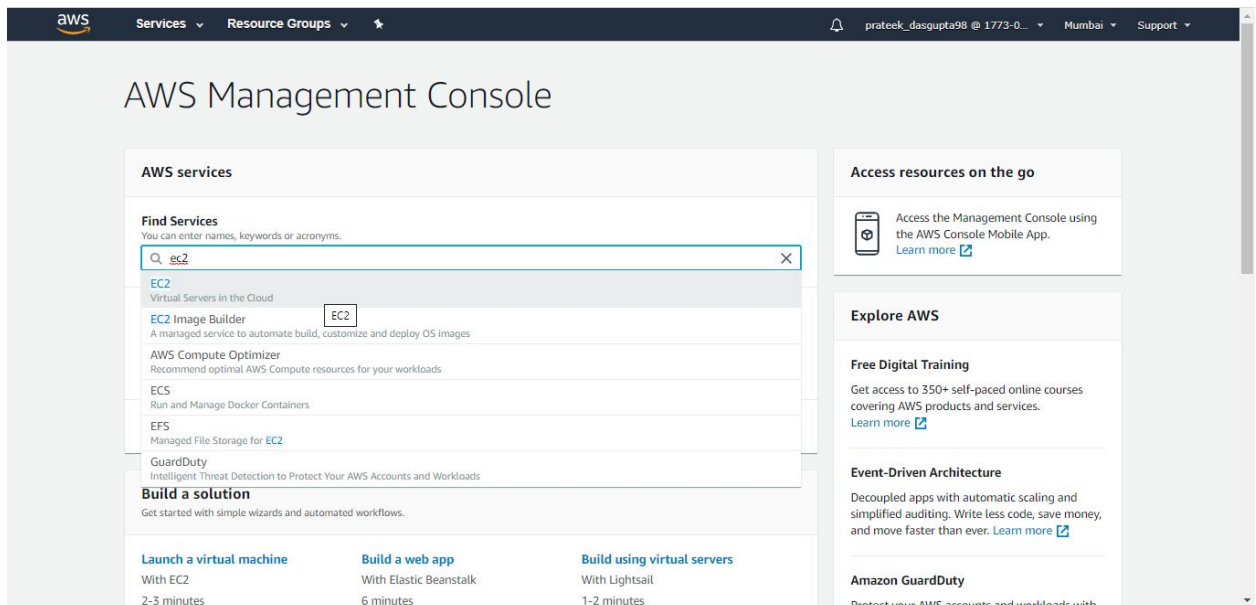


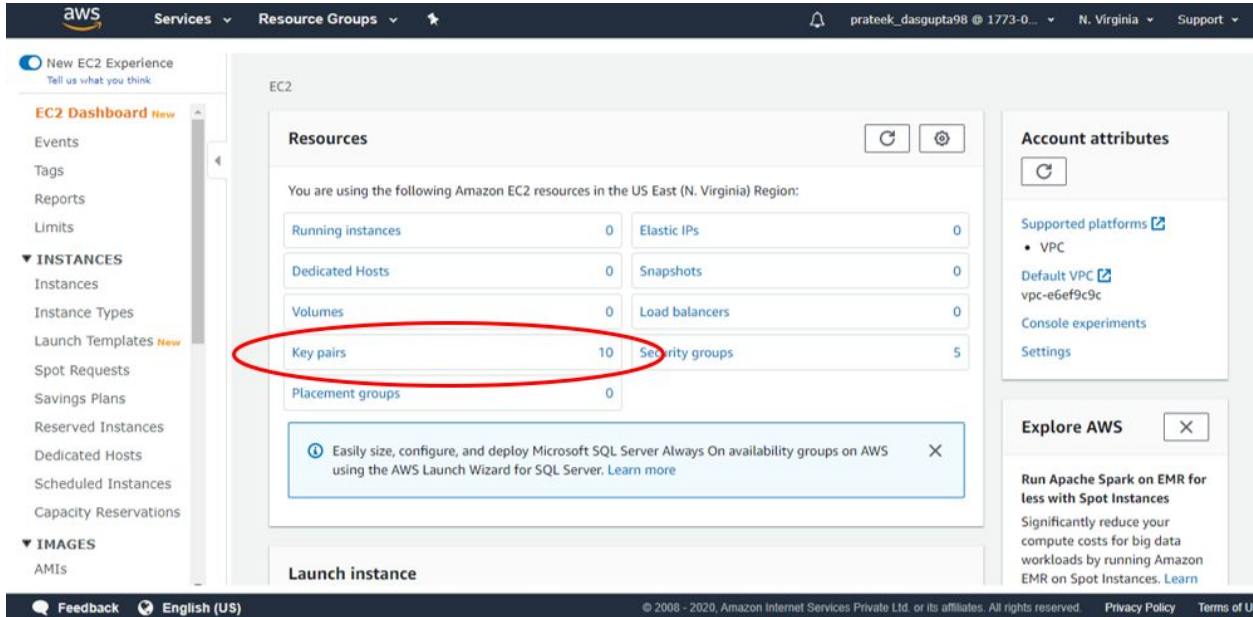
Steps to Launch an EMR Cluster

Note: Make sure that the location selected for running your account is **US East (N. Virginia)** us-east-1 else the EMR cluster won't launch

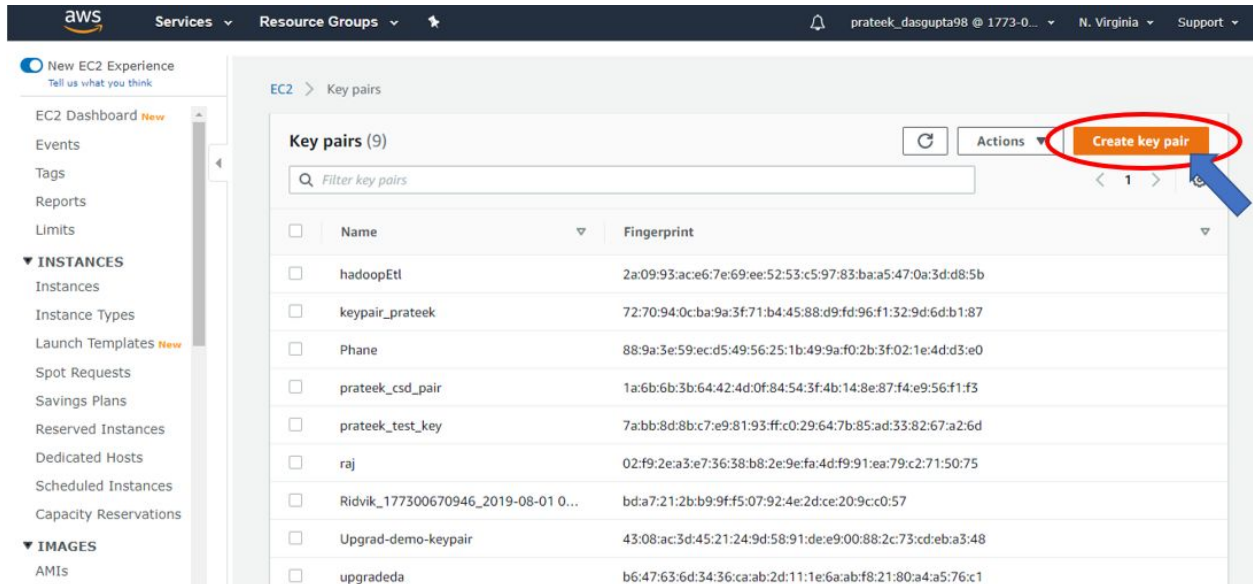
1. Log into your AWS account and open the Management Console
2. Before you create your EMR cluster, you will need to create a key-pair. This is because your EMR cluster will be running on EC2 instances and you will require a key pair to connect with your instance. Let us quickly revise how you can create your key-pair using the console.
 - a. Once you have logged in to your AWS account, search for 'EC2' under 'Find Services' and click on it.



b. Under 'Resources', click on 'Key Pairs'.



c. Now click on 'Create key pair'



d. Give a name to your key pair. In our case, we have named it as demo_key_pair and used the pem File format. Now click on 'Create key pair'.

