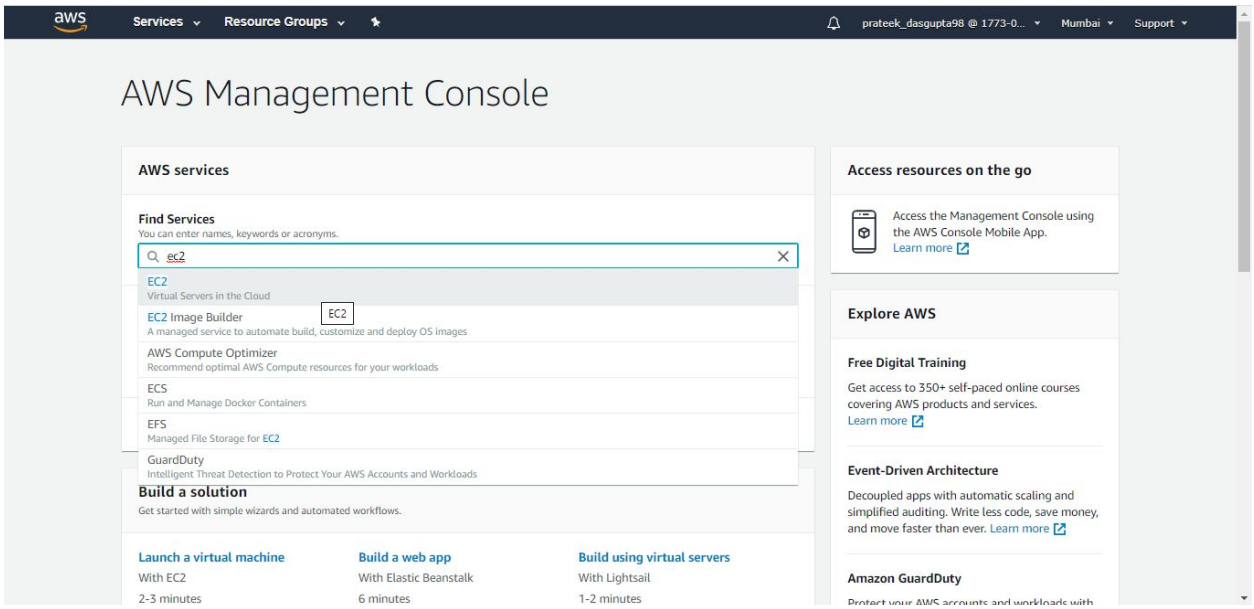
Steps to Launch an EMR Cluster

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

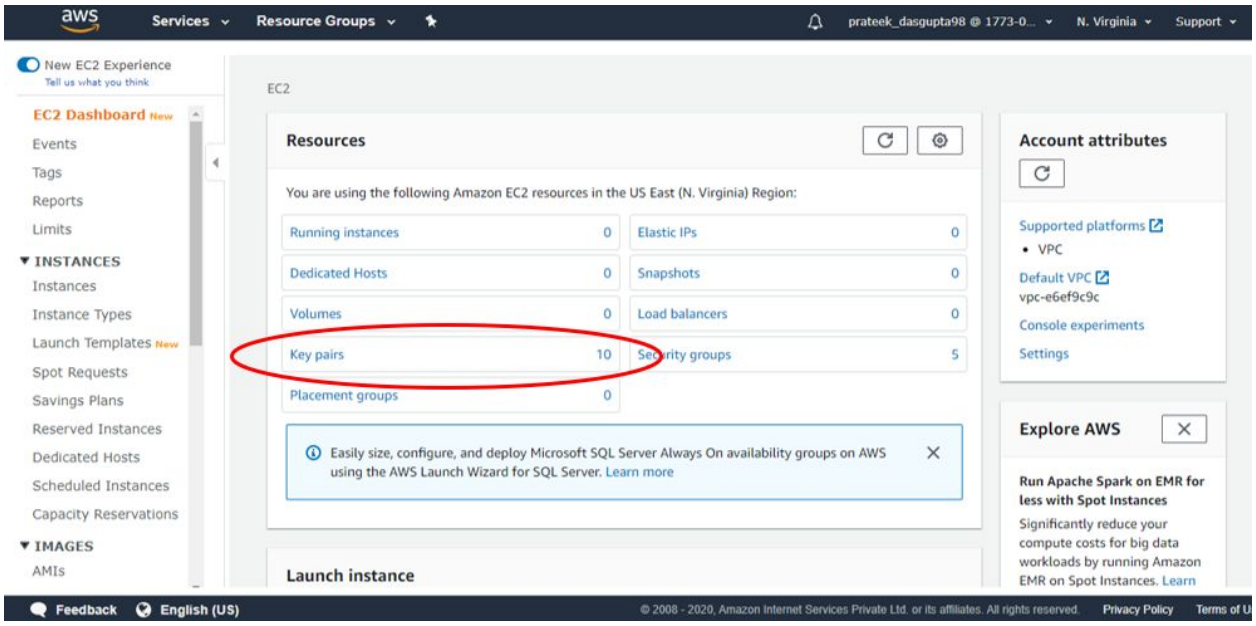
1. Log into your AWS account and open the Management Console

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.

