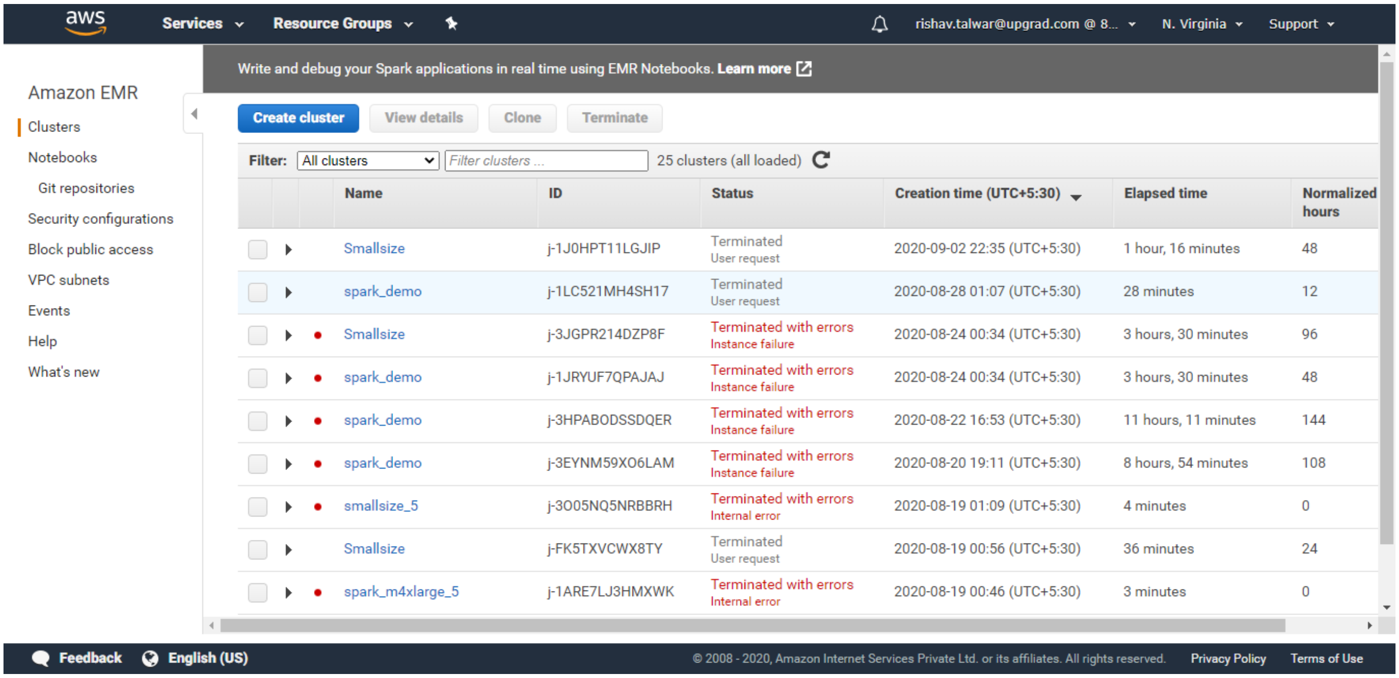
Creating an EMR Cluster on AWS

Before starting to create an EMR cluster, you will have to go to the EMR home page on your AWS account.

Once you arrive at the EMR home page, it should look something like this:



Write and debug your Spark applications in real time using EMR Notebooks. **Learn more**

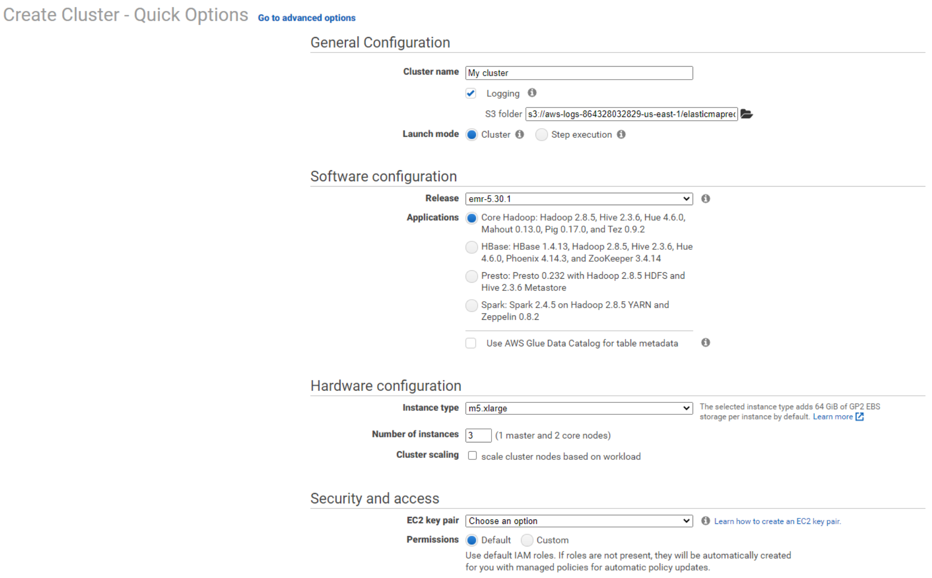
Filter: All clusters | Filter clusters ... | 25 clusters (all loaded)

			Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalized hours
□	▶		Smallsize	j-1J0HPT11LGJIP	Terminated User request	2020-09-02 22:35 (UTC+5:30)	1 hour, 16 minutes	48
□	▶		spark_demo	j-1LC521MH4SH17	Terminated User request	2020-08-28 01:07 (UTC+5:30)	28 minutes	12
□	▶	●	Smallsize	j-3JGPR214DZP8F	Terminated with errors Instance failure	2020-08-24 00:34 (UTC+5:30)	3 hours, 30 minutes	96
□	▶	●	spark_demo	j-1JRYUF7QPAJAJ	Terminated with errors Instance failure	2020-08-24 00:34 (UTC+5:30)	3 hours, 30 minutes	48
□	▶	●	spark_demo	j-3HPABODSSDQER	Terminated with errors Instance failure	2020-08-22 16:53 (UTC+5:30)	11 hours, 11 minutes	144
□	▶	●	spark_demo	j-3EYNM59XO6LAM	Terminated with errors Instance failure	2020-08-20 19:11 (UTC+5:30)	8 hours, 54 minutes	108
□	▶	●	smallsize_5	j-3O05NQ5NRBBRH	Terminated with errors Internal error	2020-08-19 01:09 (UTC+5:30)	4 minutes	0
□	▶		Smallsize	j-FK5TXVCWX8TY	Terminated User request	2020-08-19 00:56 (UTC+5:30)	36 minutes	24
□	▶	●	spark_m4xlarge_5	j-1ARE7LJ3HMXWK	Terminated with errors Internal error	2020-08-19 00:46 (UTC+5:30)	3 minutes	0

© 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.

From here, follow the steps below to create an EMR Cluster:

Step 1: First, you need to click on the blue ‘**Create cluster**’ button near the top left of the screen. Once you have clicked on this button, the following page should open on your screen:



Create Cluster - Quick Options Go to advanced options

General Configuration

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-864328032829-us-east-1/elasticmapre(

Launch mode: ● Cluster ○ Step execution

Software configuration

Release: emr-5.30.1

Applications:
- ● Core Hadoop: Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ○ HBase: HBase 1.4.13, Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- ○ Presto: Presto 0.232 with Hadoop 2.8.5 HDFS and Hive 2.3.6 Metastore
- ○ Spark: Spark 2.4.5 on Hadoop 2.8.5 YARN and Zeppelin 0.8.2

□ Use AWS Glue Data Catalog for table metadata

Hardware configuration

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

Number of instances: 3 (1 master and 2 core nodes)

Cluster scaling: □ scale cluster nodes based on workload

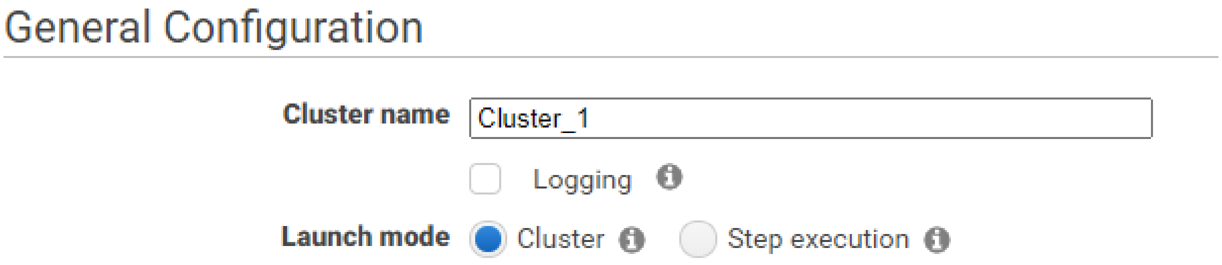
Security and access

EC2 key pair: Choose an option — Learn how to create an EC2 key pair.

Permissions: ● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

Step 2: Here, under ‘General Configuration’, write an appropriate Cluster name. Disable ‘Logging’ for this cluster, as this will lead to additional costs, and we do not need Logging services for the purposes of this module. Also, select ‘Cluster’ as the launch mode.



