EMR Workshop Lab 1 - Cluster Creation

This lab demonstrates the steps involved in creating an EMR cluster.

1. Create a VPC

1. In AWS management console

* Choose VPC from the services (you may need to type VPC in the search bar).

1. In VPC dashboard

* Choose Launch VPC Wizard and click Select.

1. Type VPC name (EMRVPC for example), and leave all the rest as default.
2. Click Create and then OK.

2. EC2 key pair

Make sure you have an EC2 key pair in the region you are using.

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair>

**To create your key pair**

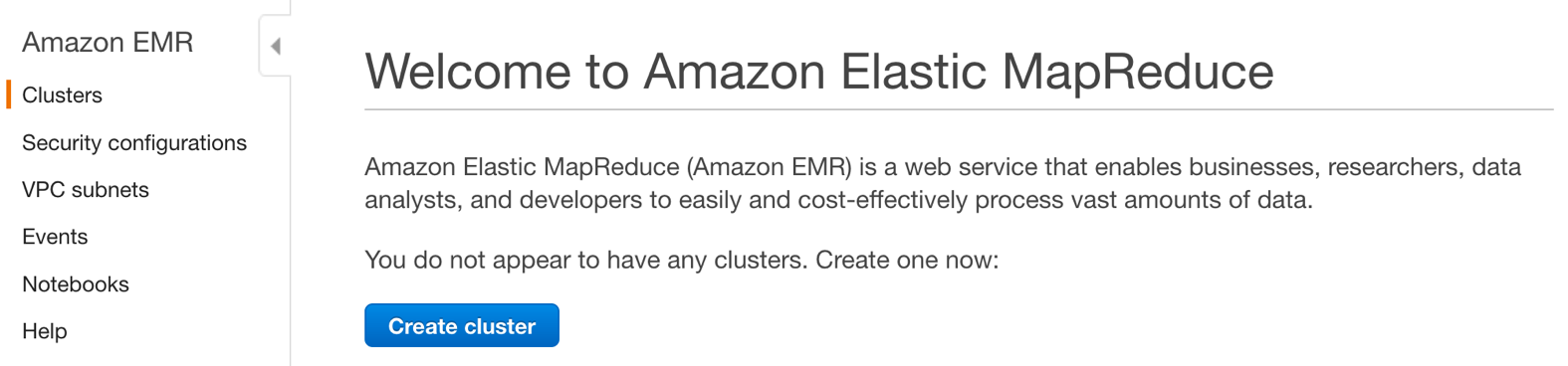
1. Open the Amazon EC2 console at <https://console.aws.amazon.com/ec2/>.
2. In the navigation pane, choose Key Pairs.
3. Choose Create key pair.
4. For Name, enter a descriptive name for the key pair.
5. For File format, choose the format in which to save the private key. To save the private key in a format that can be used with OpenSSH, choose pem. To save the private key in a format that can be used with PuTTY, choose ppk.
6. Choose Create key pair.