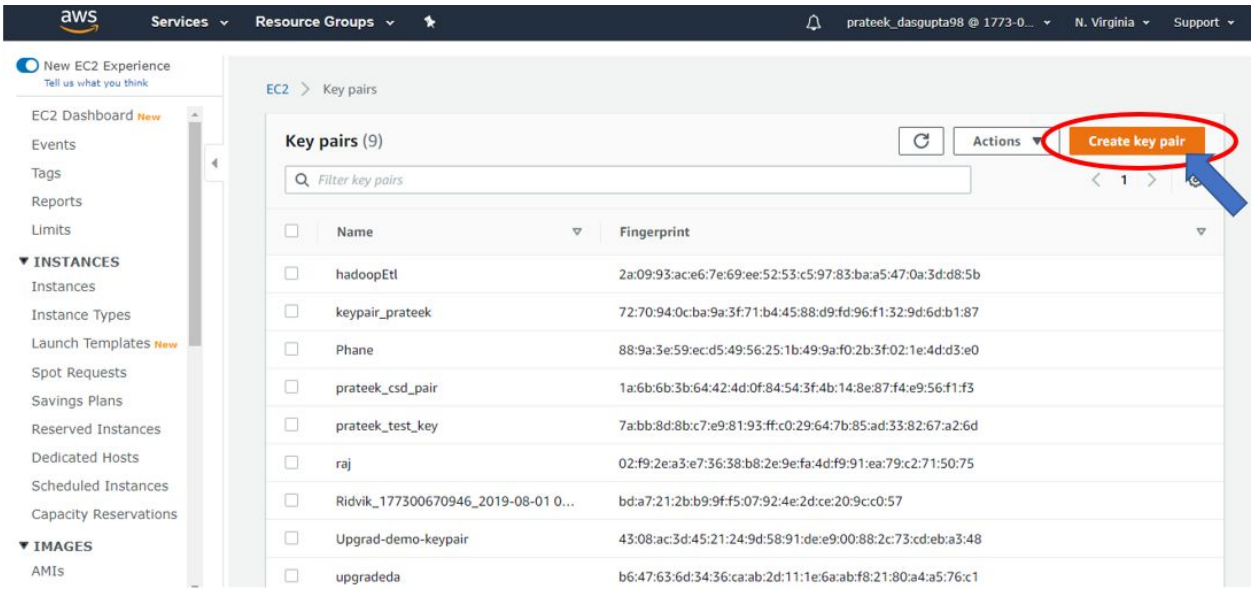
    a. Once you have logged in to your AWS account, search for '**EC2**' under '**Find Services**' and click on it.



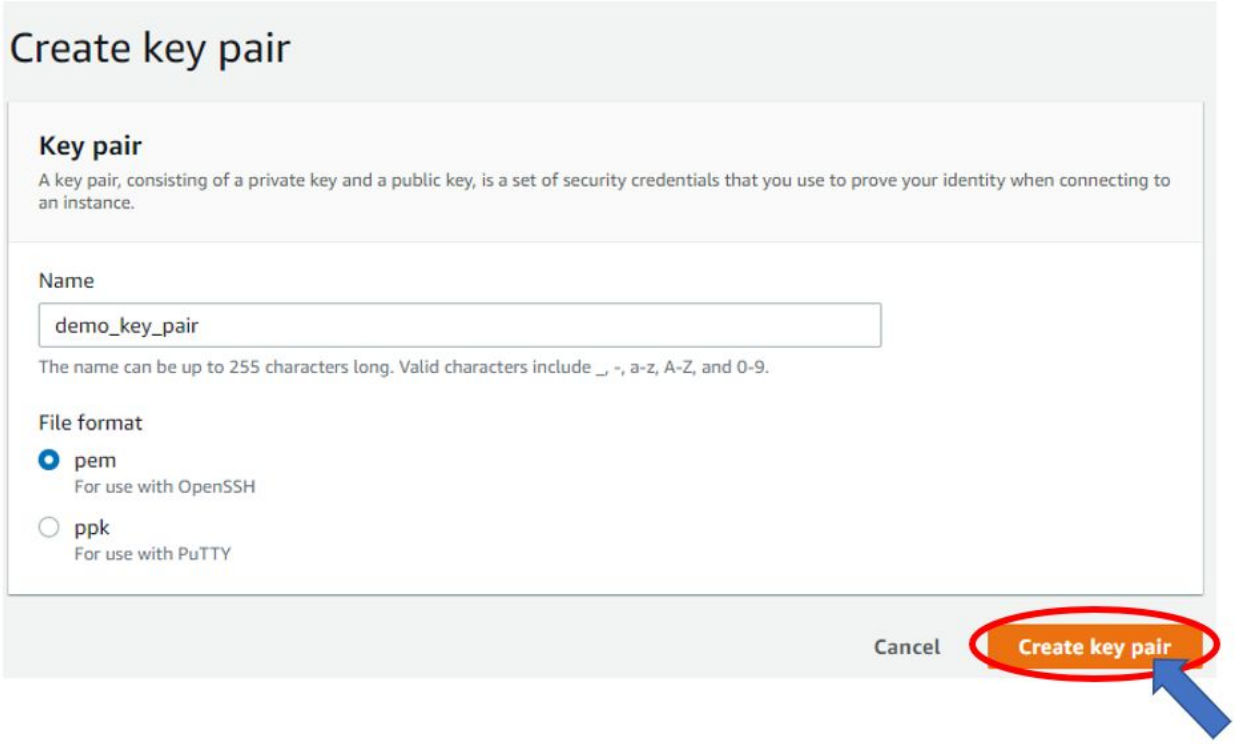
b. Under '**Resources**', click on '**Key Pairs**'.



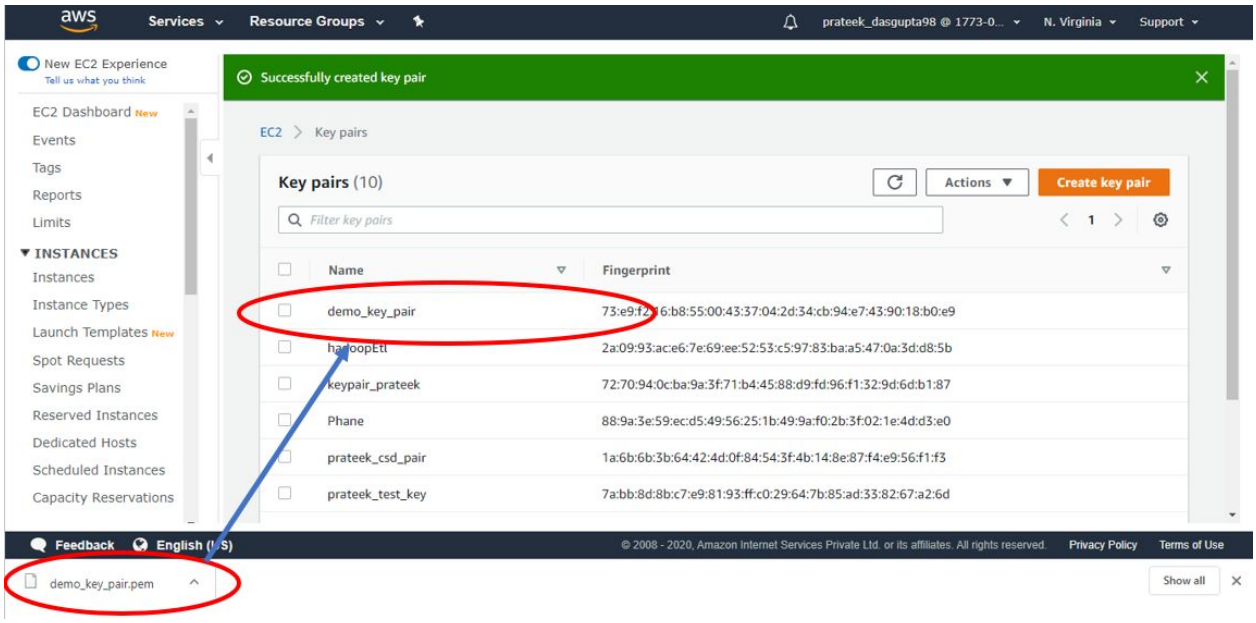
c. Now click on '**Create key pair**'