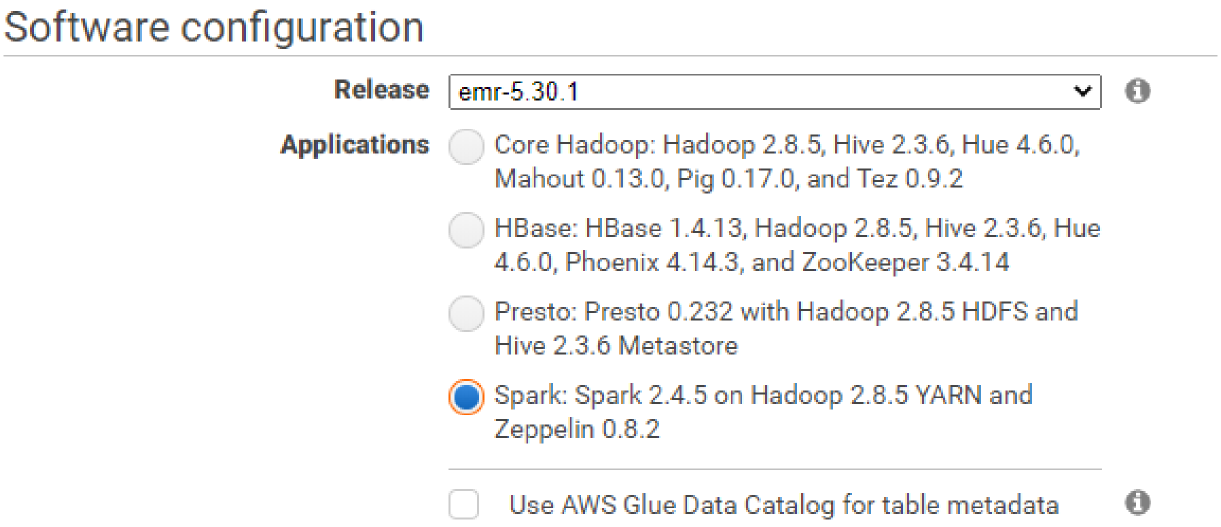
General Configuration

Cluster name: Cluster_1

□ Logging

Launch mode: ● Cluster ○ Step execution

Step 3: Under Software configuration', select "**emr-5.30.1**" as the Release. After this, you need to choose the Applications for your EMR cluster. Under 'Applications', since we will need Spark in our EMR cluster, choose the fourth option, which is for a Spark cluster.



Software configuration

Release: emr-5.30.1

Applications:

○ Core Hadoop: Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

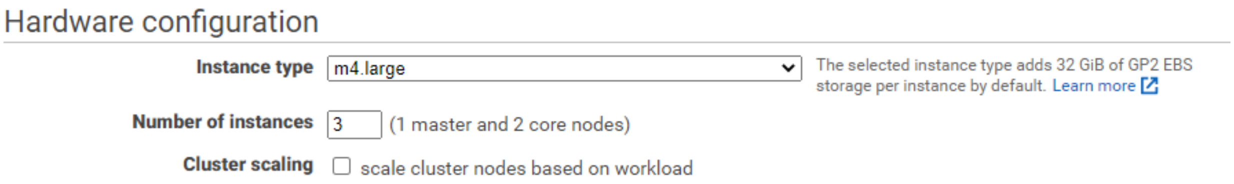
○ HBase: HBase 1.4.13, Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, Phoenix 4.14.3, and ZooKeeper 3.4.14

○ Presto: Presto 0.232 with Hadoop 2.8.5 HDFS and Hive 2.3.6 Metastore

● Spark: Spark 2.4.5 on Hadoop 2.8.5 YARN and Zeppelin 0.8.2

□ Use AWS Glue Data Catalog for table metadata

Step 4: Under 'Hardware configuration', we need to choose the instance type and the number of instances for our EMR Spark cluster. Here, please note that the type of instance and the number of instances that you choose will significantly affect the costs that you will incur when using the EMR cluster. For the purposes of this module, we will be using a **three-machine cluster** of instance type '**m4.large**'. Uncheck the checkbox for 'Cluster scaling'. Choose the configuration as shown below:



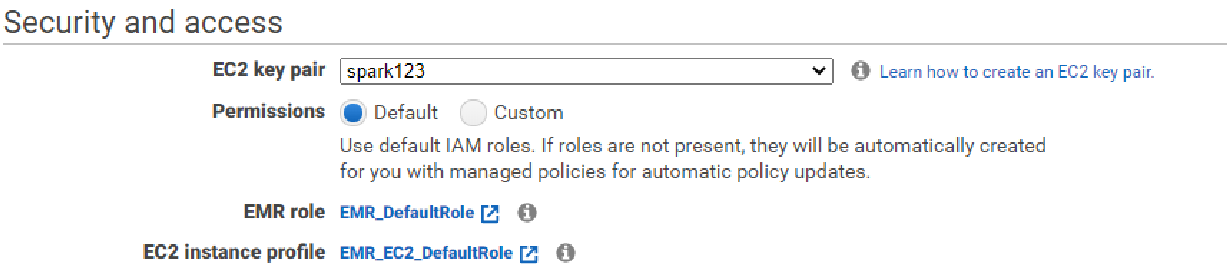
Hardware configuration

Instance type: m4.large — The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. Learn more

Number of instances: 3 (1 master and 2 core nodes)

Cluster scaling: □ scale cluster nodes based on workload

Step 5: Finally, under 'Security and access', you need to select the EC2 key pair that you have used until now in this program for practising on the EC2 machine. This will be used if you want to SSH to the Master node of your EMR cluster. For permissions, EMR role and EC2 instance profile, the default settings will be used.



