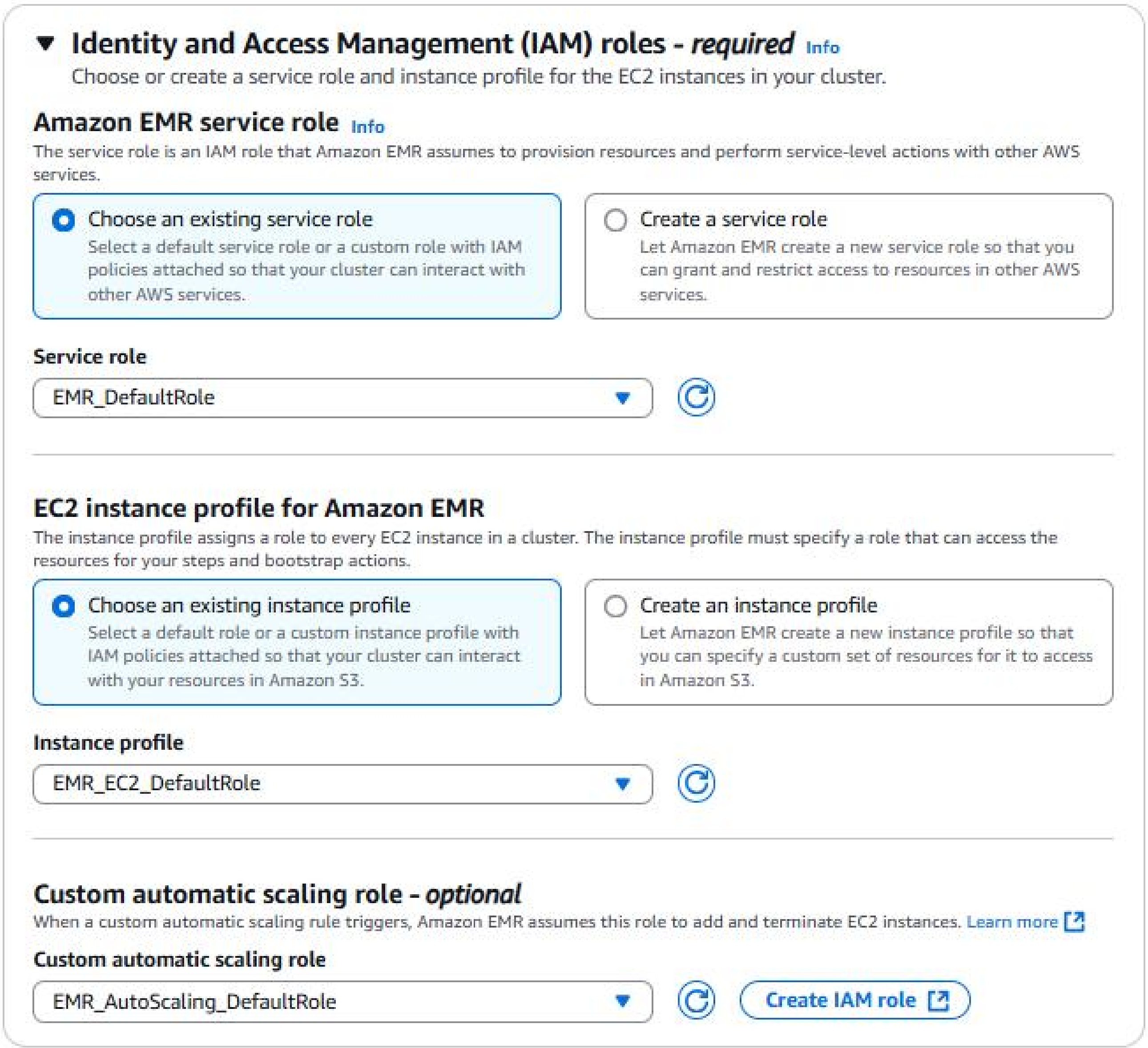
**Security configuration**

Select your cluster encryption, authentication, authorization, and instance metadata service settings.

Choose a security configuration

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

vockey



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

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

**Amazon EMR service role** Info

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**

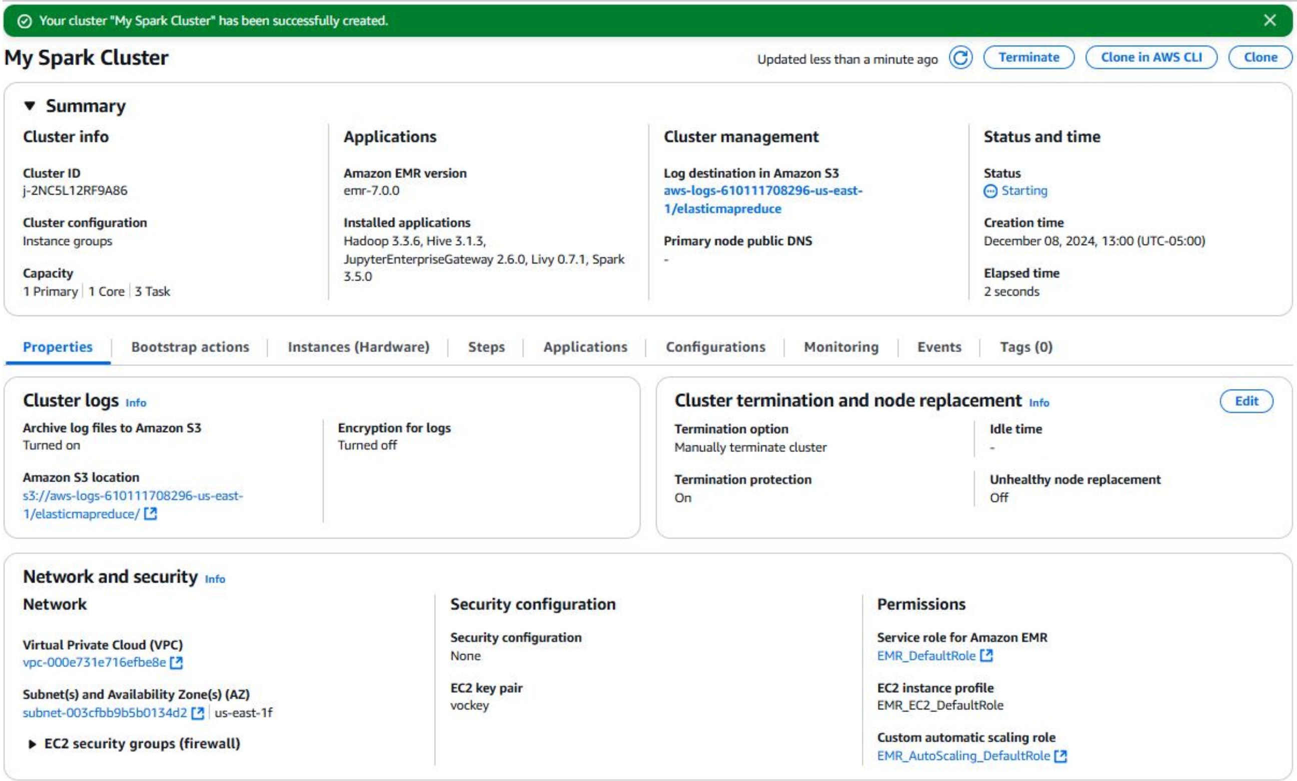
EMR_EC2_DefaultRole

**Custom automatic scaling role - *optional***

When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. Learn more

**Custom automatic scaling role**

EMR_AutoScaling_DefaultRole



Your cluster "My Spark Cluster" has been successfully created.

## My Spark Cluster

▼ Summary

**Cluster info**

Cluster ID
j-2NC5L12RF9A86

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 3 Task

**Applications**

Amazon EMR version
emr-7.0.0

Installed applications
Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Spark 3.5.0

**Cluster management**

Log destination in Amazon S3
aws-logs-610111708296-us-east-1/elasticmapreduce

Primary node public DNS
-

**Status and time**

Status
Starting

Creation time
December 08, 2024, 13:00 (UTC-05:00)

Elapsed time
2 seconds

**Cluster logs** Info

Archive log files to Amazon S3
Turned on

Amazon S3 location
s3://aws-logs-610111708296-us-east-1/elasticmapreduce/

Encryption for logs
Turned off

**Cluster termination and node replacement** Info

Termination option
Manually terminate cluster

Termination protection
On

Idle time
-

Unhealthy node replacement
Off

**Network and security** Info

**Network**

Virtual Private Cloud (VPC)
vpc-000e731e716efbe8e

Subnet(s) and Availability Zone(s) (AZ)
subnet-003cfbb9b5b0134d2 | us-east-1f

▶ EC2 security groups (firewall)

**Security configuration**

Security configuration
None

EC2 key pair
vockey

**Permissions**

Service role for Amazon EMR
EMR_DefaultRole

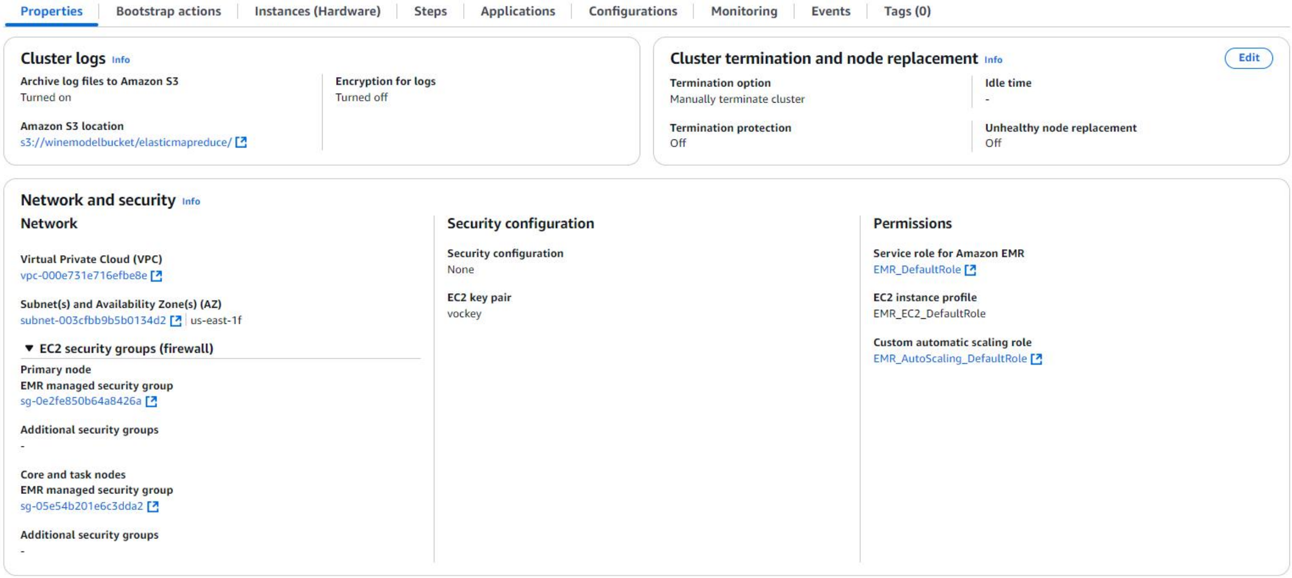
EC2 instance profile
EMR_EC2_DefaultRole

Custom automatic scaling role
EMR_AutoScaling_DefaultRole

## Step-3: Modify the security rules for the cluster

Now needs to update the inbound security rules

Do to the Network and Security -> Expand EC2 security groups -> click on core and task nodes