Create key pair

Key pair

A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name

demo_key_pair

The name can be up to 255 characters long. Valid characters include `_`, `-`, `a-z`, `A-Z`, and `0-9`.

File format

- ☒ pem
For use with OpenSSH
- ☐ ppk
For use with PuTTY

Cancel

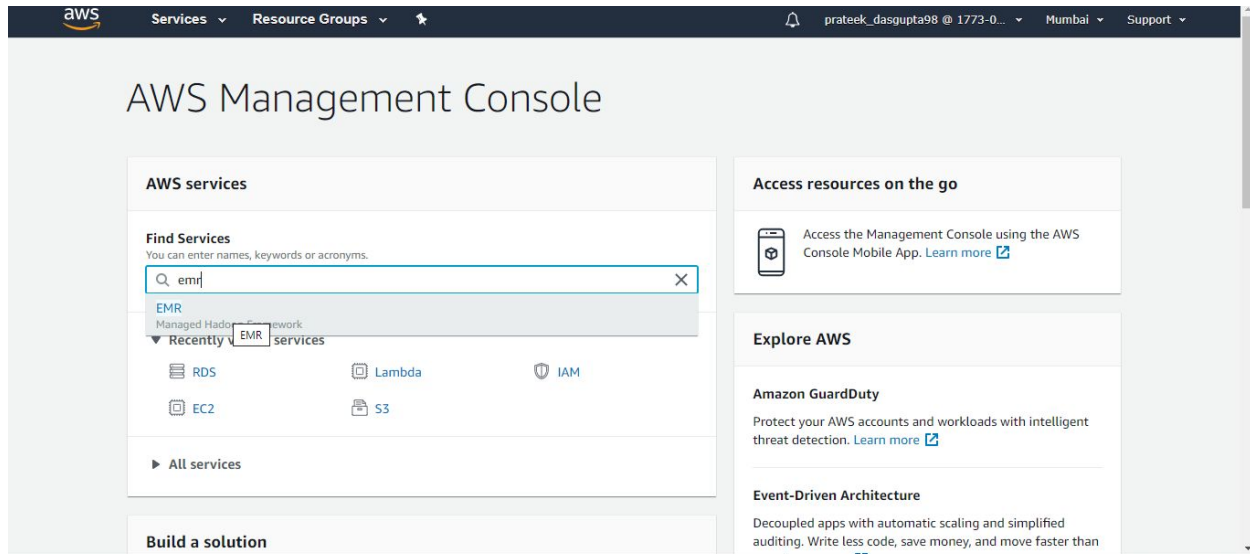
Create key pair

- e. Great! You now have your key pair and you can proceed with launching your EMR cluster.

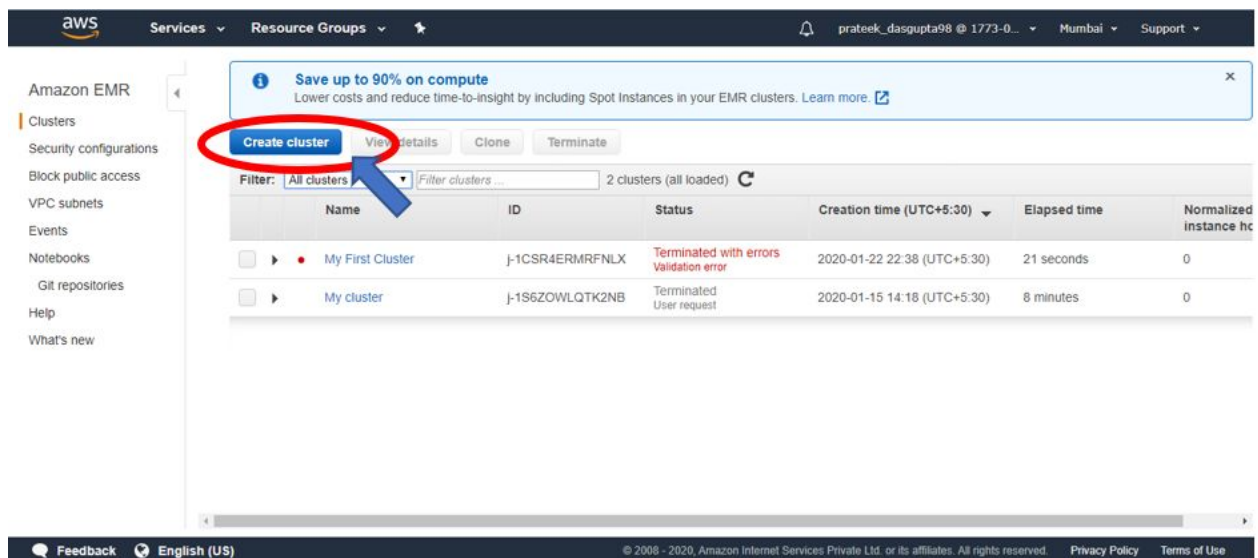
The screenshot shows the AWS Management Console interface. At the top, a green banner indicates 'Successfully created key pair'. Below this, the 'Key pairs' section is visible, showing a list of key pairs. The 'demo_key_pair' entry is circled in red. A blue arrow points from this entry to the 'demo_key_pair.pem' file in the bottom left corner of the console.