Security and access

EC2 key pair: spark123

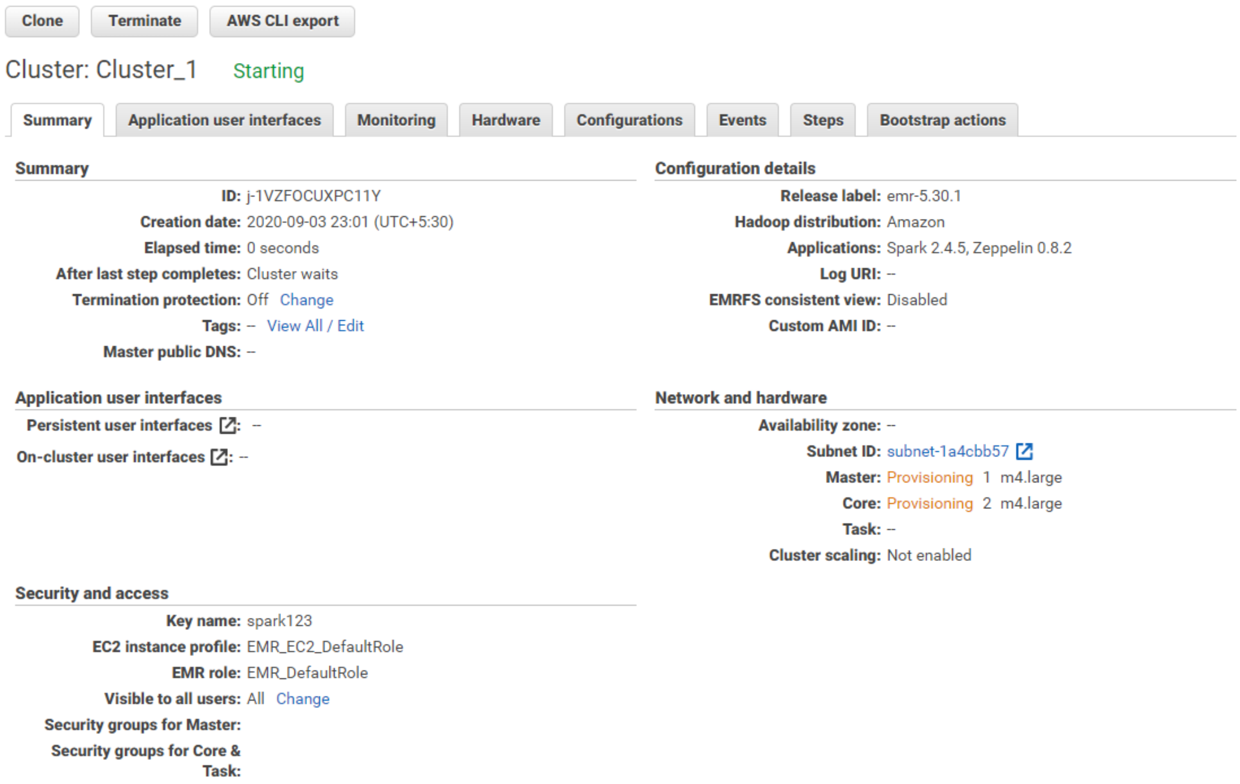
Permissions: ● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole

EC2 instance profile: EMR_EC2_DefaultRole

Step 6: You can now proceed ahead and click on the ‘Create cluster’ button and your cluster will be created. As soon as you do that, you will be taken to the screen below, which shows that your Spark EMR cluster is starting.



Cluster: Cluster_1 Starting

Summary

ID: j-1VZFOCUXPC11Y
Creation date: 2020-09-03 23:01 (UTC+5:30)
Elapsed time: 0 seconds
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: --

Configuration details

Release label: emr-5.30.1
Hadoop distribution: Amazon
Applications: Spark 2.4.5, Zeppelin 0.8.2
Log URI: --
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

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

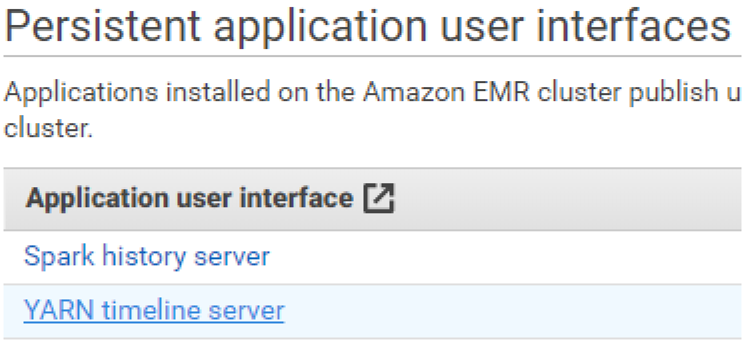
Network and hardware

Availability zone: --
Subnet ID: subnet-1a4cbb57
Master: Provisioning 1 m4.large
Core: Provisioning 2 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

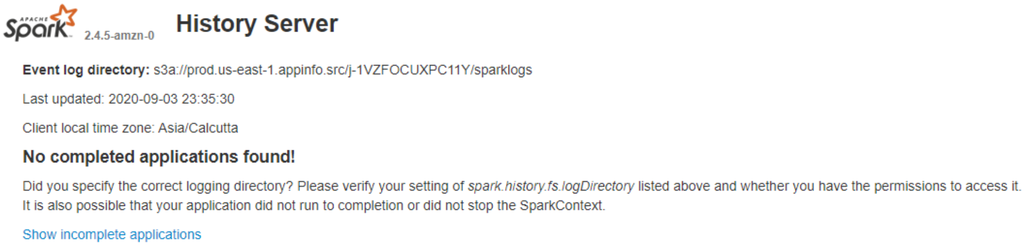
Key name: spark123
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master:
Security groups for Core & Task:

Note: If you need to access the Spark history server for your EMR cluster, you can easily get it by clicking on the ‘Application user interface’ tab and then clicking on the ‘Spark history server’ link. This will take you to the Spark history server. If you get any pop-up block error, ignore it and click on the link again.