Cluster logs

Archive log files to Amazon S3: Turned on

Encryption for logs: Turned off

Amazon S3 location: s3://winemodelbucket/elasticmapreduce/

Cluster termination and node replacement

Termination option: Manually terminate cluster

Idle time: -

Termination protection: Off

Unhealthy node replacement: Off

Network and security

Network

Virtual Private Cloud (VPC): vpc-000e731e716efbe8e

Subnet(s) and Availability Zone(s) (AZ): subnet-003cfbb9b5b0134d2 | us-east-1f

▼ EC2 security groups (firewall)

Primary node

EMR managed security group: sg-0e2fe850b64a8426a

Additional security groups: -

Core and task nodes

EMR managed security group: sg-05e54b201e6c3dda2

Additional security groups: -

Security configuration

Security configuration: None

EC2 key pair: vockey

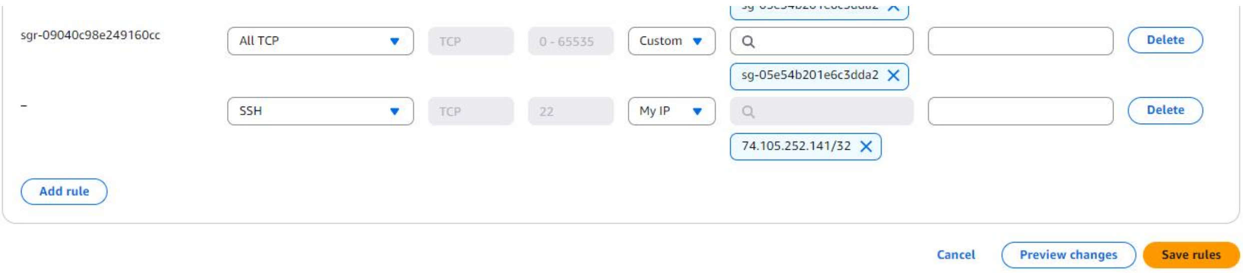
Permissions

Service role for Amazon EMR: EMR_DefaultRole

EC2 instance profile: EMR_EC2_DefaultRole

Custom automatic scaling role: EMR_AutoScaling_DefaultRole

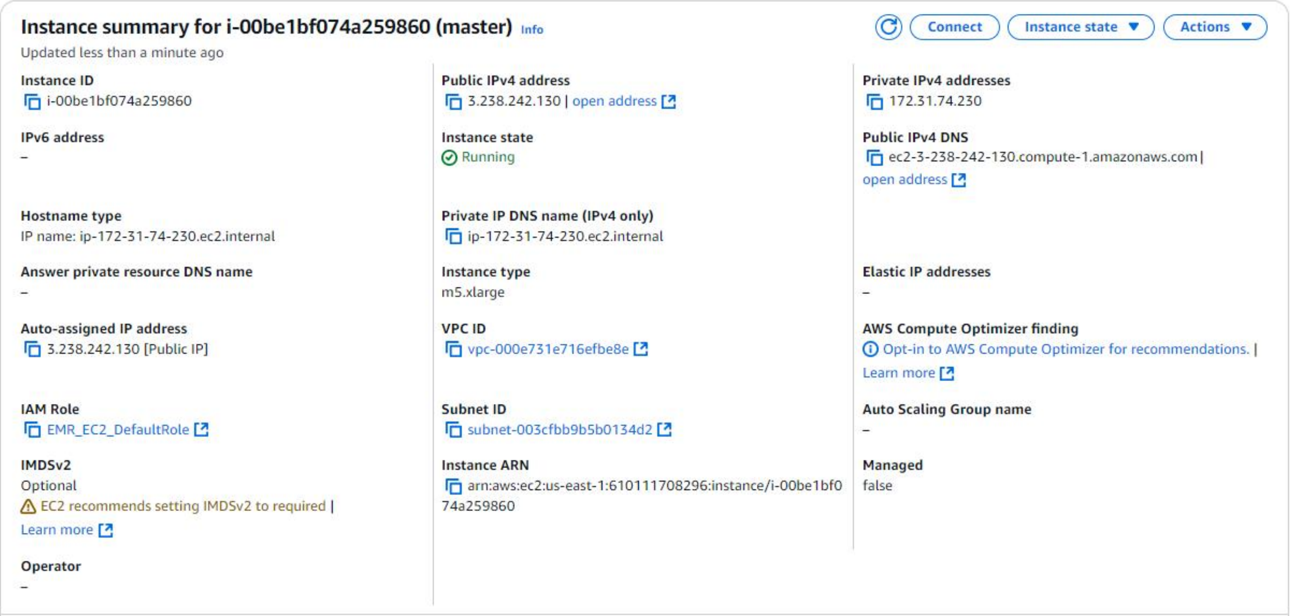
Click on edit inbound rules and add as shown in screenshot



| | | | | | | |
|---|---|---|---|---|---|---|
| sgr-09040c98e249160cc | All TCP | TCP | 0 - 65535 | Custom | sg-05e54b201e6c3dda2 | |
| - | SSH | TCP | 22 | My IP | 74.105.252.141/32 | |

## Step-4: Login to the master node and transfer all the required files into it

After selecting master node click on connect and



**Instance summary for i-00be1bf074a259860 (master)**

Updated less than a minute ago

**Instance ID**
i-00be1bf074a259860

**IPv6 address**
–

**Hostname type**
IP name: ip-172-31-74-230.ec2.internal

**Answer private resource DNS name**
–

**Auto-assigned IP address**
3.238.242.130 [Public IP]

**IAM Role**
EMR_EC2_DefaultRole

**IMDSv2**
Optional
⚠ EC2 recommends setting IMDSv2 to required | Learn more

**Operator**
–

**Public IPv4 address**
3.238.242.130 | open address

**Instance state**
Running

**Private IP DNS name (IPv4 only)**
ip-172-31-74-230.ec2.internal

**Instance type**
m5.xlarge

**VPC ID**
vpc-000e731e716efbe8e

**Subnet ID**
subnet-003cfbb9b5b0134d2

**Instance ARN**
arn:aws:ec2:us-east-1:610111708296:instance/i-00be1bf074a259860

**Private IPv4 addresses**
172.31.74.230

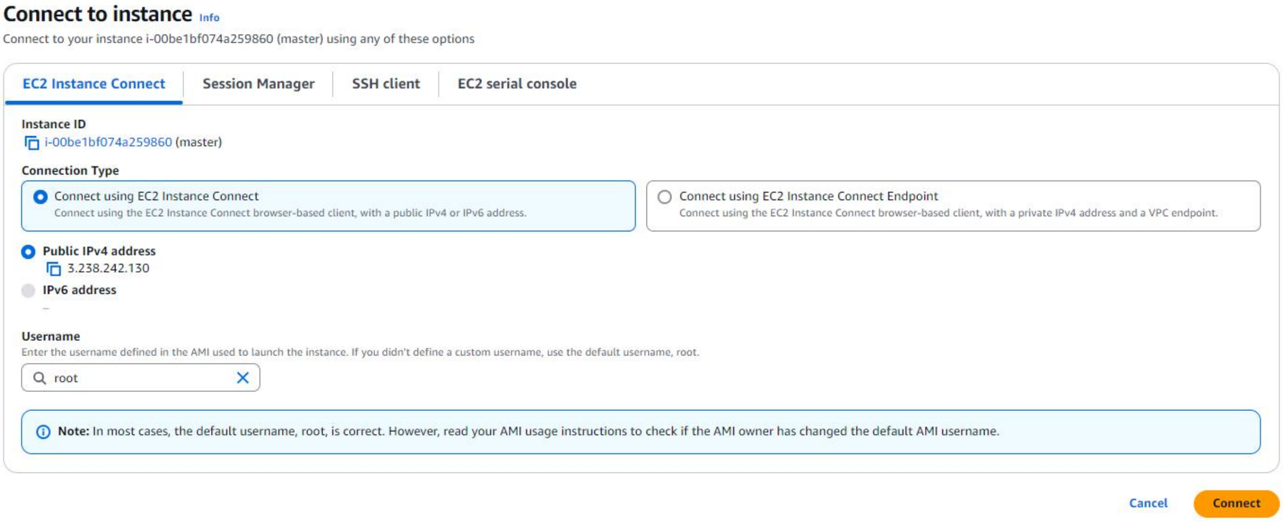
**Public IPv4 DNS**
ec2-3-238-242-130.compute-1.amazonaws.com | open address

**Elastic IP addresses**
–

**AWS Compute Optimizer finding**
Opt-in to AWS Compute Optimizer for recommendations. | Learn more

**Auto Scaling Group name**
–

**Managed**
false



**Connect to instance**

Connect to your instance i-00be1bf074a259860 (master) using any of these options

**Instance ID**
i-00be1bf074a259860 (master)

**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.

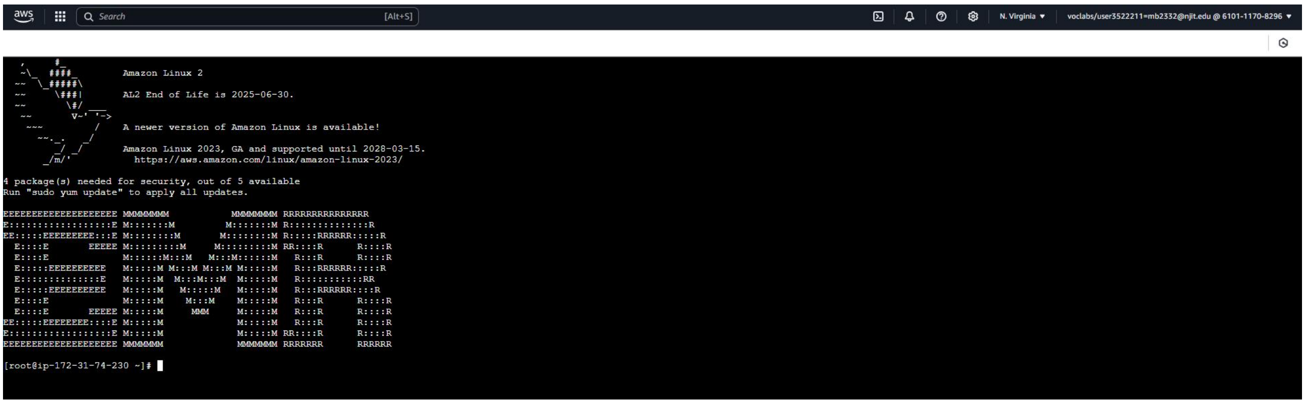
● Public IPv4 address
3.238.242.130

○ 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.