Name	Fingerprint
demo_key_pair	73:e9:f2:16:b8:55:00:43:37:04:2d:34:cb:94:e7:43:90:18:b0:e9
hadoop-ett	2a:09:93:ace6:7e:69:ee:52:53:c5:97:83:ba:a5:47:0a:3d:d8:5b
keypair-prateek	72:70:94:0c:ba:9a:3f:71:b4:45:88:d9:fd:96:f1:32:9d:6d:b1:87
Phane	88:9a:3e:59:ecd5:49:56:25:1b:49:9a:f0:2b:3f:02:1e:4d:d3:e0
prateek_csd_pair	1a:6b:6b:3b:64:42:0f:84:54:3f:4b:14:8e:87:f4:e9:56:f1:f3
prateek_test_key	7a:bb:8d:8b:c7:e9:81:93:ff:c0:29:64:7b:85:ad:33:82:67:a2:6d

3. Go to your home page and under 'Find Services', search 'EMR' and click on it



4. Once you reach the landing page of your accounts EMR cluster, you can check for any running cluster by choosing the **'Active clusters'** option next to the **'Filter'** button. Since you are interested to create a cluster, find the **'Create cluster'** button and click on it.



5. Now click on 'Go to advanced options'

aws Services Resource Groups

Create Cluster - Quick Option [Go to advanced options](#)

General Configuration

Cluster name

☒ Logging [?](#)

S3 folder

Launch mode ☒ Cluster [?](#) ☐ Step execution [?](#)

Software configuration

Release [?](#)

Applications

- ☒ Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ☐ HBase: HBase 1.4.10 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.6, Hue 4.4.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- ☐ Presto: Presto 0.227 with Hadoop 2.8.5 HDFS and Hive 2.3.6 Metastore
- ☐ Spark: Spark 2.4.4 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.2
- ☐ Use AWS Glue Data Catalog for table metadata

Feedback English (US) © 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use

6. You will find all the services under '**emr-5.29.0**'. It is advised to learn the use cases of each of these services in the Big Data Industry. You can select whichever service you will be using for your study. For the purpose of this demo, you can keep the existing services and additionally choose Spark as shown below.

aws Services Resource Groups

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release [?](#)

<input checked="" type="checkbox"/> Hadoop 2.8.5	<input type="checkbox"/> Zeppelin 0.8.2	<input type="checkbox"/> Livy 0.6.0
<input type="checkbox"/> JupyterHub 1.0.0	<input type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.9.1
<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 1.4.10	<input checked="" type="checkbox"/> Pig 0.17.0
<input checked="" type="checkbox"/> Hive 2.3.6	<input type="checkbox"/> Presto 0.227	<input type="checkbox"/> ZooKeeper 3.4.14
<input type="checkbox"/> MXNet 1.5.1	<input type="checkbox"/> Sqoop 1.4.7	<input type="checkbox"/> Mahout 0.13.0
<input checked="" type="checkbox"/> Hue 4.4.0	<input type="checkbox"/> Phoenix 4.14.3	<input type="checkbox"/> Oozie 5.1.0
<input checked="" type="checkbox"/> Spark 2.4.4	<input type="checkbox"/> HCatalog 2.3.6	<input type="checkbox"/> TensorFlow 1.14.0