Persistent application user interfaces

Applications installed on the Amazon EMR cluster publish u cluster.



History Server

Event log directory: s3a://prod.us-east-1.appinfo.src/j-1VZFOCUXPC11Y/sparklogs

Last updated: 2020-09-03 23:35:30

Client local time zone: Asia/Calcutta

No completed applications found!

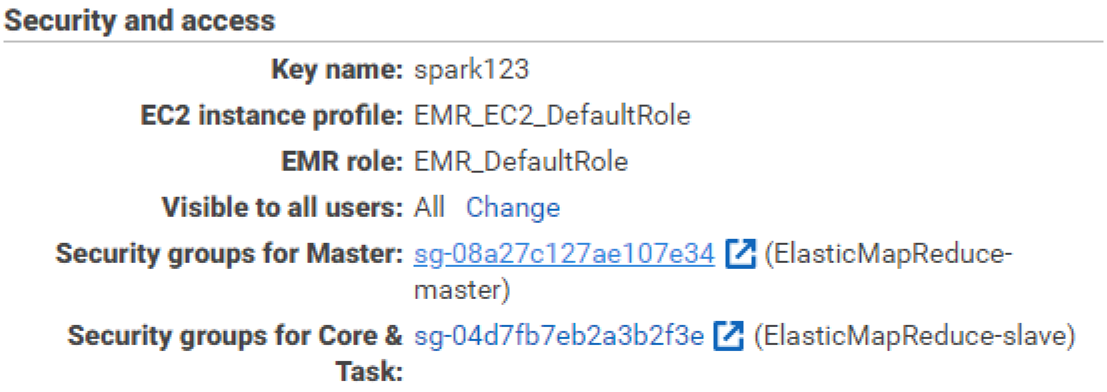
Did you specify the correct logging directory? Please verify your setting of *spark.history.fs.logDirectory* listed above and whether you have the permissions to access it.
It is also possible that your application did not run to completion or did not stop the SparkContext.

Show incomplete applications

Setting Up the Security Group for the Master Node

You need to configure the security group for the Master node if you want to SSH to it. You will be doing this in the third session wherein you will be running the spark-submit commands on the Master node of the EMR cluster. The steps for setting up the security group of the Master node are the same as that for normal EC2 instances.

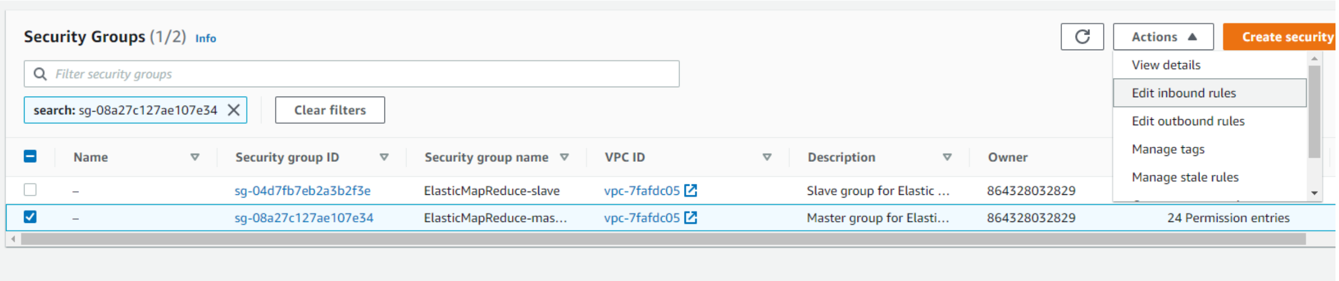
Step 1: First, click on the security group link that is displayed after the text 'Security groups for Master'. This will take you to the security group configuration screen.



Security and access

Key name: spark123
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-08a27c127ae107e34 (ElasticMapReduce-master)
Security groups for Core & Task: sg-04d7fb7eb2a3b2f3e (ElasticMapReduce-slave)

Step 2: Next, click the checkbox for the Master group for EMR and then click on Actions, followed by clicking on Edit inbound rules, as shown below:

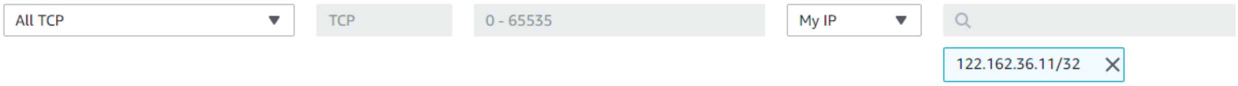


Security Groups (1/2) Info

search: sg-08a27c127ae107e34

☐	Name	Security group ID	Security group name	VPC ID	Description	Owner	
☐	–	sg-04d7fb7eb2a3b2f3e	ElasticMapReduce-slave	vpc-7fafdc05	Slave group for Elastic ...	864328032829	
☑	–	sg-08a27c127ae107e34	ElasticMapReduce-mas...	vpc-7fafdc05	Master group for Elasti...	864328032829	24 Permission entries

Step 3: Here, you need to add another Rule. Select the type of rule as ‘ALL TCP’ and source as ‘My IP’. After this, save the rules and this step complete.