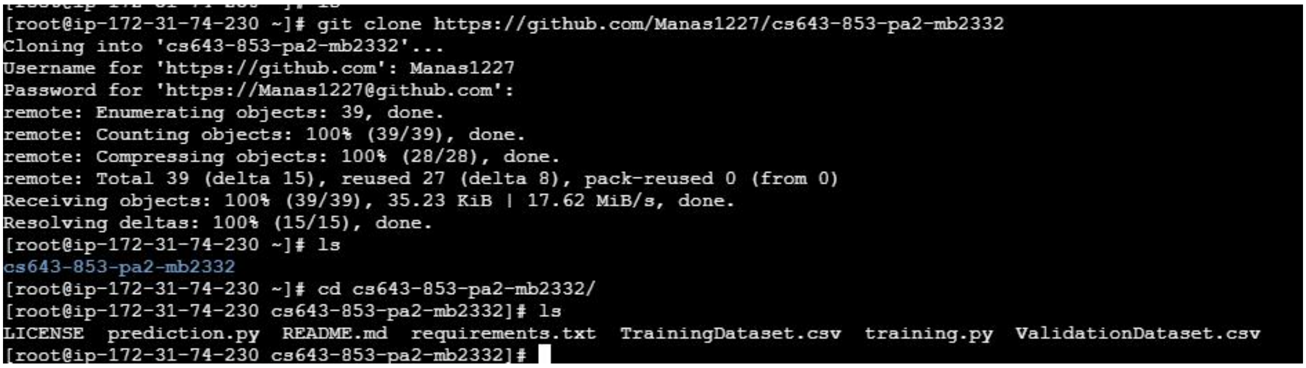
```
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

4 package(s) needed for security, out of 5 available
Run "sudo yum update" to apply all updates.

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

[root@ip-172-31-74-230 ~]# 
```

## Step-5: Execute the training.py ML application

Enter command to install git and enter yes when ask
sudo yum install git

Now clone with the github repository
git clone git clone https://github.com/Manas1227/cs643-853-pa2-mb2332

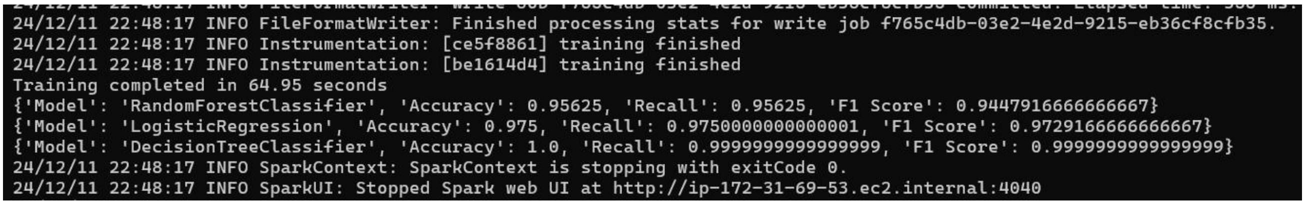
Now change to the new directory cs643-853-pa2-mb2332 and check it clone correctly
cd cs643-853-pa2-mb2332
ls



```
[root@ip-172-31-74-230 ~]# git clone https://github.com/Manas1227/cs643-853-pa2-mb2332
Cloning into 'cs643-853-pa2-mb2332'...
Username for 'https://github.com': Manas1227
Password for 'https://Manas1227@github.com':
remote: Enumerating objects: 39, done.
remote: Counting objects: 100% (39/39), done.
remote: Compressing objects: 100% (28/28), done.
remote: Total 39 (delta 15), reused 27 (delta 8), pack-reused 0 (from 0)
Receiving objects: 100% (39/39), 35.23 KiB | 17.62 MiB/s, done.
Resolving deltas: 100% (15/15), done.
[root@ip-172-31-74-230 ~]# ls
cs643-853-pa2-mb2332
[root@ip-172-31-74-230 ~]# cd cs643-853-pa2-mb2332/
[root@ip-172-31-74-230 cs643-853-pa2-mb2332]# ls
LICENSE  prediction.py  README.md  requirements.txt  TrainingDataset.csv  training.py  ValidationDataset.csv
[root@ip-172-31-74-230 cs643-853-pa2-mb2332]#
```

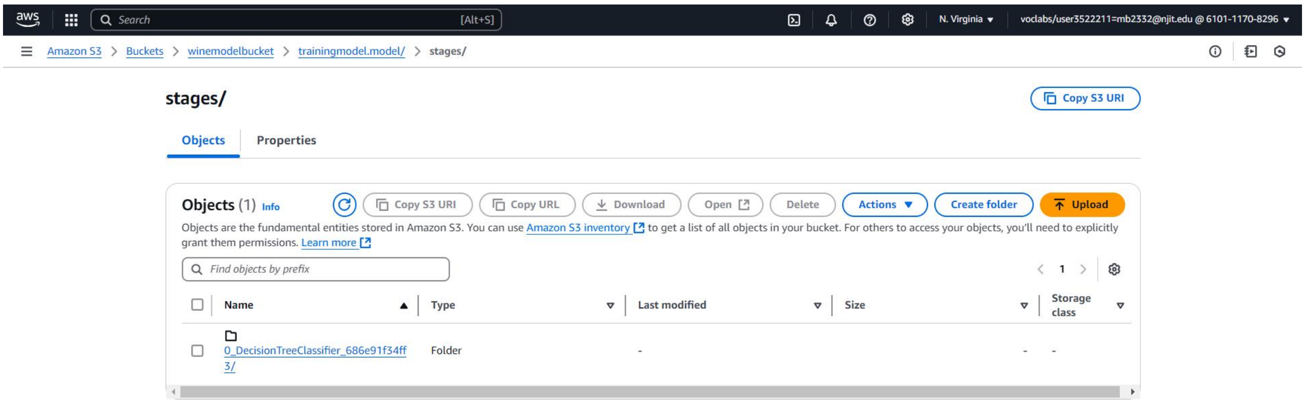
Now install all required libraries by followed command
sudo pip3 install -r requirements.txt

Run the training file
sudo spark-submit training.py



```
24/12/11 22:48:17 INFO FileFormatWriter: Finished processing stats for write job f765c4db-03e2-4e2d-9215-eb36cf8cfb35.
24/12/11 22:48:17 INFO Instrumentation: [ce5f8861] training finished
24/12/11 22:48:17 INFO Instrumentation: [be1614d4] training finished
Training completed in 64.95 seconds
{'Model': 'RandomForestClassifier', 'Accuracy': 0.95625, 'Recall': 0.95625, 'F1 Score': 0.9447916666666667}
{'Model': 'LogisticRegression', 'Accuracy': 0.975, 'Recall': 0.9750000000000001, 'F1 Score': 0.9729166666666667}
{'Model': 'DecisionTreeClassifier', 'Accuracy': 1.0, 'Recall': 0.9999999999999999, 'F1 Score': 0.9999999999999999}
24/12/11 22:48:17 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/11 22:48:17 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-69-53.ec2.internal:4040
```

Training model will train the ML model using EMR cluster and upload the best model to S3 bucket "trainingmodel.model"



Amazon S3 > Buckets > winemodelbucket > trainingmodel.model/ > stages/

stages/

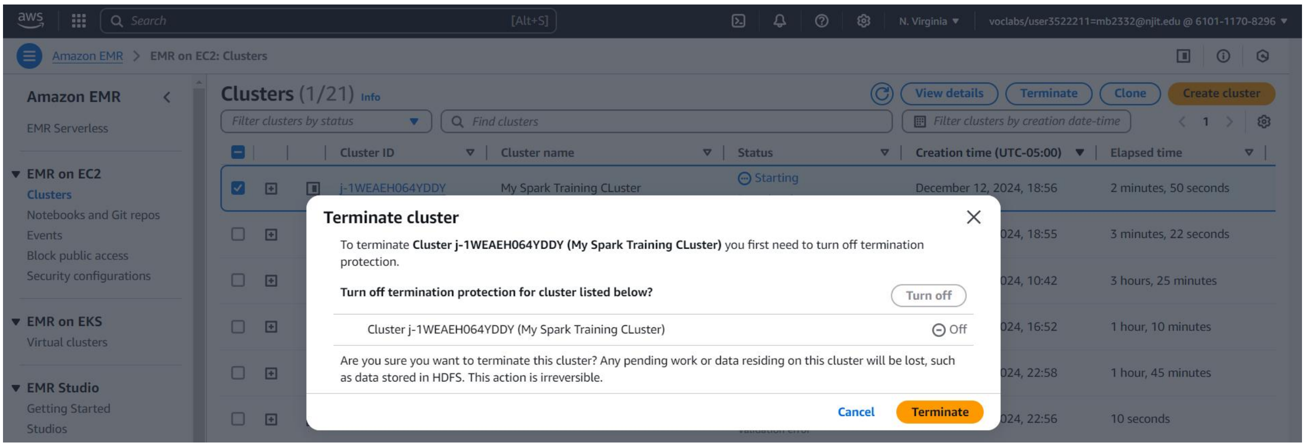
Objects (1) Info

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

| □ | Name | Type | Last modified | Size | Storage class |
|---|---|---|---|---|---|
| □ | 0_DecisionTreeClassifier_686e91f34ff3/ | Folder | - | - | - |

### Step-6: Terminate the EMR cluster

Terminate the EMR cluster after sucessfully competed our training part
If asked **Turn off** the protection and **Terminate** the cluster



| ☑ | Cluster ID | Cluster name | Status | Creation time (UTC-05:00) | Elapsed time |
|---|---|---|---|---|---|
| ☑ | j-1WEAEH064YDDY | My Spark Training CLuster | Starting | December 12, 2024, 18:56 | 2 minutes, 50 seconds |
| ☐ | | | | 024, 18:55 | 3 minutes, 22 seconds |
| ☐ | | | | 024, 10:42 | 3 hours, 25 minutes |
| ☐ | | | | 024, 16:52 | 1 hour, 10 minutes |
| ☐ | | | | 024, 22:58 | 1 hour, 45 minutes |
| ☐ | | | | 024, 22:56 | 10 seconds |

**Terminate cluster**

To terminate **Cluster j-1WEAEH064YDDY (My Spark Training CLuster)** you first need to turn off termination protection.

**Turn off termination protection for cluster listed below?** Turn off

Cluster j-1WEAEH064YDDY (My Spark Training CLuster) Off

Are you sure you want to terminate this cluster? Any pending work or data residing on this cluster will be lost, such as data stored in HDFS. This action is irreversible.