Multiple master nodes (optional)

☐ Use multiple master nodes to improve cluster availability. [Learn more](#)

AWS Glue Data Catalog settings (optional)

☐ Use for Hive table metadata [?](#)

☐ Use for Spark table metadata [?](#)

Edit software settings [?](#)

☒ Enter configuration ☐ Load JSON from S3

Feedback English (US) © 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Scroll to the bottom of the page and click on 'Next'.

7. You would now have reached the hardware configuration page where you can **define your clusters and nodes** which will look something like this.

Root device EBS volume size GiB ⓘ

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. [Learn more about instance purchasing options](#)

Node type	Instance type	Instance count	Purchasing option
Master Master - 1 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ⓘ Add configuration settings ⓘ	1 Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▼
Core Core - 2 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ⓘ Add configuration settings ⓘ	<input type="text" value="2"/> Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▼
Task ✕ Task - 3 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ⓘ Add configuration settings ⓘ	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▼

[+ Add task instance group](#)

Feedback English (US) © 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use

8. The next step is to choose the Instance types for your cluster and nodes. Click on the pencil-shaped under the '**Instance Type**' column.

Root device EBS volume size GiB ⓘ

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. [Learn more about instance purchasing options](#)

Node type	Instance type	Instance count	Purchasing option
Master Master - 1 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ⓘ Add configuration settings ⓘ	1 Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▼
Core Core - 2 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ⓘ Add configuration settings ⓘ	<input type="text" value="2"/> Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▼
Task ✕ Task - 3 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ⓘ Add configuration settings ⓘ	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▼

[+ Add task instance group](#)

You will find a number of instance types with their corresponding RAM and Memory. Click on this [link](#) to learn about available Instance Types.

9. For the purpose of this demo, select the options shown in the screenshot below (**m4.large** for the Master and Core node) and click **Next**.

Use-m4.large Master- one instance and Core also as m4.large with 2 instances.

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master Master - 1	m4.large 4 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not available for Master
Core Core - 2	m4.large 4 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	2 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not enabled
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	0 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not enabled

10. It is advisable to rename your cluster-name to a unique name so that later on it becomes easy to search for the cluster. For this demo, we have named it 'Demo-Cluster'. You should be having a page as shown below. Click on **Next**.

General Options

Cluster name

☒ Logging
S3 folder

☒ Debugging

☒ Termination protection

Tags

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

Additional Options

☐ EMRFS consistent view

Custom AMI ID

Bootstrap Actions

Cancel

Previous

Next

11. Select the key pair that you have created and click **Create cluster**. In our case, we have named it as 'demo_key_pair'

The screenshot shows the 'Create Cluster - Advanced Options' page in the AWS Management Console. The 'Security Options' section is highlighted, showing the 'EC2 key pair' dropdown menu with 'demo_key_pair' selected. Below this, the 'Permissions' section shows the 'Default' radio button selected. The 'Create cluster' button at the bottom right is circled in red, and a blue arrow points to it.

12. AWS has started launching your cluster. It will take about 10-15 minutes to launch this cluster. You will be presented with a page like this as shown below.

The screenshot shows the 'Cluster: Demo-Cluster' page in the Amazon EMR console. The cluster is in the 'Starting' state. The page displays various details including connections, master public DNS, history service, tags, summary, and configuration details.