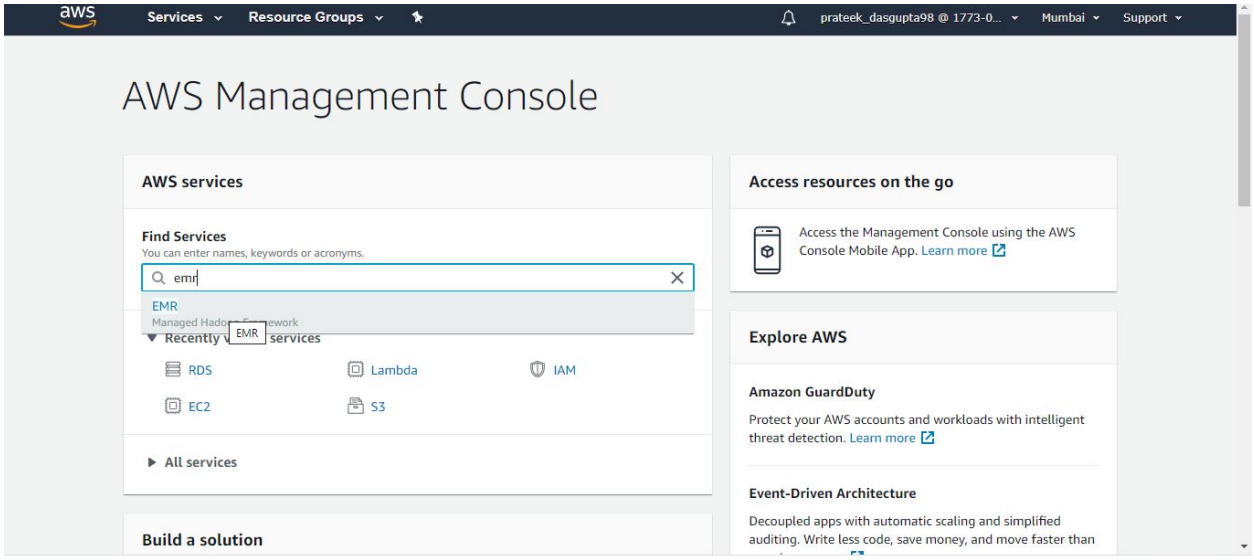
d. Give a name to your key pair. In our case, we have named it as demo_key_pair and used the pem File format. Now click on '**Create key pair**'.



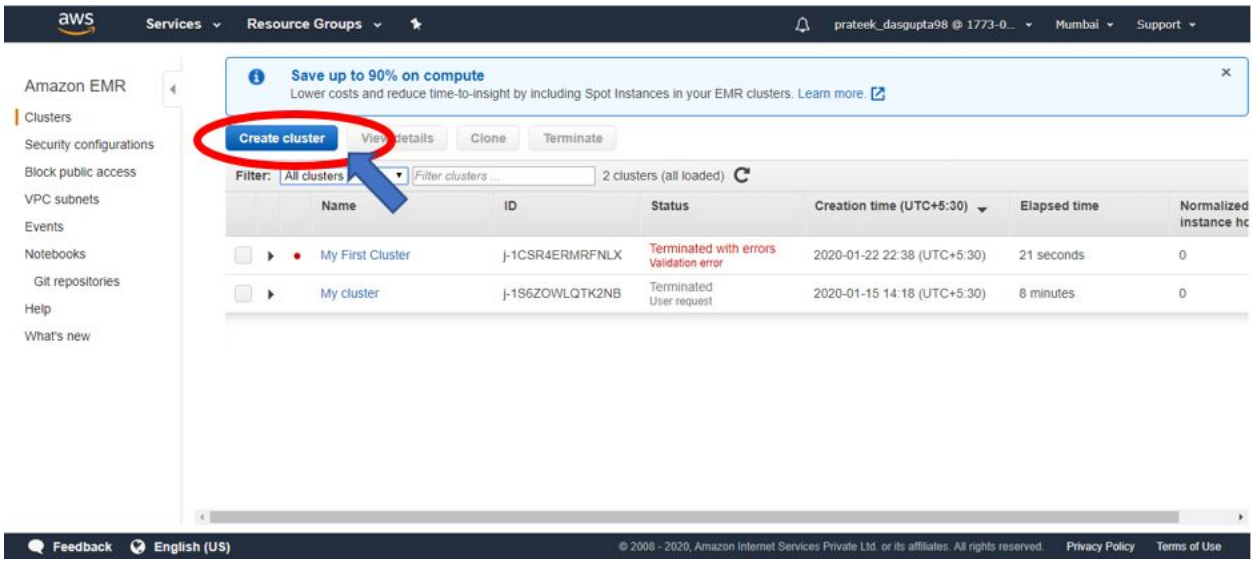
e. Great! You now have your key pair and you can proceed with launching your EMR cluster.