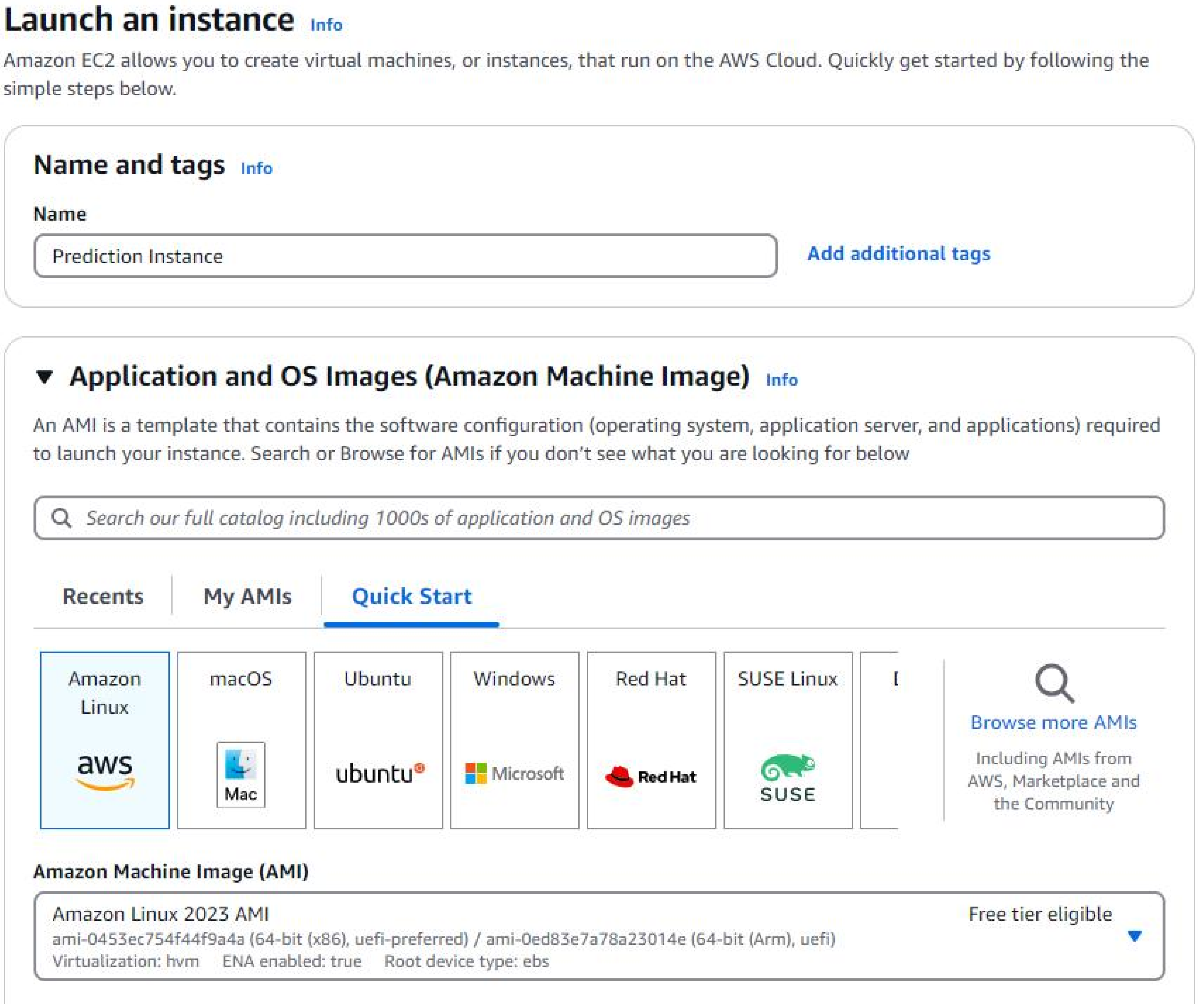
Cancel Terminate

## SECTION 2: AWS Cloud setup for running the prediction ML application - without Docker

Step-1: Create standalone EC2 instance on AWS

Redirect to **EC2** on AWS console and click on **Launch Instance**

Enter Name and Select Amazon Linux-2023 AMI



Launch an instance Info

Amazon EC2 allows you to create virtual machines, or instances, that run on the AWS Cloud. Quickly get started by following the simple steps below.

Name and tags Info

Name: Prediction Instance

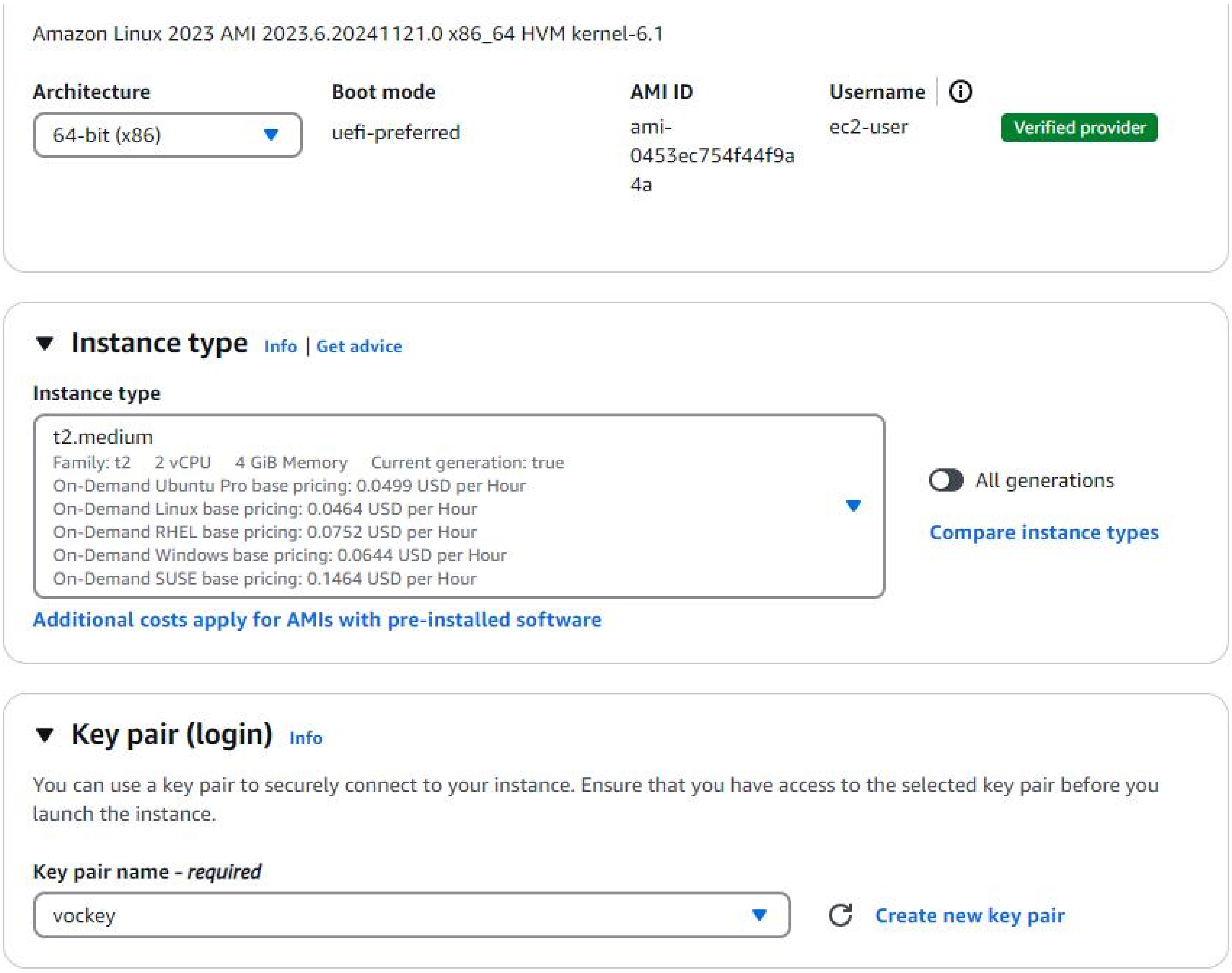
Application and OS Images (Amazon Machine Image) Info

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. Search or Browse for AMIs if you don't see what you are looking for below

Amazon Machine Image (AMI)

Amazon Linux 2023 AMI — Free tier eligible
ami-0453ec754f44f9a4a (64-bit (x86), uefi-preferred) / ami-0ed83e7a78a23014e (64-bit (Arm), uefi)
Virtualization: hvm ENA enabled: true Root device type: ebs

Select t2.medium as a Instance type



Amazon Linux 2023 AMI 2023.6.20241121.0 x86_64 HVM kernel-6.1

| Architecture | Boot mode | AMI ID | Username |
|---|---|---|---|
| 64-bit (x86) | uefi-preferred | ami-0453ec754f44f9a4a | ec2-user |

▼ Instance type

Instance type

t2.medium
Family: t2 2 vCPU 4 GiB Memory Current generation: true
On-Demand Ubuntu Pro base pricing: 0.0499 USD per Hour
On-Demand Linux base pricing: 0.0464 USD per Hour
On-Demand RHEL base pricing: 0.0752 USD per Hour
On-Demand Windows base pricing: 0.0644 USD per Hour
On-Demand SUSE base pricing: 0.1464 USD per Hour

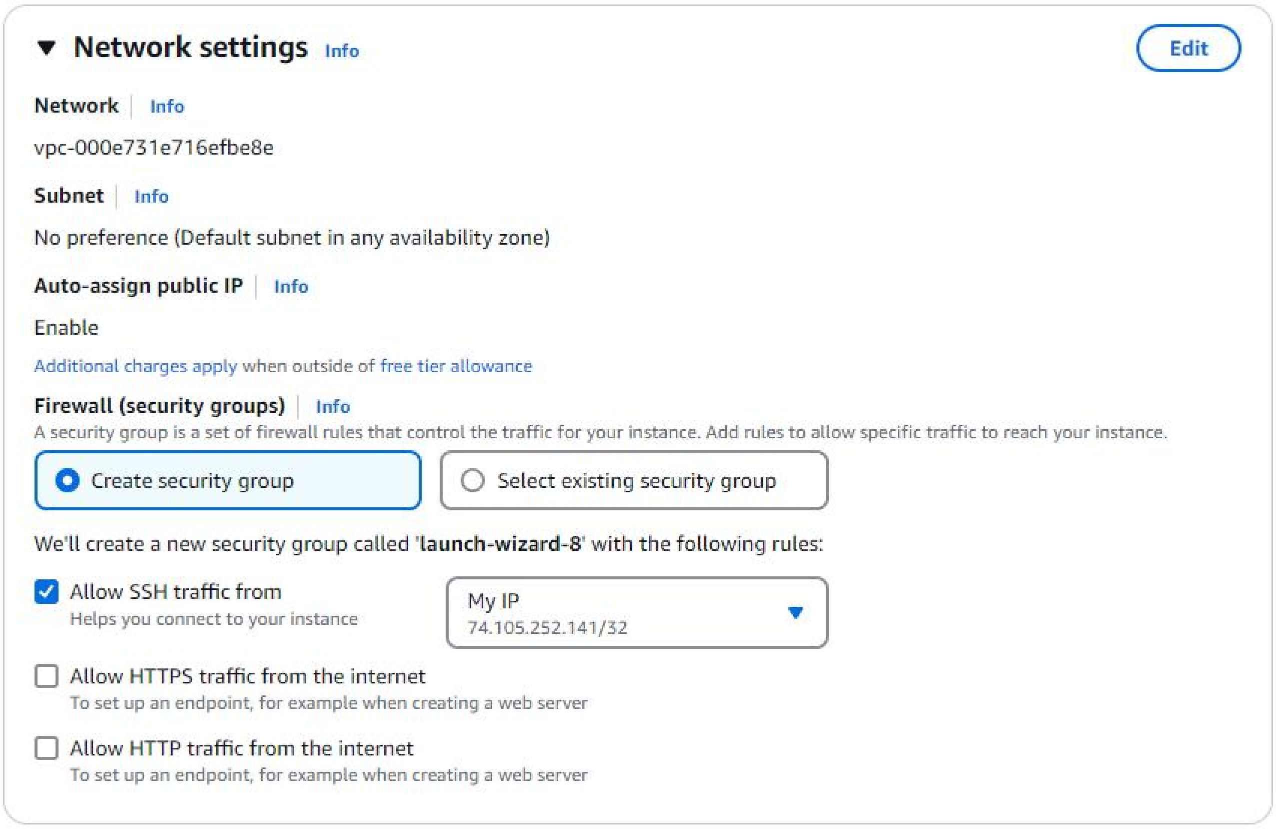
▼ Key pair (login)

You can use a key pair to securely connect to your instance. Ensure that you have access to the selected key pair before you launch the instance.

Key pair name - *required*

vockey

Create or choose from existing security group to inbound rule to accept ssh connection from MyIp
Also configure with 15 GB storage as we are going to install all the required dependencies



▼ Network settings Info

Edit

Network | Info

vpc-000e731e716efbe8e

Subnet | Info

No preference (Default subnet in any availability zone)

Auto-assign public IP | Info

Enable

Additional charges apply when outside of free tier allowance

Firewall (security groups) | Info

A security group is a set of firewall rules that control the traffic for your instance. Add rules to allow specific traffic to reach your instance.