Summary	Configuration details
ID: j-K00ZIH4FQWJY	Release label: emr-5.29.0
Creation date: 2020-02-19 11:21 (UTC+5:30)	Hadoop distribution: Amazon 2.8.5
Elapsed time: 1 minute	Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0, Spark 2.4.4
After last step completes: Cluster waits	Log URI: s3://aws-logs-177300670946-us-east-1/elasticmapreduce/
Termination protection: On	EMRFS consistent view: Disabled
	Custom AMI ID: --

Steps to follow before performing SSH to the master node

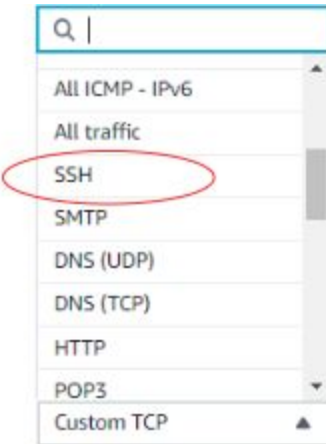
1. Under the cluster information page click on the **security groups of the master node**

Network and hardware	Security and access
Availability zone: us-east-1d	Key name: phanendra_sanskar
Subnet ID: subnet-38133064	EC2 instance profile: EMR_EC2_DefaultRole
Master: Running 1 m4.large	EMR role: EMR_DefaultRole
Core: Running 2 m4.large	Auto Scaling role: EMR_AutoScaling_DefaultRole
Task: --	Visible to all users: All Change
	Security groups for sg-024a40edec2182c0d Master: (ElasticMapReduce-master)
	Security groups for sg-08b9414f92bc12611 Core & Task: (ElasticMapReduce-slave)

- Clicking on the security group and you will land on a similar page. Here click on the security group of the **Elastic Mapreduce-master node** as highlighted in the image.

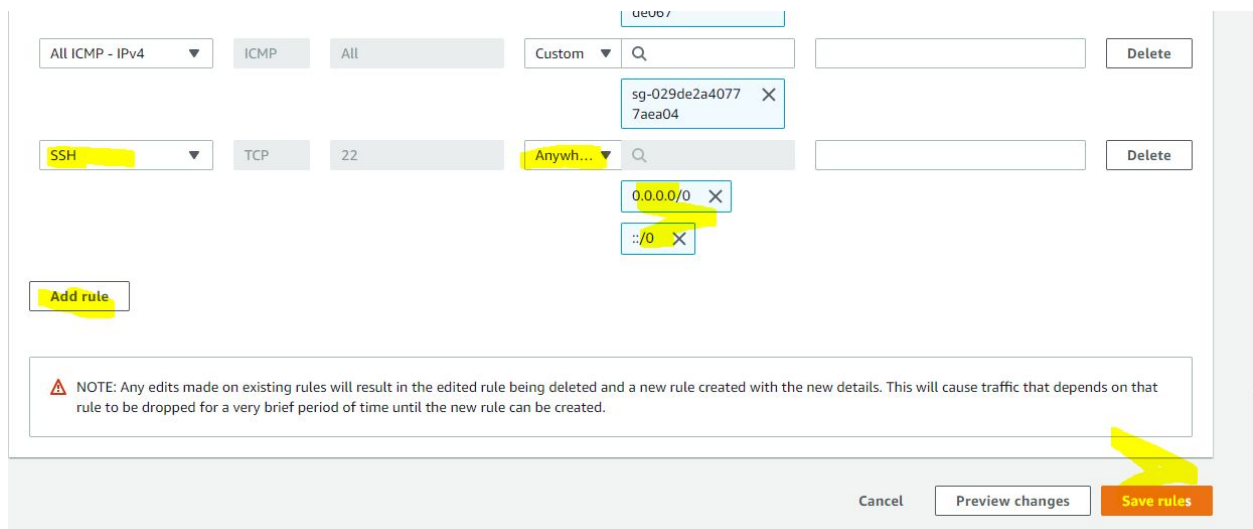
Security Groups (2) Info				
<div> <input type="text" value="Filter security groups"/> < 1 > </div> <div> <input type="text" value="search: sg-024a40edec2182c0d"/> <input type="button" value="Clear filters"/> </div>				
<input type="checkbox"/>	Security group ID ▲	Security group name ▼	VPC ID ▼	Description ▼
<input type="checkbox"/>	sg-024a40edec2182c0d	ElasticMapReduce-mas...	vpc-c03143ba	Master group for Elasti...
<input type="checkbox"/>	sg-08b9414f92bc12611	ElasticMapReduce-slave	vpc-c03143ba	Slave group for Elastic ...