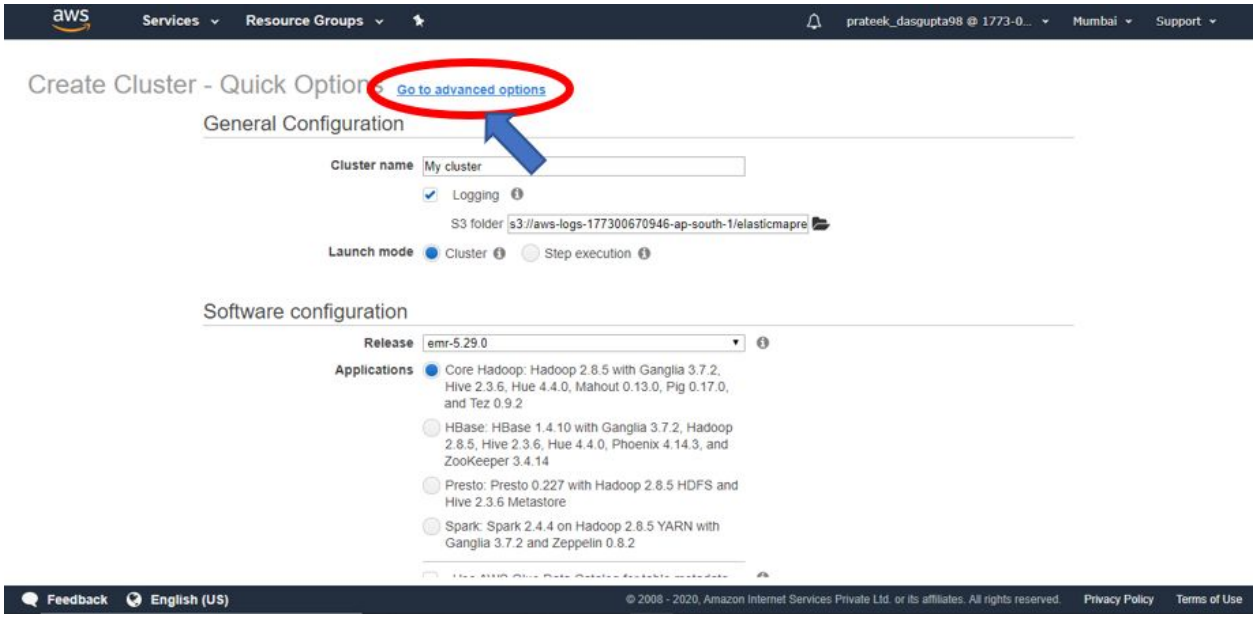
3. Go to your home page and under '**Find Services**', search **'EMR'** and click on it



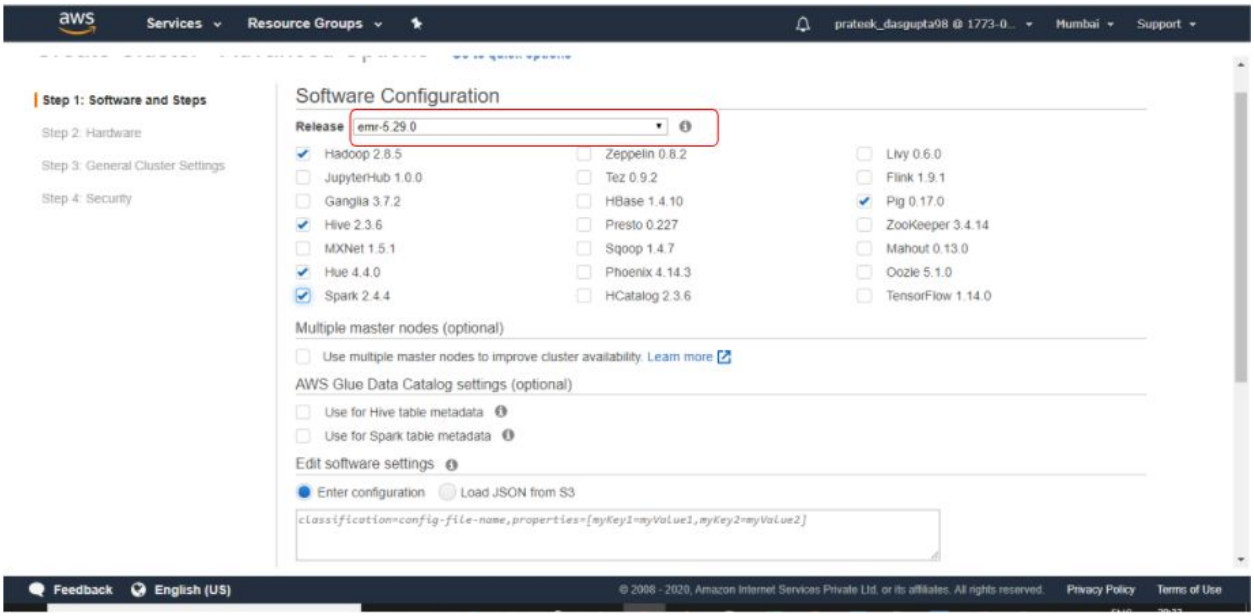
4. Once you reach the landing page of your accounts EMR cluster, you can check for any running cluster by choosing the '**Active clusters**' option next to the **'Filter'** button. Since you are interested to create a cluster, find the **'Create cluster**' button and click on it.