● Create security group ○ Select existing security group

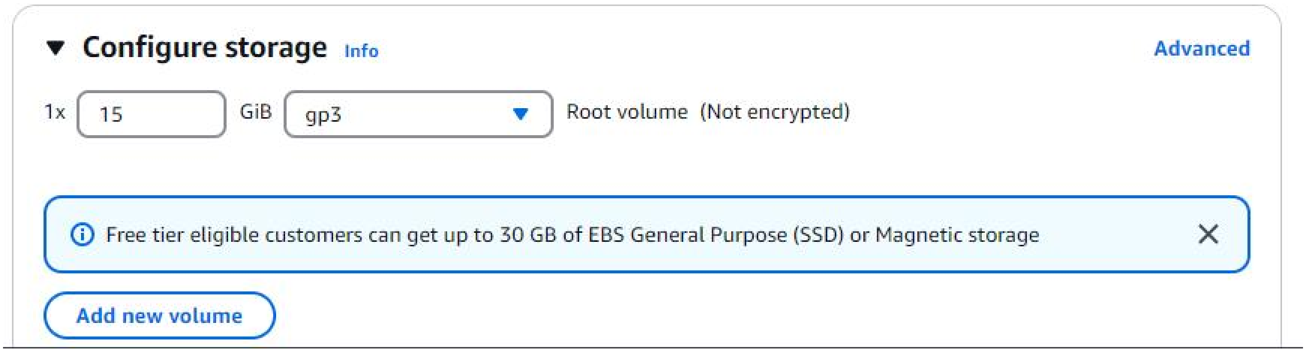
We'll create a new security group called '**launch-wizard-8**' with the following rules:

☑ Allow SSH traffic from
Helps you connect to your instance

My IP
74.105.252.141/32

□ Allow HTTPS traffic from the internet
To set up an endpoint, for example when creating a web server

□ Allow HTTP traffic from the internet
To set up an endpoint, for example when creating a web server



▼ Configure storage Info

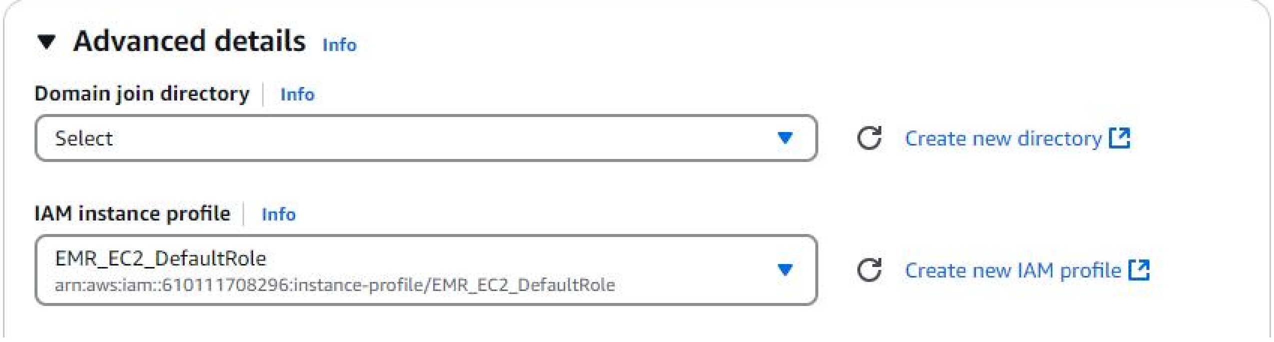
Advanced

1x 15 GiB gp3 Root volume (Not encrypted)

ⓘ Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage ✕

Add new volume

Make sure to select IAM role with all required AWS services access



▼ Advanced details Info

Domain join directory | Info

Select

Create new directory

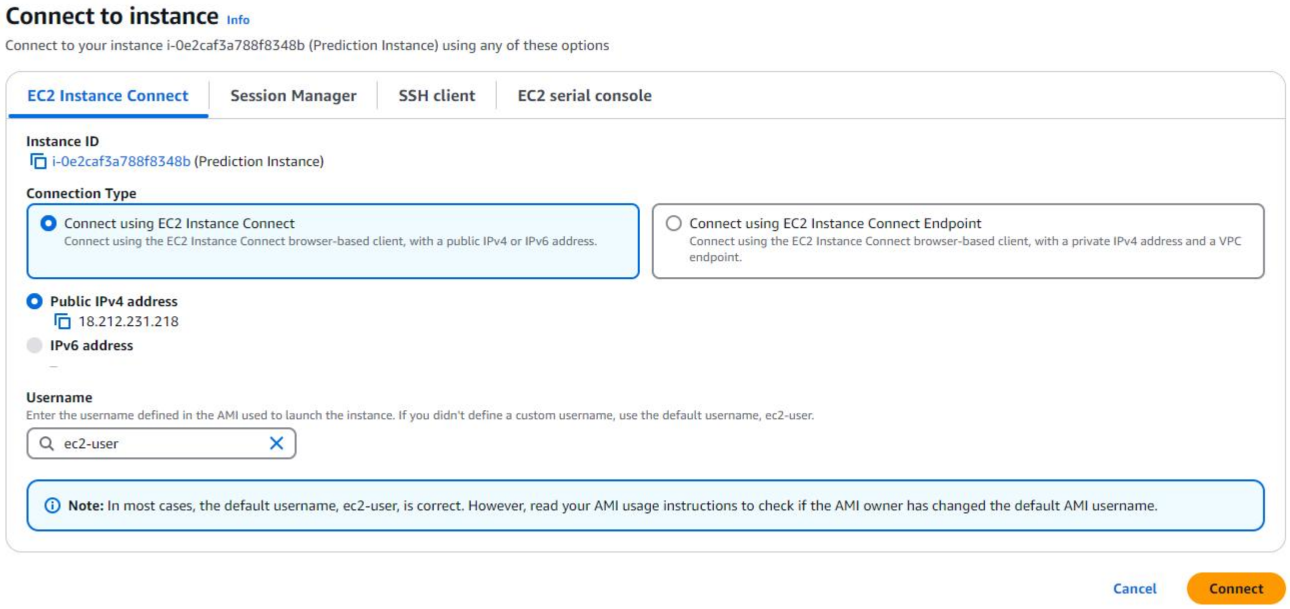
IAM instance profile | Info

EMR_EC2_DefaultRole
arn:aws:iam::610111708296:instance-profile/EMR_EC2_DefaultRole

Create new IAM profile

Step-2: Login to the EC2 instance and setup environment

Click on Connect and connect to the instance with your preferred options



**Connect to instance** Info

Connect to your instance i-0e2caf3a788f8348b (Prediction Instance) using any of these options

**Instance ID**
i-0e2caf3a788f8348b (Prediction Instance)

**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**
18.212.231.218

○ **IPv6 address**
–

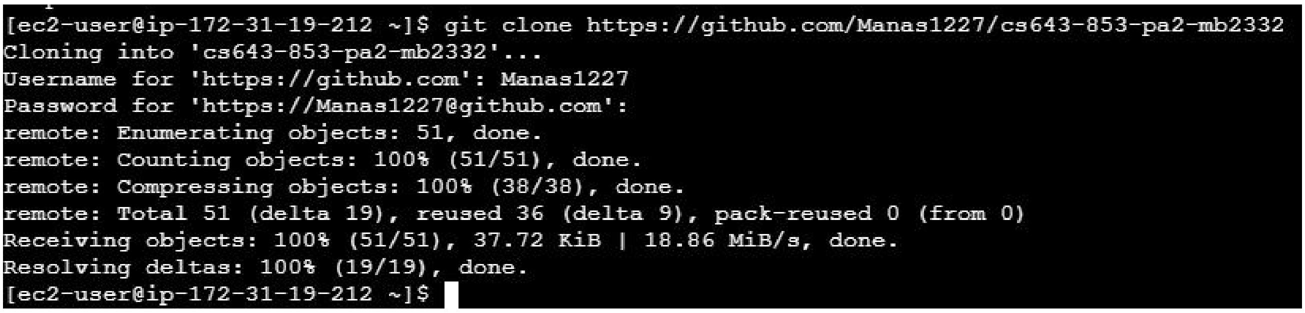
**Username**
Enter the username defined in the AMI used to launch the instance. If you didn't define a custom username, use the default username, ec2-user.

ec2-user

**Note:** In most cases, the default username, ec2-user, is correct. However, read your AMI usage instructions to check if the AMI owner has changed the default AMI username.

Now install Github to clone with the repository
sudo yum install git

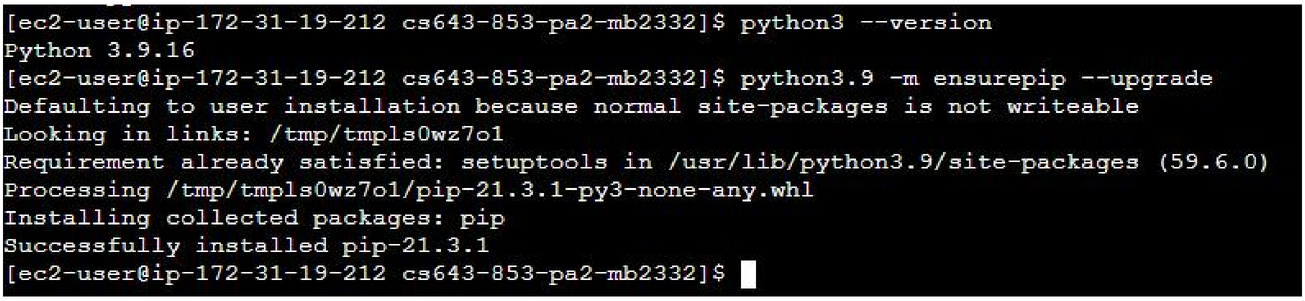
Clone with the github repository to access all the program files on EC2 instance
git clone https://github.com/Manas1227/cs643-853-pa2-mb2332



```
[ec2-user@ip-172-31-19-212 ~]$ git clone https://github.com/Manas1227/cs643-853-pa2-mb2332
Cloning into 'cs643-853-pa2-mb2332'...
Username for 'https://github.com': Manas1227
Password for 'https://Manas1227@github.com':
remote: Enumerating objects: 51, done.
remote: Counting objects: 100% (51/51), done.
remote: Compressing objects: 100% (38/38), done.
remote: Total 51 (delta 19), reused 36 (delta 9), pack-reused 0 (from 0)
Receiving objects: 100% (51/51), 37.72 KiB | 18.86 MiB/s, done.
Resolving deltas: 100% (19/19), done.
[ec2-user@ip-172-31-19-212 ~]$
```