3. Launch EMR Cluster

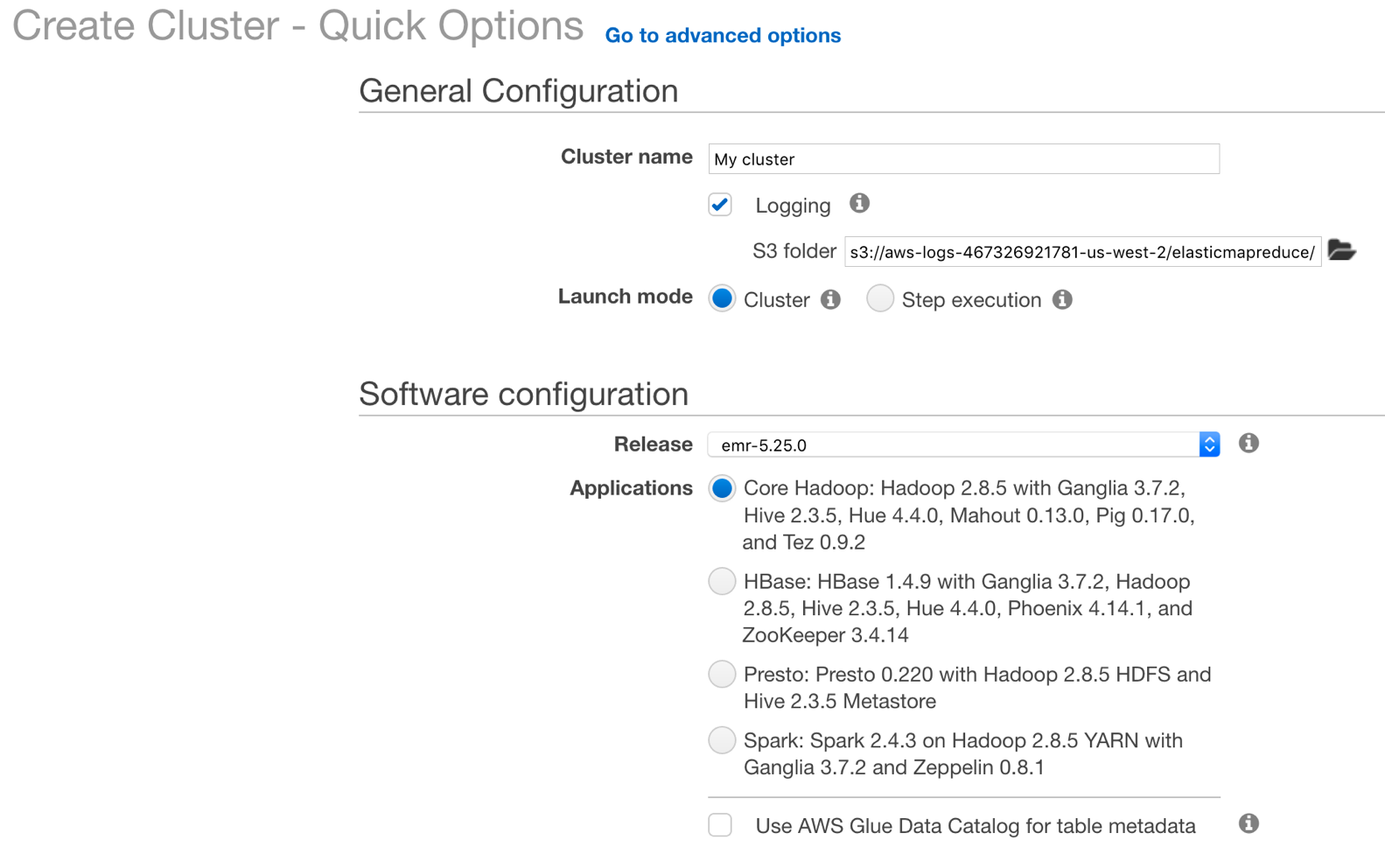
Open the Amazon EMR console at

https://console.aws.amazon.com/elasticmapreduce/

1. Click ‘Create cluster’

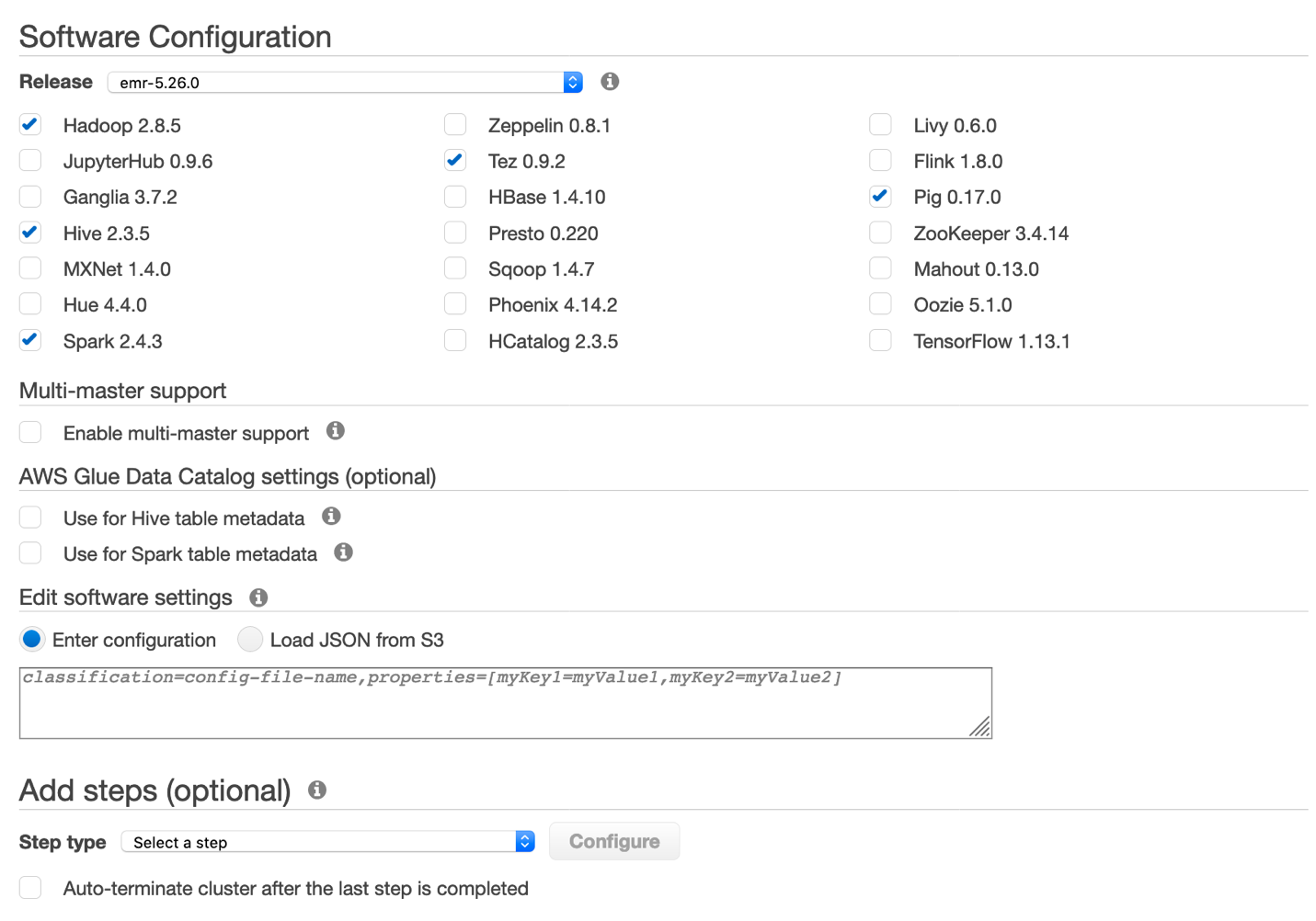