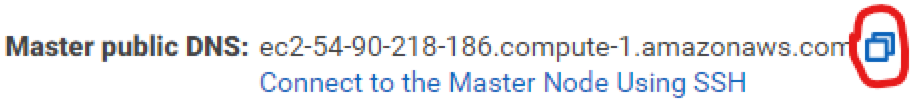
All TCP ▼ | TCP | 0 - 65535 | My IP ▼ | 122.162.36.11/32 ✕

Connecting to the Master Node of Your Spark EMR Cluster

Once your cluster is up and running, you can connect to the Spark EMR cluster using SSH. The steps for this are as follows:

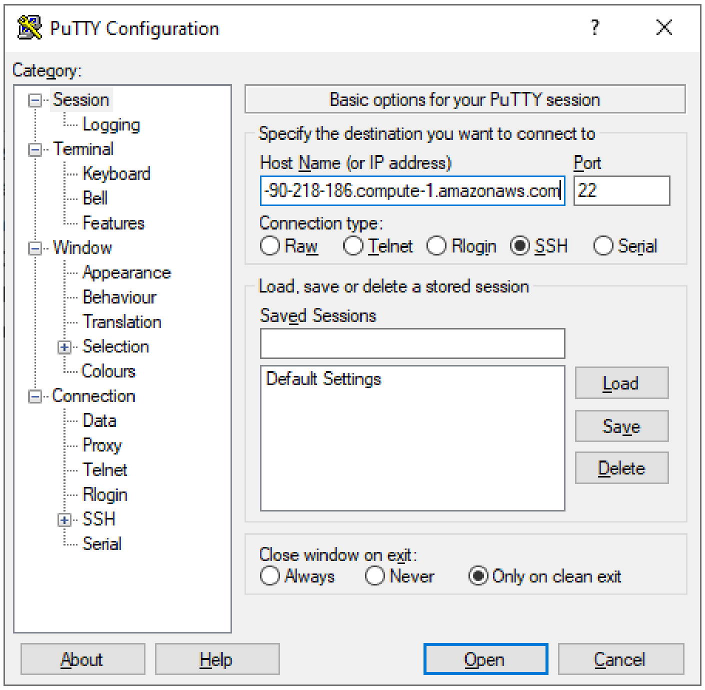
For Windows Users

Step 1: First, you need to copy the Master public DNS from the cluster Summary page. You can do this by clicking on the blue square icon adjacent to the DNS address to the right.



Master public DNS: ec2-54-90-218-186.compute-1.amazonaws.com

Step 2: After this, the steps to connect to the Master node are the same as the steps for connecting to an EC2 instance. First, open PuTTY and then paste the DNS address in the ‘Host Name’ field.



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): -90-218-186.compute-1.amazonaws.com

Port: 22

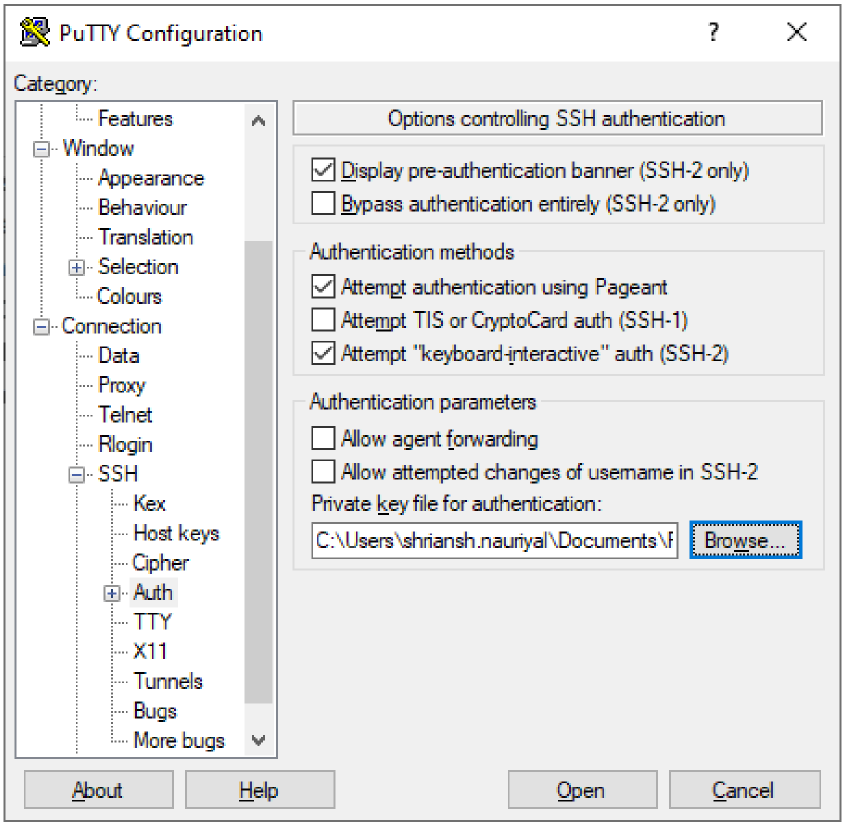
Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions: Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

Step 3: Next, proceed to opening the SSH category, and under that, click on ‘Auth’ and then browse for the ppk file that you have configured to be used for this cluster.



Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)
□ Bypass authentication entirely (SSH-2 only)

Authentication methods

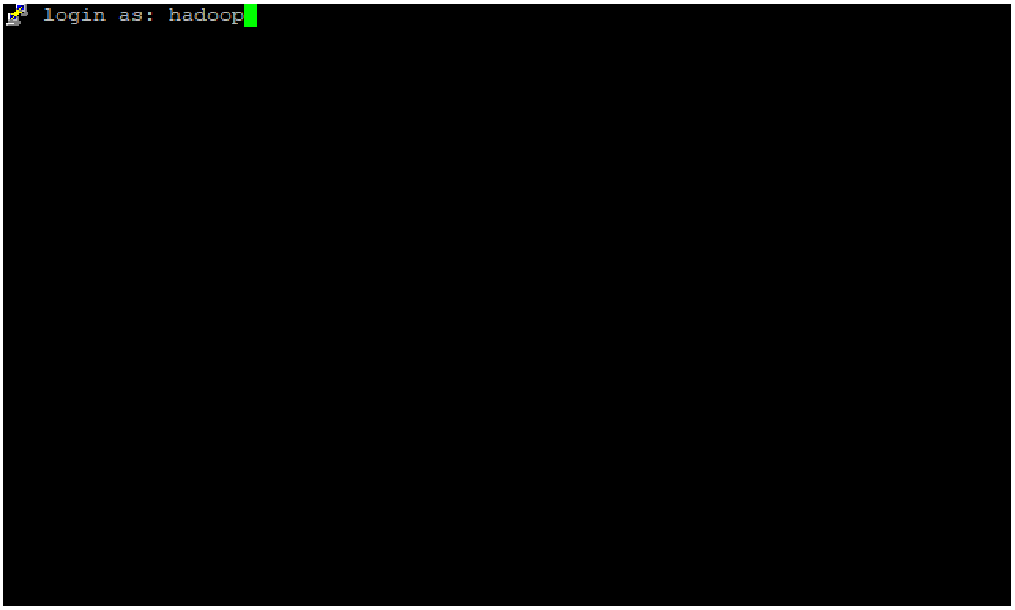
☑ Attempt authentication using Pageant
□ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

□ Allow agent forwarding
□ Allow attempted changes of username in SSH-2

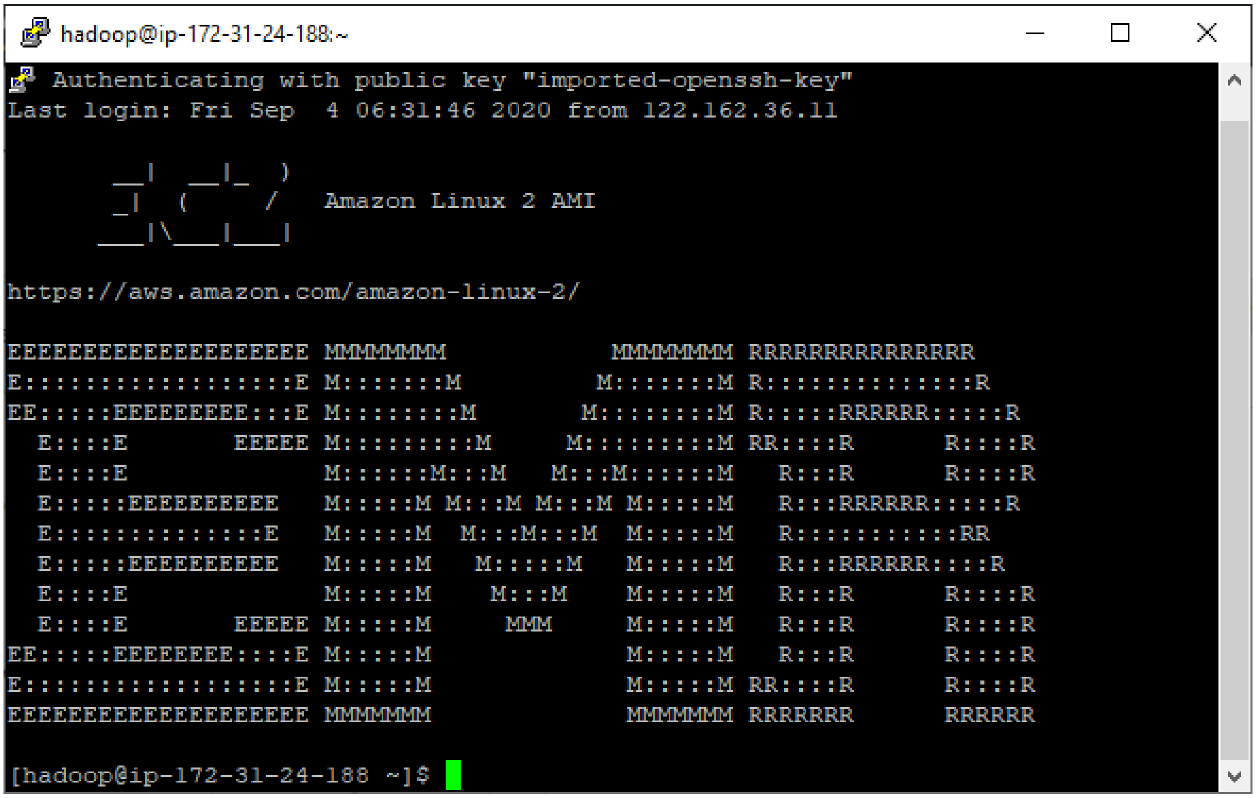
Private key file for authentication:
C:\Users\shriansh.nauriyal\Documents\I

Step 4: Next, click on ‘Open’. Click ‘Yes’ if any pop-ups appear and then in ‘login as’, you need to type ‘**hadoop**’, as you did for accessing an EC2 instance, and then click on Enter.



```
login as: hadoop
```

With this, you have gained access to the Master node of your EMR cluster.



```
Authenticating with public key "imported-openssh-key"
Last login: Fri Sep  4 06:31:46 2020 from 122.162.36.11

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

https://aws.amazon.com/amazon-linux-2/

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

[hadoop@ip-172-31-24-188 ~]$
```

For Mac/Linux Users

Step 1: First, you need to open the terminal on your machines.