Go to the new project directory
cd cs643-853-pa2-mb2332/

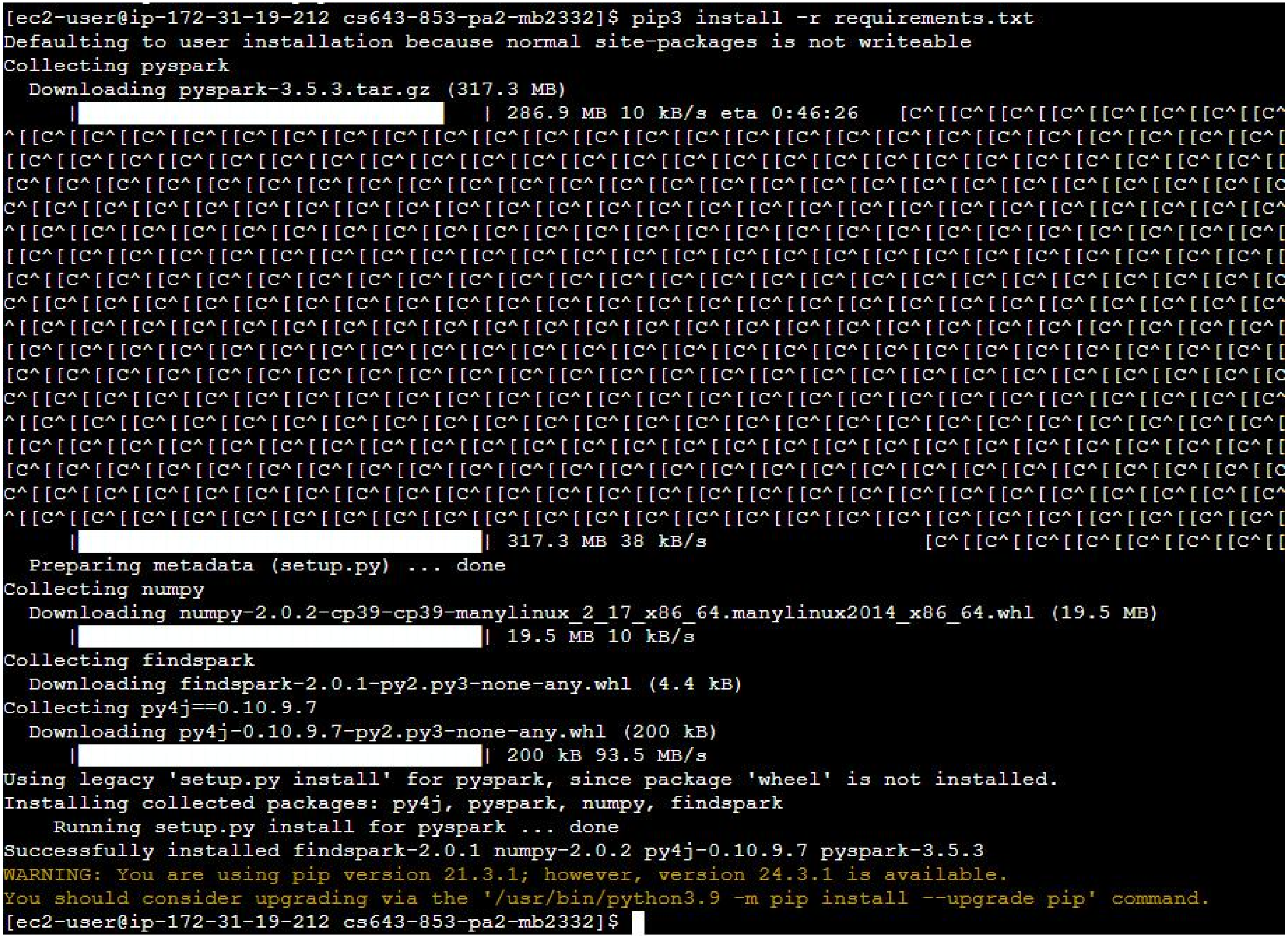
Check the Python and pip version and if not install them on the EC2
python3 --version
python3.9 -m ensurepip --upgrade



```
[ec2-user@ip-172-31-19-212 cs643-853-pa2-mb2332]$ python3 --version
Python 3.9.16
[ec2-user@ip-172-31-19-212 cs643-853-pa2-mb2332]$ python3.9 -m ensurepip --upgrade
Defaulting to user installation because normal site-packages is not writeable
Looking in links: /tmp/tmpls0wz7o1
Requirement already satisfied: setuptools in /usr/lib/python3.9/site-packages (59.6.0)
Processing /tmp/tmpls0wz7o1/pip-21.3.1-py3-none-any.whl
Installing collected packages: pip
Successfully installed pip-21.3.1
[ec2-user@ip-172-31-19-212 cs643-853-pa2-mb2332]$
```

Install all the requirements listed in requirements.txt file

pip3 install -r requirements.txt



```
[ec2-user@ip-172-31-19-212 cs643-853-pa2-mb2332]$ pip3 install -r requirements.txt
Defaulting to user installation because normal site-packages is not writeable
Collecting pyspark
  Downloading pyspark-3.5.3.tar.gz (317.3 MB)
     |█████████████████████████         | 286.9 MB 10 kB/s eta 0:46:26    [C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^
^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^
[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[
[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[
C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C
^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^
[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[
[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[
C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C
^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^
[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[
[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[
C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C
^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^
[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[
[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[
C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C
^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^[[C^
     |████████████████████████████████| 317.3 MB 38 kB/s              [C^[[C^[[C^[[C^[[C^[[C^[[
  Preparing metadata (setup.py) ... done
Collecting numpy
  Downloading numpy-2.0.2-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (19.5 MB)
     |████████████████████████████████| 19.5 MB 10 kB/s
Collecting findspark
  Downloading findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Collecting py4j==0.10.9.7
  Downloading py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
     |████████████████████████████████| 200 kB 93.5 MB/s
Using legacy 'setup.py install' for pyspark, since package 'wheel' is not installed.
Installing collected packages: py4j, pyspark, numpy, findspark
    Running setup.py install for pyspark ... done
Successfully installed findspark-2.0.1 numpy-2.0.2 py4j-0.10.9.7 pyspark-3.5.3
WARNING: You are using pip version 21.3.1; however, version 24.3.1 is available.
You should consider upgrading via the '/usr/bin/python3.9 -m pip install --upgrade pip' command.
[ec2-user@ip-172-31-19-212 cs643-853-pa2-mb2332]$ █
```

Install java on EC2 instance

sudo yum install java-11-amazon-corretto-devel

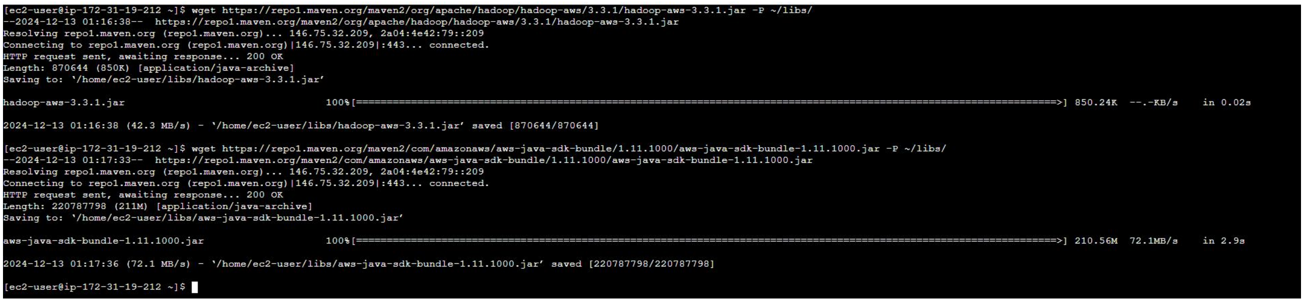
Add Java home and path variables in below file and save it

nano ~/.bashrc

export JAVA_HOME=/usr/lib/jvm/java-11-amazon-corretto.x86_64
export PATH=$PATH:$JAVA_HOME/bin

Now download required jars

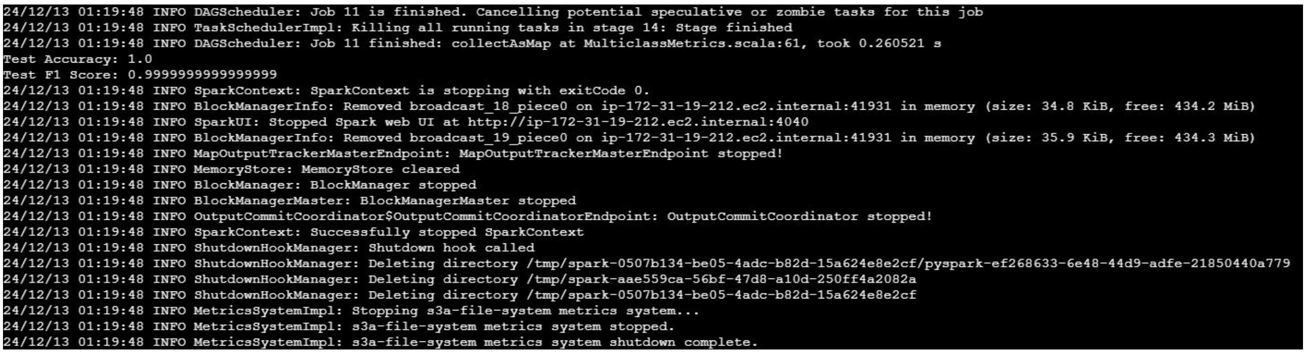
wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.1/hadoop-aws-3.3.1.jar -P ~/libs/
wget https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.11.1000/aws-java-sdk-bundle-1.11.1000.jar -P ~/libs/



```
[ec2-user@ip-172-31-19-212 ~]$ wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.1/hadoop-aws-3.3.1.jar -P ~/libs/
--2024-12-13 01:16:38--  https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.1/hadoop-aws-3.3.1.jar
Resolving repo1.maven.org (repo1.maven.org)... 146.75.32.209, 2a04:4e42:79::209
Connecting to repo1.maven.org (repo1.maven.org)|146.75.32.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 870644 (850K) [application/java-archive]
Saving to: '/home/ec2-user/libs/hadoop-aws-3.3.1.jar'

hadoop-aws-3.3.1.jar                  100%[===================================================================================>] 850.24K  --.-KB/s    in 0.02s

2024-12-13 01:16:38 (42.3 MB/s) - '/home/ec2-user/libs/hadoop-aws-3.3.1.jar' saved [870644/870644]

[ec2-user@ip-172-31-19-212 ~]$ wget https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.11.1000/aws-java-sdk-bundle-1.11.1000.jar -P ~/libs/
--2024-12-13 01:17:33--  https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.11.1000/aws-java-sdk-bundle-1.11.1000.jar
Resolving repo1.maven.org (repo1.maven.org)... 146.75.32.209, 2a04:4e42:79::209
Connecting to repo1.maven.org (repo1.maven.org)|146.75.32.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 220787798 (211M) [application/java-archive]
Saving to: '/home/ec2-user/libs/aws-java-sdk-bundle-1.11.1000.jar'

aws-java-sdk-bundle-1.11.1000.jar     100%[===================================================================================>] 210.56M  72.1MB/s    in 2.9s

2024-12-13 01:17:36 (72.1 MB/s) - '/home/ec2-user/libs/aws-java-sdk-bundle-1.11.1000.jar' saved [220787798/220787798]

[ec2-user@ip-172-31-19-212 ~]$ █
```

Now run the prediction.py file

spark-submit --jars /home/ec2-user/libs/hadoop-aws-3.3.1.jar,/home/ec2-user/libs/aws-java-sdk-bundle-1.11.1000.jar prediction.py