- Clicking on the security group will land you on the corresponding security information page. Click on **edit inbound rules** to add a new rule



The “**Type**” field will be **SSH**”, and the **Source** will be “**Anywhere**” for this rule. For frequent testing, you can avoid using My IP address and choose “Anywhere” while adding rules in the Security Group.

After adding the rule do not forget to click **save rules** at the bottom of the window

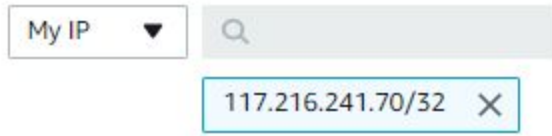


Then access the master node using putty[Windows Users] or the terminal [MAC/Linux Users].

Note: Avoid choosing the My IP for EMR cluster.

What happens if you choose My IP?

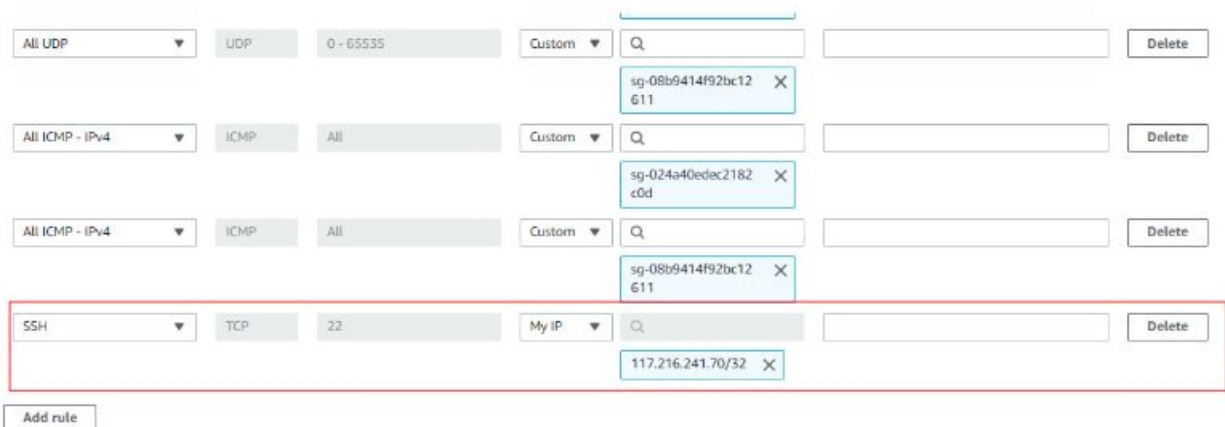
If you Choose **My IP** under the source field. This will automatically load your IP address in the adjacent blank column.



My IP ▼

117.216.241.70/32 X

On addition of the rule and choosing the appropriate options as shown below, click on **save rule** [at the bottom of the screen] to successfully add the rule



All UDP	UDP	0 - 65535	Custom	Q		Delete
				sg-08b9414f92bc12611		
All ICMP - IPv4	ICMP	All	Custom	Q		Delete
				sg-024a40edec2182c0d		
All ICMP - IPv4	ICMP	All	Custom	Q		Delete
				sg-08b9414f92bc12611		
SSH	TCP	22	My IP	Q		Delete
				117.216.241.70/32		

Add rule

Now under the list of inbound rules appearing under the master node security group you can see the newly added rule.

Inbound rules				Edit inbound rules
Type	Protocol	Port range	Source	Description - optional
All TCP	TCP	0 - 65535	sg-024a40edec2182c0d (ElasticMapReduce-master)	-
All TCP	TCP	0 - 65535	sg-08b9414f92bc12611 (ElasticMapReduce-slave)	-
SSH	TCP	22	117.216.241.70/32	-
Custom TCP	TCP	8443	207.171.167.25/32	-
Custom TCP	TCP	8443	54.240.217.8/29	-
Custom TCP	TCP	8443	72.21.196.64/29	-
Custom TCP	TCP	8443	72.21.198.64/29	-
Custom TCP	TCP	8443	54.240.217.16/29	-

On adding this rule it enables you to perform an SSH to the master node of the cluster.