1. Click ‘Go to advanced options’



**Step 1: Software and Steps**

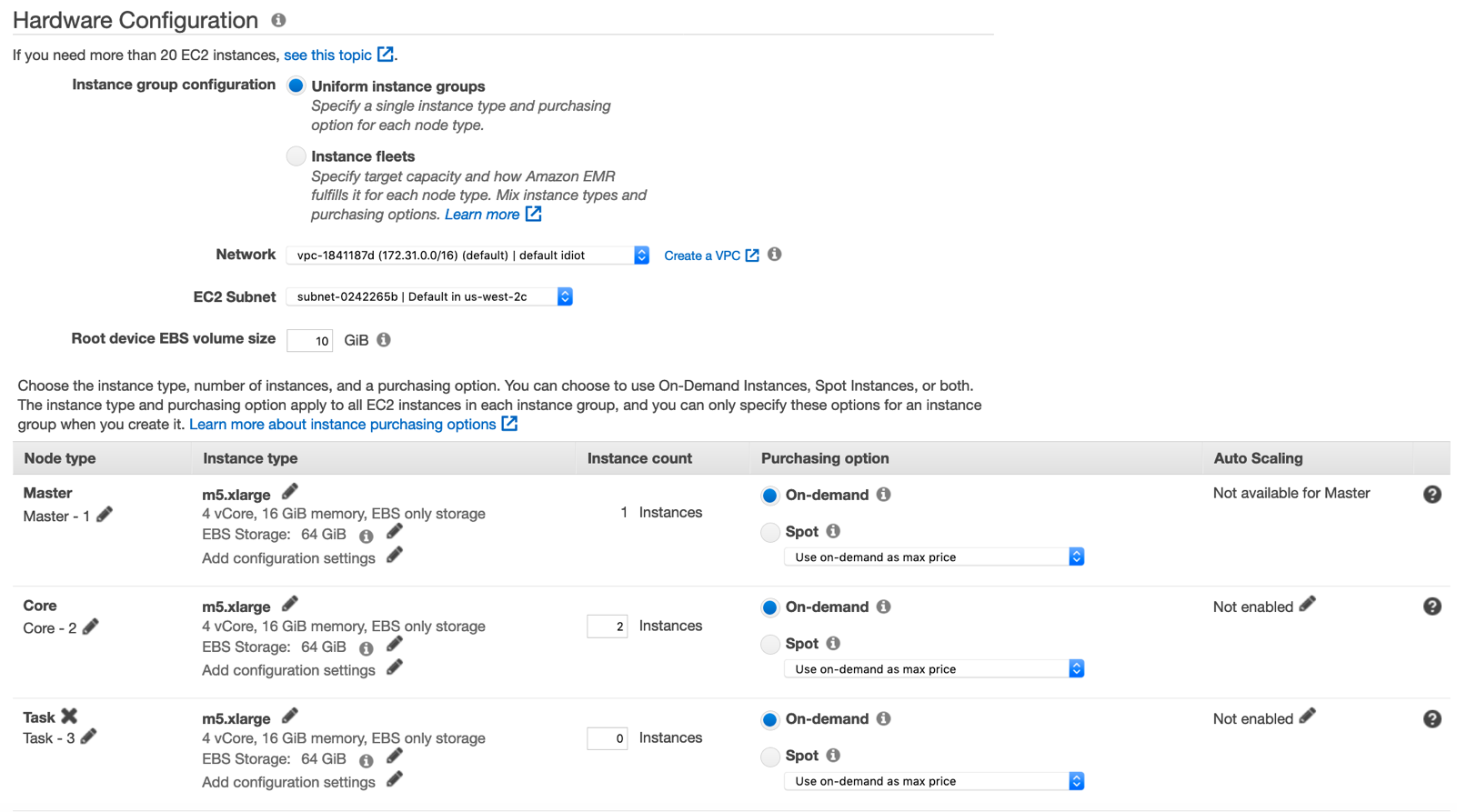
|  |  |
| --- | --- |
| Release | Leave as default |
| Software Configuration | Ensure that the following are checked:   * Hadoop * Hive * Tez * Pig * Spark |
| Multi-master support | Leave as default |
| AWS Glue Data Catalog settings | Leave as default |
| Edit Software Settings | Leave as default |
| page2image33210944Add Steps | Leave as default |