5. Now click on **'Go to advanced options'**

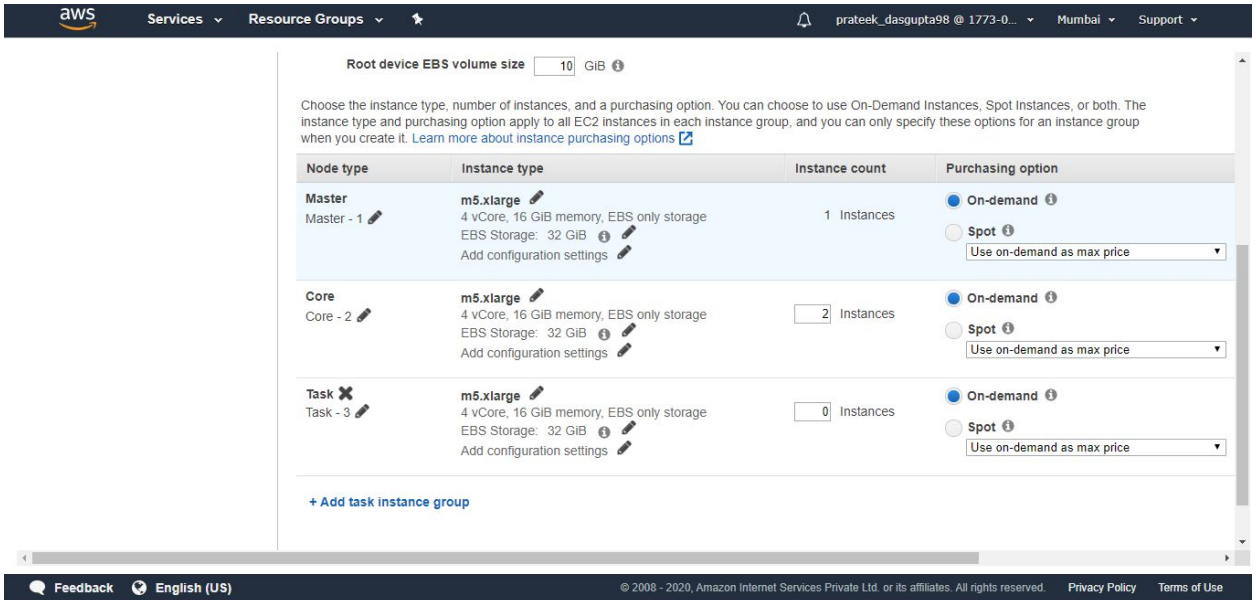


6. You will find all the services under '**emr-5.29.0**'. It is advised to learn the use cases of each of these services in the Big Data Industry. You can select whichever service you will be using for your study. For the purpose of this demo, you can keep the existing services and additionally choose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

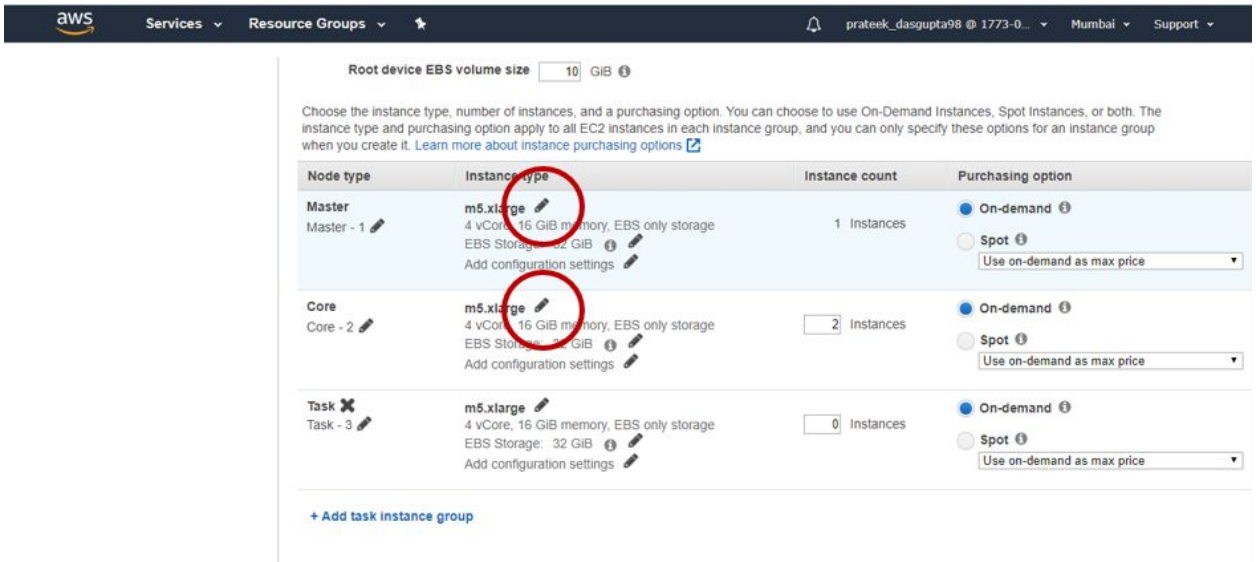


Scroll to the bottom of the page and click on **'Next'**.

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



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.



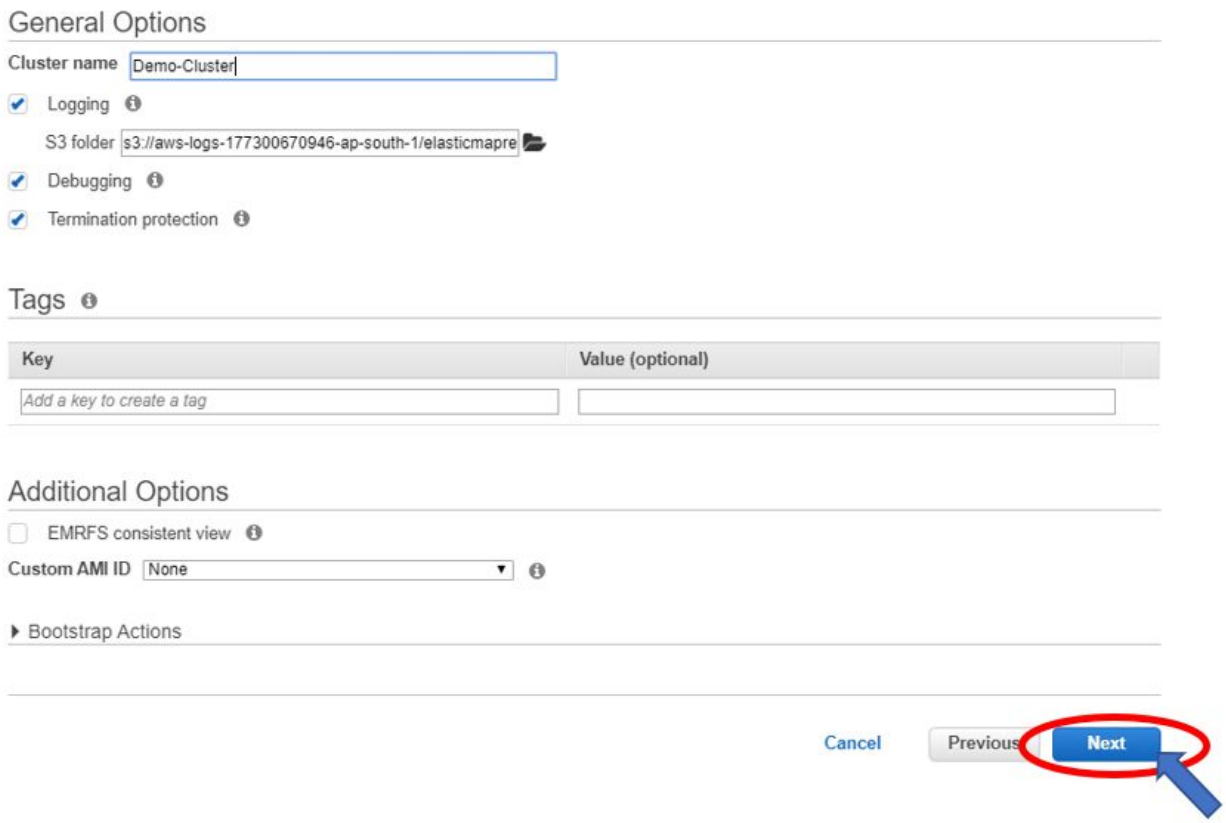
You will find a number of instance types with their corresponding RAM and Memory. Click on this link to learn about available Instance Types.