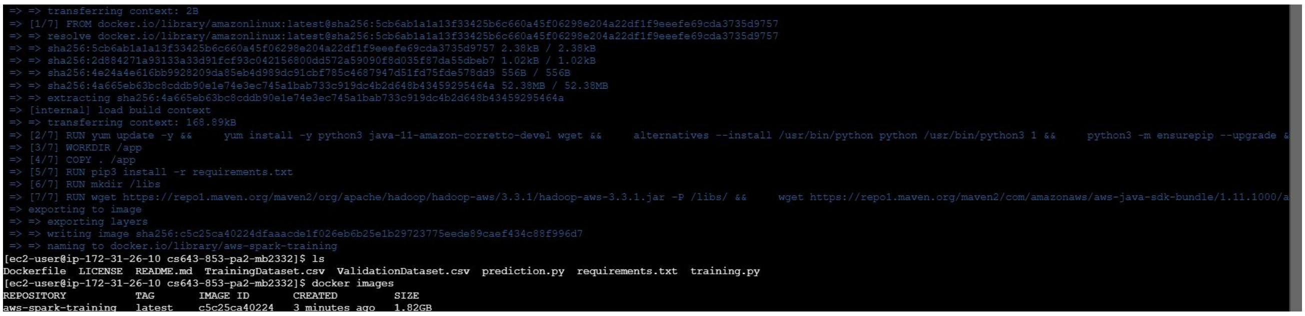
```
24/12/13 01:19:48 INFO DAGScheduler: Job 11 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/13 01:19:48 INFO TaskSchedulerImpl: Killing all running tasks in stage 14: Stage finished
24/12/13 01:19:48 INFO DAGScheduler: Job 11 finished: collectAsMap at MulticlassMetrics.scala:61, took 0.260521 s
Test Accuracy: 1.0
Test F1 Score: 0.9999999999999999
24/12/13 01:19:48 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/13 01:19:48 INFO BlockManagerInfo: Removed broadcast_18_piece0 on ip-172-31-19-212.ec2.internal:41931 in memory (size: 34.8 KiB, free: 434.2 MiB)
24/12/13 01:19:48 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-19-212.ec2.internal:4040
24/12/13 01:19:48 INFO BlockManagerInfo: Removed broadcast_19_piece0 on ip-172-31-19-212.ec2.internal:41931 in memory (size: 35.9 KiB, free: 434.3 MiB)
24/12/13 01:19:48 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/13 01:19:48 INFO MemoryStore: MemoryStore cleared
24/12/13 01:19:48 INFO BlockManager: BlockManager stopped
24/12/13 01:19:48 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/13 01:19:48 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/13 01:19:48 INFO SparkContext: Successfully stopped SparkContext
24/12/13 01:19:48 INFO ShutdownHookManager: Shutdown hook called
24/12/13 01:19:48 INFO ShutdownHookManager: Deleting directory /tmp/spark-0507b134-be05-4adc-b82d-15a624e8e2cf/pyspark-ef268633-6e48-44d9-adfe-21850440a779
24/12/13 01:19:48 INFO ShutdownHookManager: Deleting directory /tmp/spark-aae559ca-56bf-47d8-a10d-250ff4a2082a
24/12/13 01:19:48 INFO ShutdownHookManager: Deleting directory /tmp/spark-0507b134-be05-4adc-b82d-15a624e8e2cf
24/12/13 01:19:48 INFO MetricsSystemImpl: Stopping s3a-file-system metrics system...
24/12/13 01:19:48 INFO MetricsSystemImpl: s3a-file-system metrics system stopped.
24/12/13 01:19:48 INFO MetricsSystemImpl: s3a-file-system metrics system shutdown complete.
```

As shown above, DecisionTree Model from s3 was used to predict the label for ValidationDataset.csv

## SECTION 3: Create Docker image and push it on Docker repository

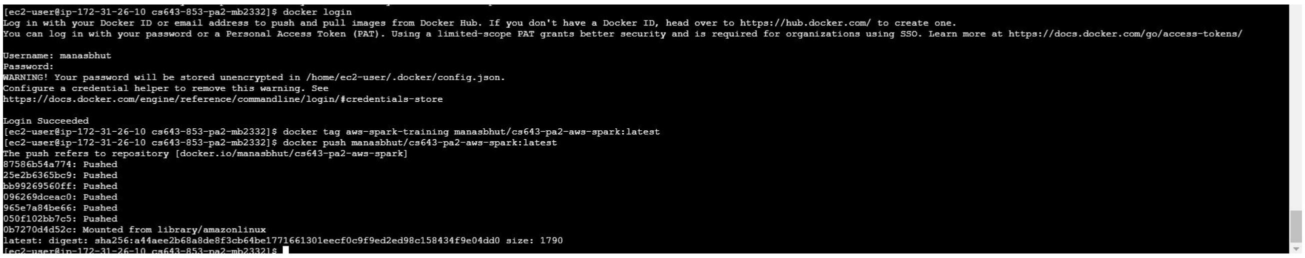
### Step-1: Create docker image by following command:

docker build -t aws-spark-training .
docker images



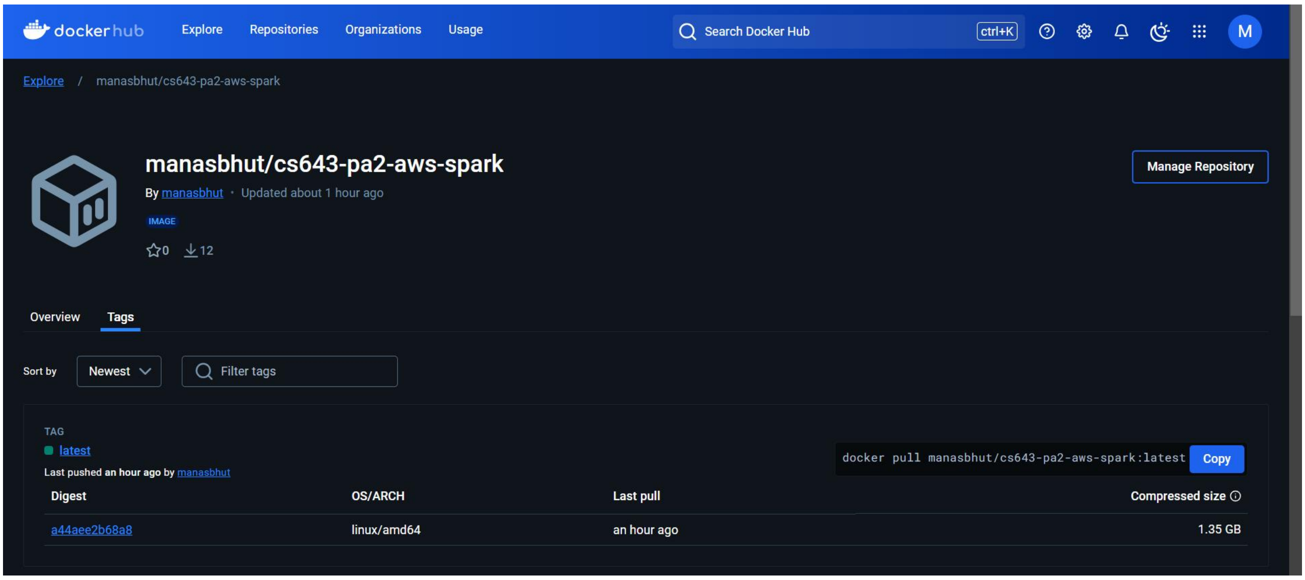
```
=> => transferring context: 2B
=> [1/7] FROM docker.io/library/amazonlinux:latest@sha256:5cb6ab1a1a13f33425b6c660a45f06298e204a22df1f9eeefe69cda3735d9757
=> => resolve docker.io/library/amazonlinux:latest@sha256:5cb6ab1a1a13f33425b6c660a45f06298e204a22df1f9eeefe69cda3735d9757
=> => sha256:5cb6ab1a1a13f33425b6c660a45f06298e204a22df1f9eeefe69cda3735d9757 2.38kB / 2.38kB
=> => sha256:2d884271a93133a33d91fcf93c042156800dd572a59090f8d035f87da55dbeb7 1.02kB / 1.02kB
=> => sha256:4e24a4e616bb9928209da85eb4d989dc91cbf785c4687947d51fd75fde578dd9 556B / 556B
=> => sha256:4a665eb63bc8cddb90e1e74e3ec745a1bab733c919dc4b2d648b43459295464a 52.38MB / 52.38MB
=> => extracting sha256:4a665eb63bc8cddb90e1e74e3ec745a1bab733c919dc4b2d648b43459295464a
=> [internal] load build context
=> => transferring context: 168.89kB
=> [2/7] RUN yum update -y &&     yum install -y python3 java-11-amazon-corretto-devel wget &&     alternatives --install /usr/bin/python python /usr/bin/python3 1 &&     python3 -m ensurepip --upgrade &
=> [3/7] WORKDIR /app
=> [4/7] COPY . /app
=> [5/7] RUN pip3 install -r requirements.txt
=> [6/7] RUN mkdir /libs
=> [7/7] RUN wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.1/hadoop-aws-3.3.1.jar -P /libs/ &&     wget https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.11.1000/a
=> exporting to image
=> => exporting layers
=> => writing image sha256:c5c25ca40224dfaaacde1f026eb6b25e1b29723775eede89caef434c88f996d7
=> => naming to docker.io/library/aws-spark-training
[ec2-user@ip-172-31-26-10 cs643-853-pa2-mb2332]$ ls
Dockerfile  LICENSE  README.md  TrainingDataset.csv  ValidationDataset.csv  prediction.py  requirements.txt  training.py
[ec2-user@ip-172-31-26-10 cs643-853-pa2-mb2332]$ docker images
REPOSITORY           TAG       IMAGE ID       CREATED         SIZE
aws-spark-training   latest    c5c25ca40224   3 minutes ago   1.82GB
```

### Step-2: Upload docker image to docker repository



```
[ec2-user@ip-172-31-26-10 cs643-853-pa2-mb2332]$ docker login
Log in with your Docker ID or email address to push and pull images from Docker Hub. If you don't have a Docker ID, head over to https://hub.docker.com/ to create one.
You can log in with your password or a Personal Access Token (PAT). Using a limited-scope PAT grants better security and is required for organizations using SSO. Learn more at https://docs.docker.com/go/access-tokens/

Username: manasbhut
Password:
WARNING! Your password will be stored unencrypted in /home/ec2-user/.docker/config.json.
Configure a credential helper to remove this warning. See
https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
[ec2-user@ip-172-31-26-10 cs643-853-pa2-mb2332]$ docker tag aws-spark-training manasbhut/cs643-pa2-aws-spark:latest
[ec2-user@ip-172-31-26-10 cs643-853-pa2-mb2332]$ docker push manasbhut/cs643-pa2-aws-spark:latest
The push refers to repository [docker.io/manasbhut/cs643-pa2-aws-spark]
87586b54a774: Pushed
25e2b6365bc9: Pushed
bb99269560ff: Pushed
096269dceac0: Pushed
965e7a84be66: Pushed
050f102bb7c5: Pushed
0b7270d4d52c: Mounted from library/amazonlinux
latest: digest: sha256:a44aee2b68a8de8f3cb64be1771661301eecf0c9f9ed2ed98c158434f9e04dd0 size: 1790
[ec2-user@ip-172-31-26-10 cs643-853-pa2-mb2332]$
```

### Step-3: Go to docker repository to verify the uploaded image



manasbhut/cs643-pa2-aws-spark

By manasbhut · Updated about 1 hour ago

TAG

🟩 latest

Last pushed an hour ago by manasbhut

docker pull manasbhut/cs643-pa2-aws-spark:latest

| Digest | OS/ARCH | Last pull | Compressed size |
|---|---|---|---|
| a44aee2b68a8 | linux/amd64 | an hour ago | 1.35 GB |