But every time you are cloning the cluster or connecting to the **cluster after restarting the laptop**, the first thing to do is edit the security groups and update your current IP address.

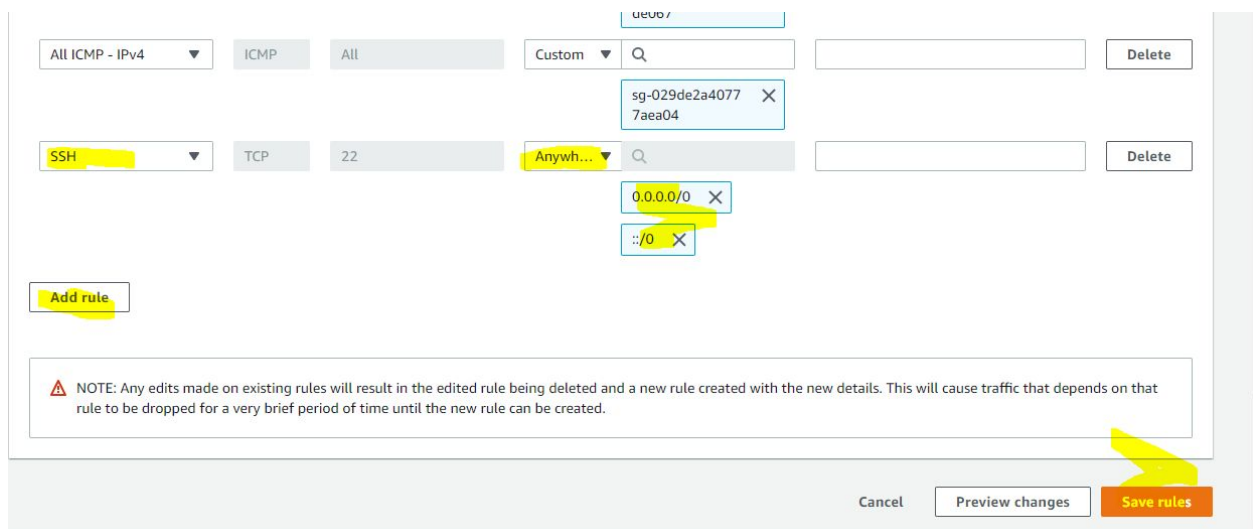
The screenshot shows the AWS Security Groups console with three inbound rules. The third rule, 'SSH', is highlighted with a red box. This rule is configured with 'Type' as SSH, 'Protocol' as TCP, and 'Port range' as 22. The 'Source' field is set to 'My IP'. The 'Delete' button is visible to the right of the rule. Above the highlighted rule, two other rules are visible: 'All TCP' (Port range 0 - 65535) and 'Custom TCP' (Port range 0 - 65535). Both of these rules have their 'Source' field set to a specific security group ID (sg-024a40edec2182c0d and sg-08b9414f92bc12611 respectively).

As shown in the figure you need to edit the rule corresponding to **SSH type** rule and change the option from custom to **My IP** and click on **save rules**.

This will ensure you to do a successful SSH or successfully cloning to a new cluster.

Important - General Practice

To avoid this hassle every time you clone a cluster or every time you restart the laptop a common practice followed while studying/ or internal testing is choosing the option anywhere instead of custom or My IP. In the actual development environment, this should be avoided because it leaves the cluster vulnerable and any IP address can access the cluster.



The screenshot shows the AWS IAM console rule configuration interface. It features two rule configurations. The first rule is for 'All ICMP - IPv4' with protocol 'ICMP' and action 'All'. The second rule is for 'SSH' with protocol 'TCP' and port '22'. The source for the SSH rule is set to 'Anywhere...'. Below the rules, there is a note: 'NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.' At the bottom right, there are buttons for 'Cancel', 'Preview changes', and 'Save rules'. A yellow arrow points to the 'Save rules' button.

0e0b7

All ICMP - IPv4 ICMP All Custom Q Delete

sg-029de2a40777aea04 X

SSH TCP 22 Anywh... Q Delete

0.0.0.0/0 X

:::0 X

Add rule

⚠ NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Cancel Preview changes Save rules