Step 2: Next, you need to type the following command:

```
ssh -i <path to pem file> hadoop@<Master Public DNS>
```

For example, in my case, I can execute the following command to log in to the Master node:

```
ssh -i ~/spark123.pem hadoop@ec2-54-226-100-62.compute-1.amazonaws.com
```

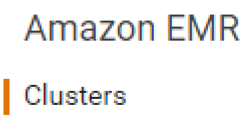
Step 3: Finally, type ‘Yes’ to dismiss any warnings that may be displayed:

Termination of an EMR Cluster

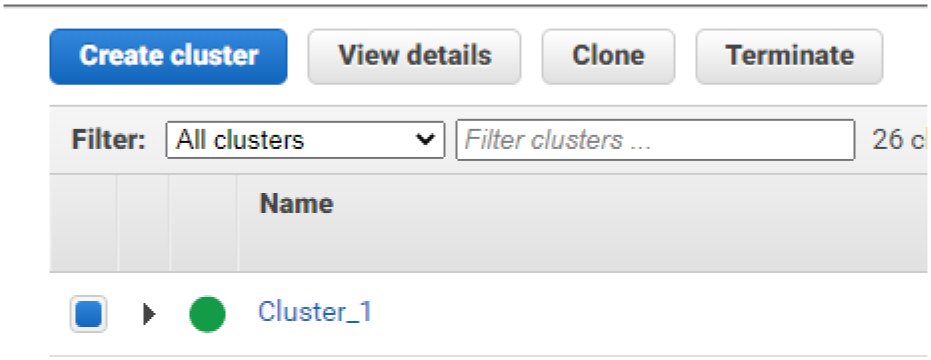
Once you have completed all your coding exercises on your EMR cluster, it is very important that you terminate the cluster. EMR is quite a costly service, and it must be kept activated only for the time that you are using it to run Spark jobs.

You can easily terminate your EMR cluster by following the steps below:

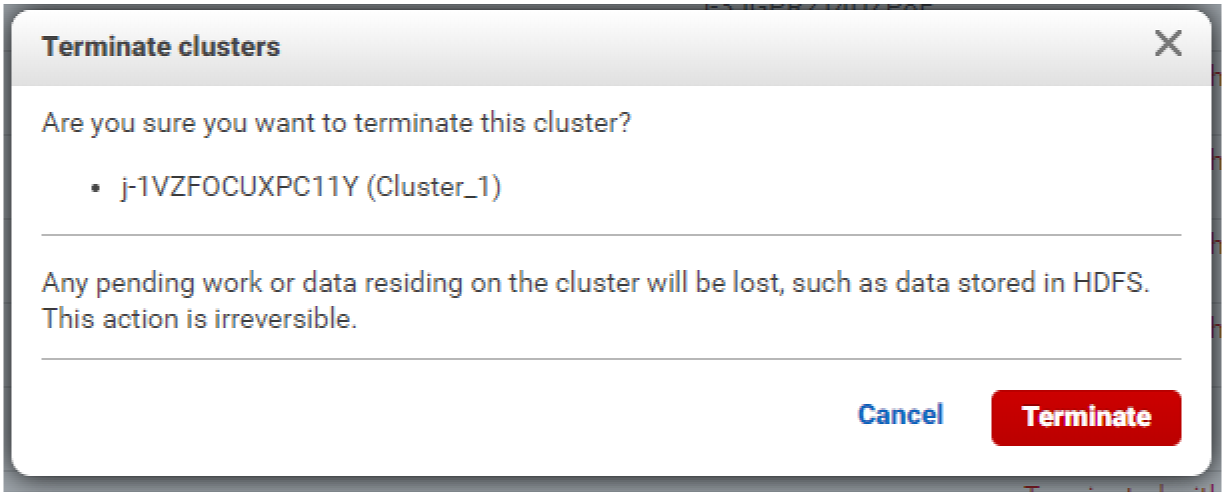
Step 1: You need to go to the cluster home page, which contains an entire list of all the EMR clusters in your account. You can do this by clicking on the ‘Clusters’ button to the top left under Amazon EMR.



Step 2: Next, you need to click the checkbox for your Spark EMR cluster and then click on the ‘Terminate’ button.



Step 3: A pop-up will appear, asking you to confirm whether you want to terminate the EMR cluster. Click on the 'Terminate' button in the pop-up. Once you click this button, your Spark EMR cluster will be terminated.



Terminate clusters

Are you sure you want to terminate this cluster?

- j-1VZFOCUXPC11Y (Cluster_1)

Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.