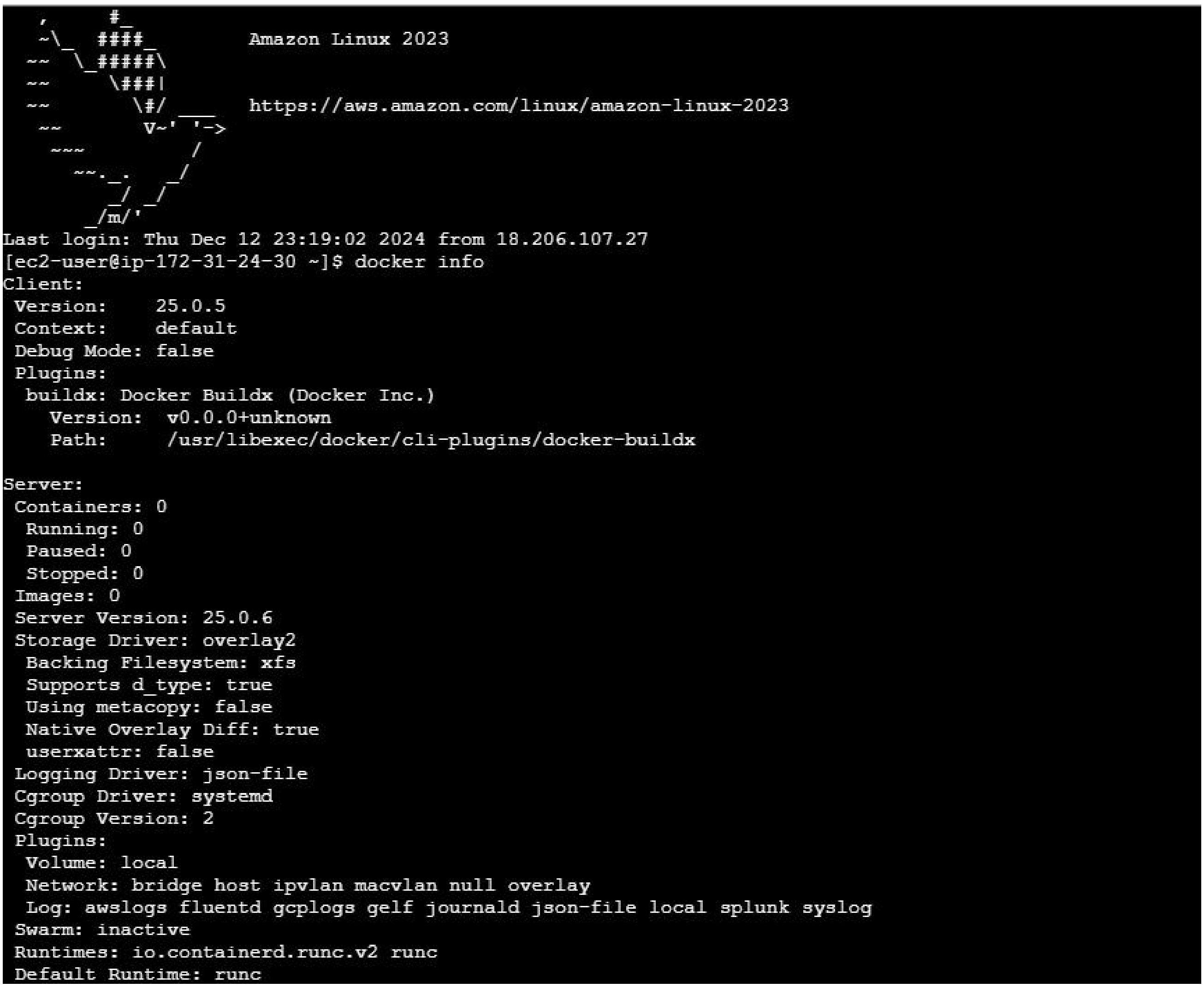
## SECTION 4: Run docker image on newly created EC2 instance

**Step-1: Install and run docker on EC2 instance**

sudo yum install docker -y
sudo service docker start
sudo usermod -aG docker ec2-user

**Step-2: Verify by running below command**

docker info



```
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Thu Dec 12 23:19:02 2024 from 18.206.107.27
[ec2-user@ip-172-31-24-30 ~]$ docker info
Client:
 Version:    25.0.5
 Context:    default
 Debug Mode: false
 Plugins:
  buildx: Docker Buildx (Docker Inc.)
    Version:  v0.0.0+unknown
    Path:     /usr/libexec/docker/cli-plugins/docker-buildx

Server:
 Containers: 0
  Running: 0
  Paused: 0
  Stopped: 0
 Images: 0
 Server Version: 25.0.6
 Storage Driver: overlay2
  Backing Filesystem: xfs
  Supports d_type: true
  Using metacopy: false
  Native Overlay Diff: true
  userxattr: false
 Logging Driver: json-file
 Cgroup Driver: systemd
 Cgroup Version: 2
 Plugins:
  Volume: local
  Network: bridge host ipvlan macvlan null overlay
  Log: awslogs fluentd gcplogs gelf journald json-file local splunk syslog
 Swarm: inactive
 Runtimes: io.containerd.runc.v2 runc
 Default Runtime: runc
```

If it still giving error than try to exit the Ec2 instance and login again

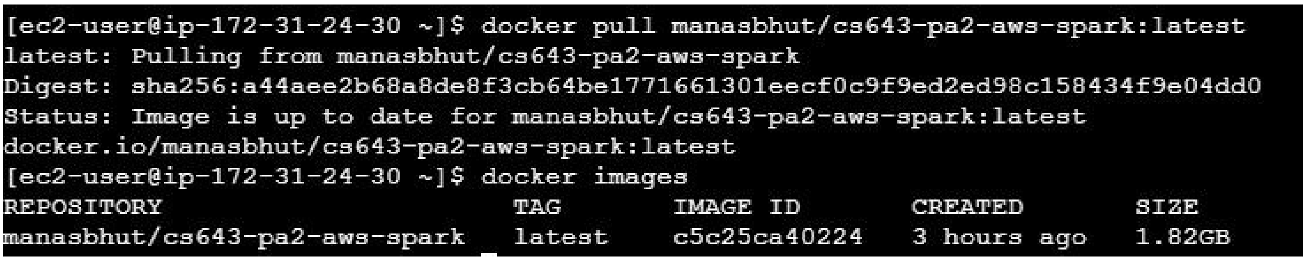
exit

Now login again to the Ec2 instance and verify docker is up and running

docker info

Pull the latest image from Docker repository

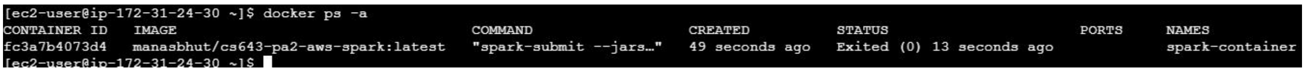
docker pull manasbhut/cs643-pa2-aws-spark:latest
docker images



```
[ec2-user@ip-172-31-24-30 ~]$ docker pull manasbhut/cs643-pa2-aws-spark:latest
latest: Pulling from manasbhut/cs643-pa2-aws-spark
Digest: sha256:a44aee2b68a8de8f3cb64be1771661301eecf0c9f9ed2ed98c158434f9e04dd0
Status: Image is up to date for manasbhut/cs643-pa2-aws-spark:latest
docker.io/manasbhut/cs643-pa2-aws-spark:latest
[ec2-user@ip-172-31-24-30 ~]$ docker images
REPOSITORY                      TAG       IMAGE ID       CREATED         SIZE
manasbhut/cs643-pa2-aws-spark   latest    c5c25ca40224   3 hours ago     1.82GB
```

Run the pulled image by following command

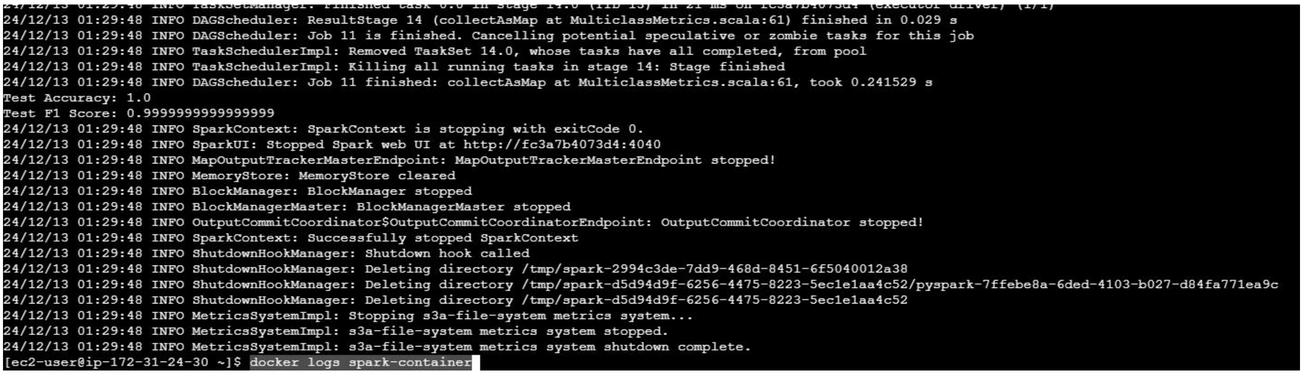
docker run -d --name spark-container manasbhut/cs643-pa2-aws-spark:latest
docker ps -a



```
[ec2-user@ip-172-31-24-30 ~]$ docker ps -a
CONTAINER ID   IMAGE                                  COMMAND                  CREATED          STATUS                      PORTS     NAMES
fc3a7b4073d4   manasbhut/cs643-pa2-aws-spark:latest   "spark-submit --jars…"   49 seconds ago   Exited (0) 13 seconds ago             spark-container
[ec2-user@ip-172-31-24-30 ~]$
```

Run following command to see the console output

docker logs spark-container



```
24/12/13 01:29:48 INFO TaskSetManager: Finished task 0.0 in stage 14.0 (TID 15) in 21 ms on fc3a7b4073d4 (executor driver) (1/1)
24/12/13 01:29:48 INFO DAGScheduler: ResultStage 14 (collectAsMap at MulticlassMetrics.scala:61) finished in 0.029 s
24/12/13 01:29:48 INFO DAGScheduler: Job 11 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/13 01:29:48 INFO TaskSchedulerImpl: Removed TaskSet 14.0, whose tasks have all completed, from pool
24/12/13 01:29:48 INFO TaskSchedulerImpl: Killing all running tasks in stage 14: Stage finished
24/12/13 01:29:48 INFO DAGScheduler: Job 11 finished: collectAsMap at MulticlassMetrics.scala:61, took 0.241529 s
Test Accuracy: 1.0
Test F1 Score: 0.9999999999999999
24/12/13 01:29:48 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/13 01:29:48 INFO SparkUI: Stopped Spark web UI at http://fc3a7b4073d4:4040
24/12/13 01:29:48 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/13 01:29:48 INFO MemoryStore: MemoryStore cleared
24/12/13 01:29:48 INFO BlockManager: BlockManager stopped
24/12/13 01:29:48 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/13 01:29:48 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/13 01:29:48 INFO SparkContext: Successfully stopped SparkContext
24/12/13 01:29:48 INFO ShutdownHookManager: Shutdown hook called
24/12/13 01:29:48 INFO ShutdownHookManager: Deleting directory /tmp/spark-2994c3de-7dd9-468d-8451-6f5040012a38
24/12/13 01:29:48 INFO ShutdownHookManager: Deleting directory /tmp/spark-d5d94d9f-6256-4475-8223-5ec1e1aa4c52/pyspark-7ffebe8a-6ded-4103-b027-d84fa771ea9c
24/12/13 01:29:48 INFO ShutdownHookManager: Deleting directory /tmp/spark-d5d94d9f-6256-4475-8223-5ec1e1aa4c52
24/12/13 01:29:48 INFO MetricsSystemImpl: Stopping s3a-file-system metrics system...
24/12/13 01:29:48 INFO MetricsSystemImpl: s3a-file-system metrics system stopped.
24/12/13 01:29:48 INFO MetricsSystemImpl: s3a-file-system metrics system shutdown complete.
[ec2-user@ip-172-31-24-30 ~]$ docker logs spark-container
```