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**.

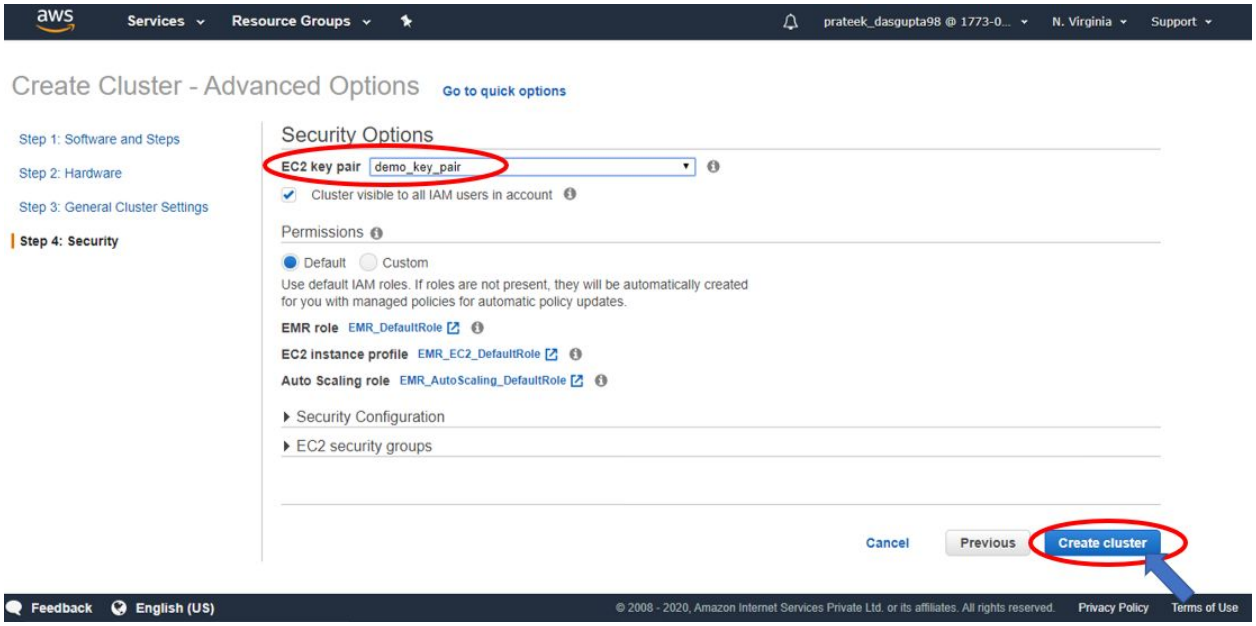
Use-m4.large Master- one instance and Core also as m4.large with 2 instances.

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	On-demand 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	On-demand 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	On-demand Spot Use on-demand as max price	Not enabled

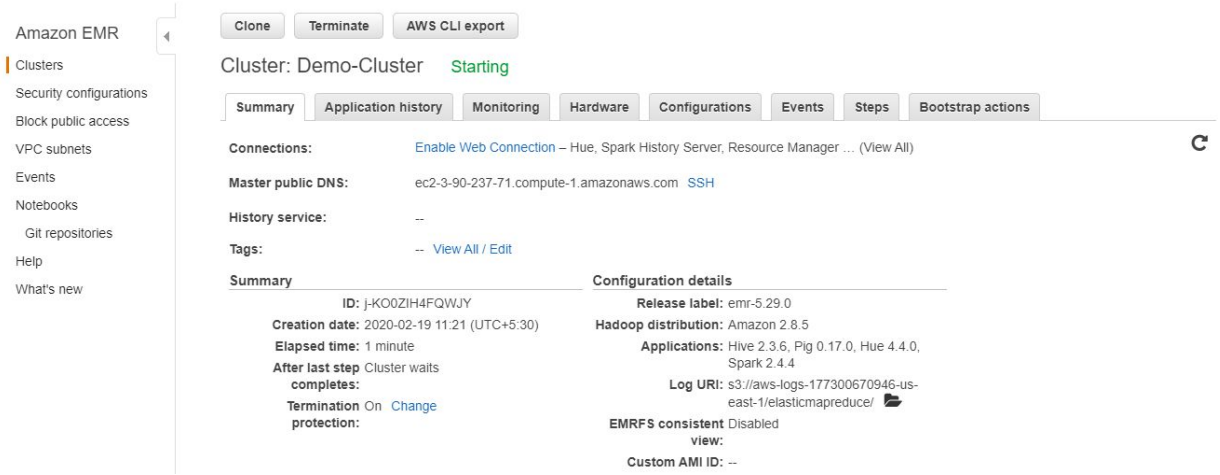
10. It is advisable to rename your cluster-name to a unique name so that later on it becomes easy to search for the cluster. For this demo, we have named it 'Demo-Cluster'. You should be having a page as shown below. Click on **Next**.



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

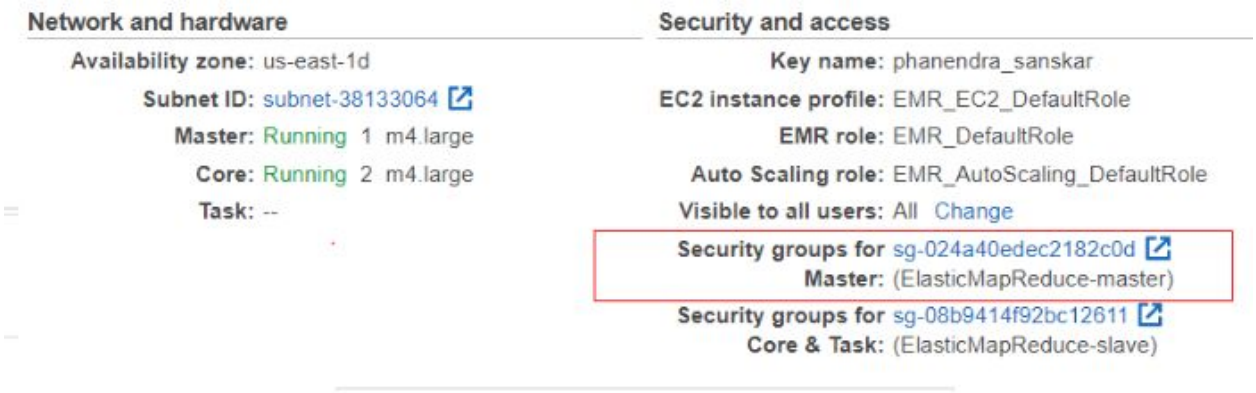


12. AWS has started launching your cluster. It will take about 10-15 minutes to launch this cluster. You will be presented with a page like this as shown below.

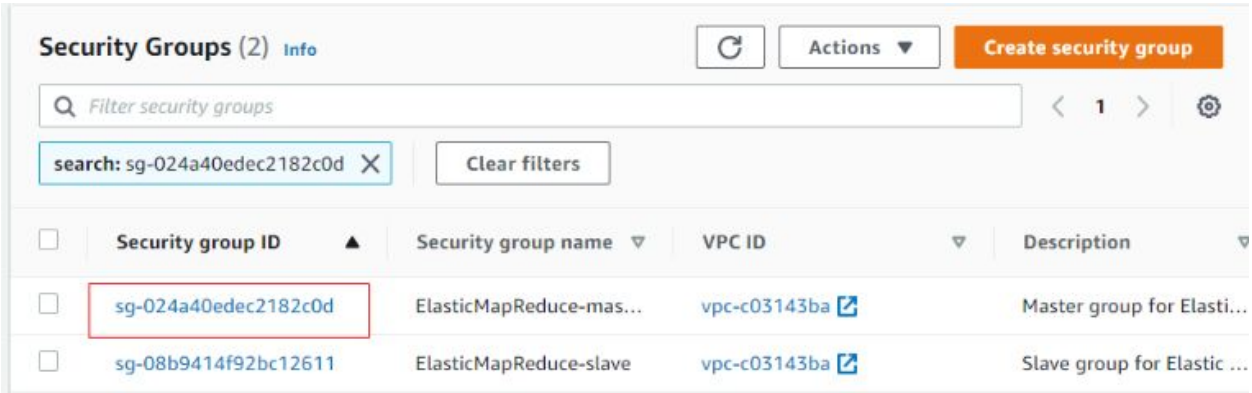


Steps to follow before performing SSH to the master node

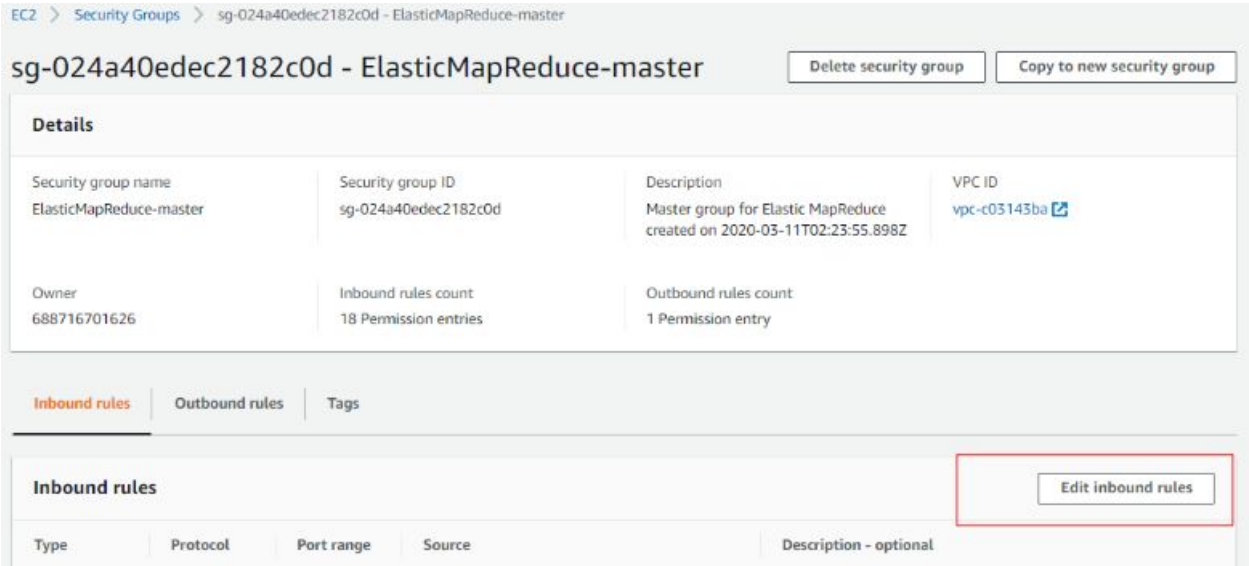
1. Under the cluster information page click on the **security groups of the master node**



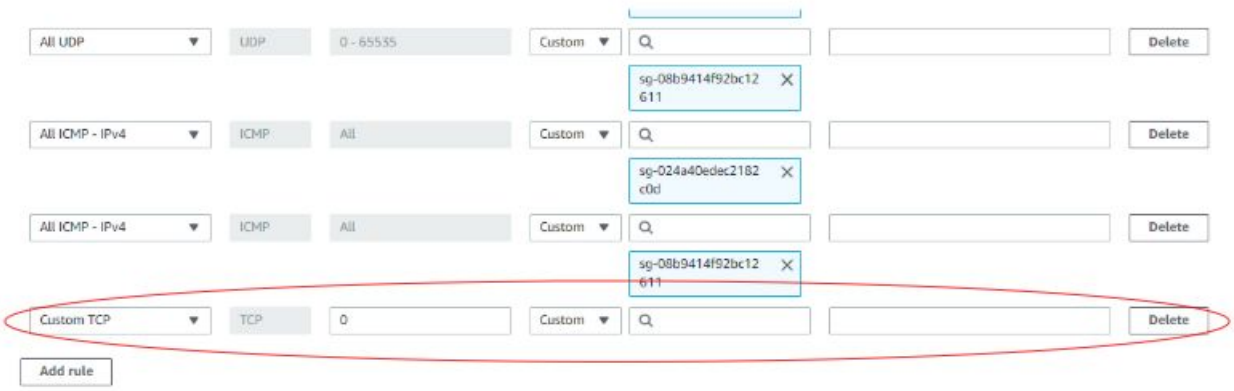
2. Clicking on the security group and you will land on a similar page. Here click on the security group of the Elastic Mapreduce-master node as highlighted in the image.



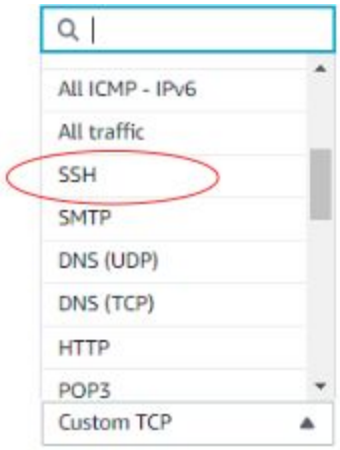
3. Clicking on the security group will land you on the corresponding security information page. Click on edit inbound rules to add a new rule



4. This will take you list of existing rules where you have the option to delete the existing rules [Clicking on delete on the extreme right-hand side] or add a new rule by clicking on **Add rule** towards the bottom of all the rules. Clicking on the add rule will add a new row as shown in the figure below.

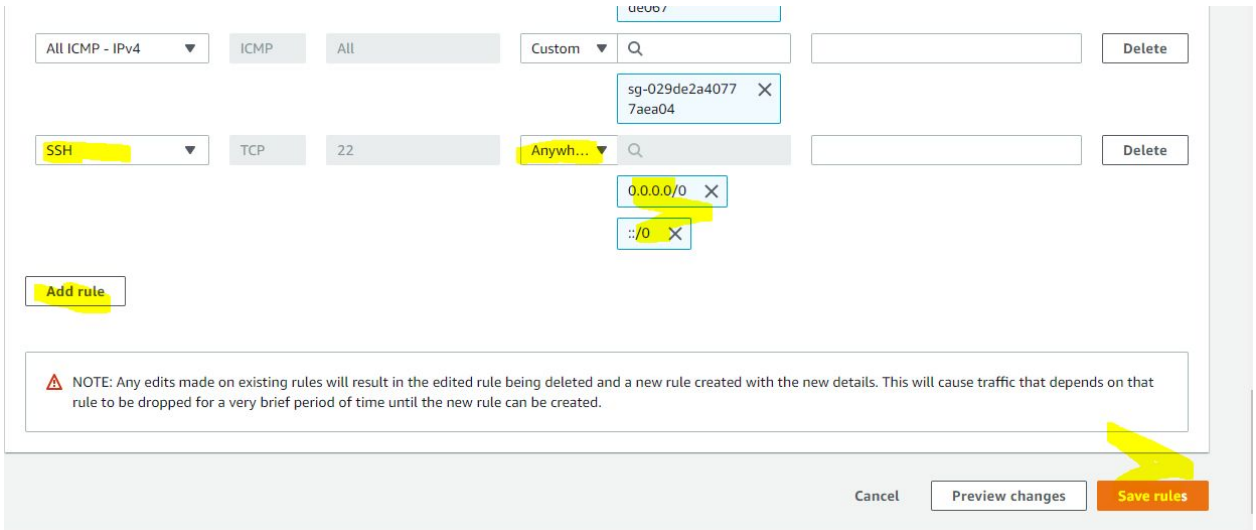


Under the type field of the newly added row select **SSH**



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

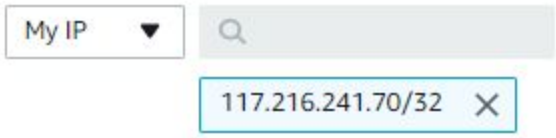


Then access the master node using putty[Windows Users] or the terminal [MAC/Linux Users].

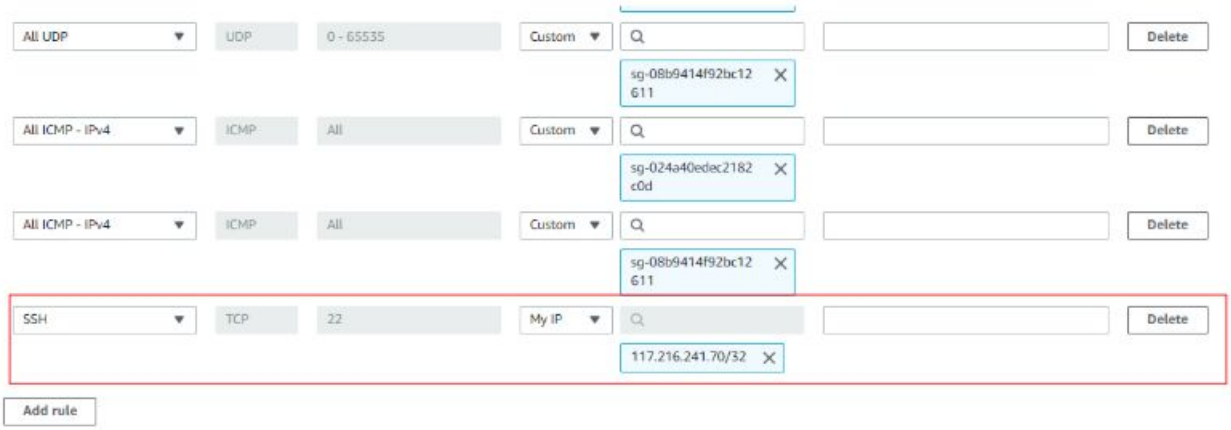
Note: Avoid choosing the My IP for EMR cluster.

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.



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



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.

All TCP	TCP	0 - 65535	Custom	sg-024a40edec2182c0d		Delete
All TCP	TCP	0 - 65535	Custom	sg-08b9414f92bc12611		Delete
SSH	TCP	22	My IP	117.198.157.204/32		Delete

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.