1. Click ‘Next’

**Step 2: Hardware Configuration**

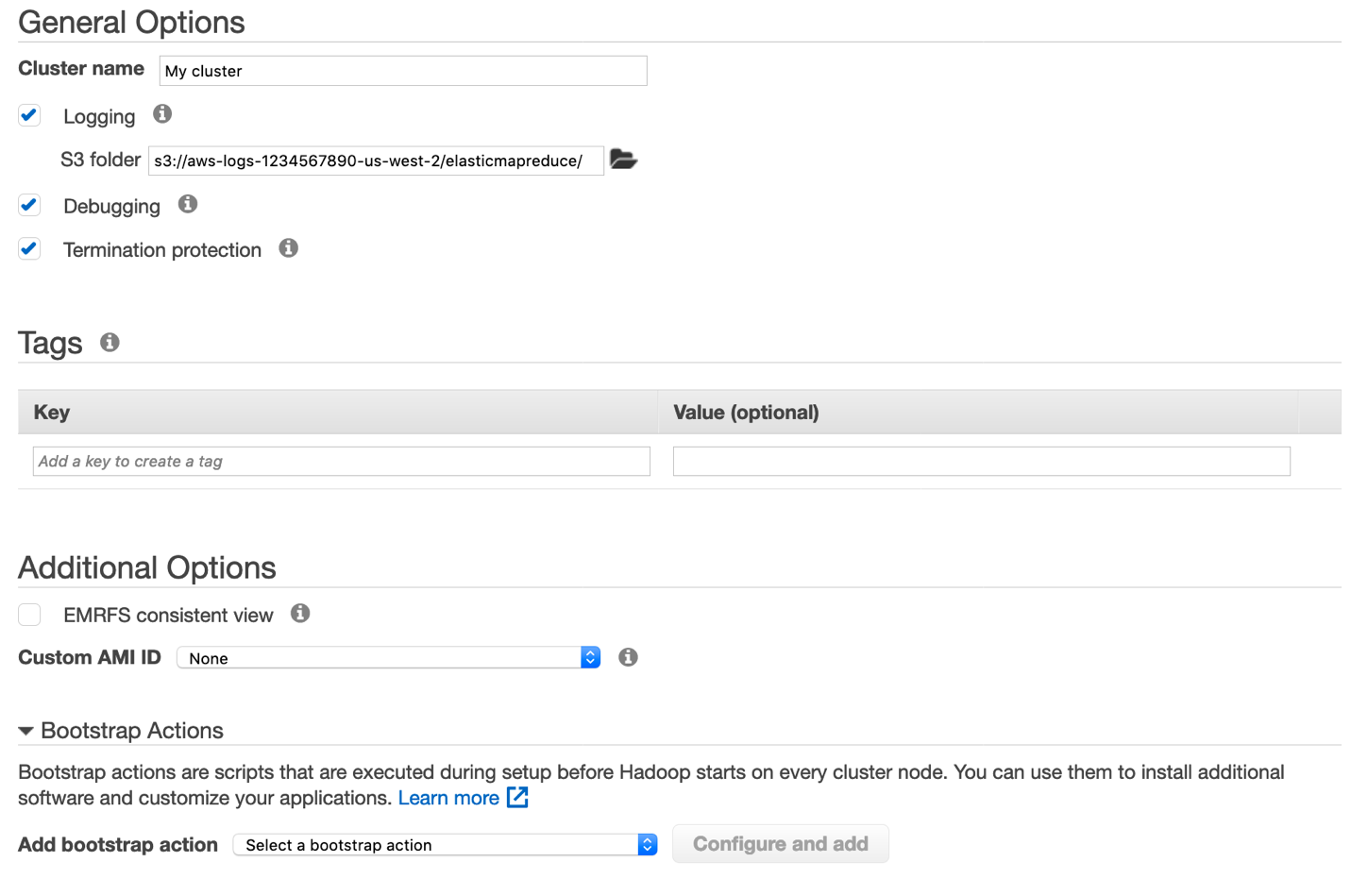
|  |  |
| --- | --- |
| Instance group configuration | Leave as default |
| Network | Select VPC created in first step |
| page3image18512928EC2 Subnet | Choose the public subnet |
| Root device EBS volume size | Leave as default (10GB) |
| page3image18734608  Instances  page3image18734816page3image33226752 | Set the cluster instances and counts as follows:   * Master: m5.xlarge, count = 1 * Core: m5.xlarge, count = 2 * Task: m5.xlarge, count = 0   page3image18513552 |



d) Click ‘Next’

**Step 3: General Cluster Settings**

|  |  |
| --- | --- |
| Cluster name | page4image18721760Name your cluster |
| Logging | Leave checked Choose a bucket in the same region |
| Debugging | Leave checked |
| Termination Protection | Leave checked |
| Tags | Leave blank |
| EMRFS Consistent View | page4image18725088Leave unchecked |
| page4image33285696Bootstrap actions | Leave alone |

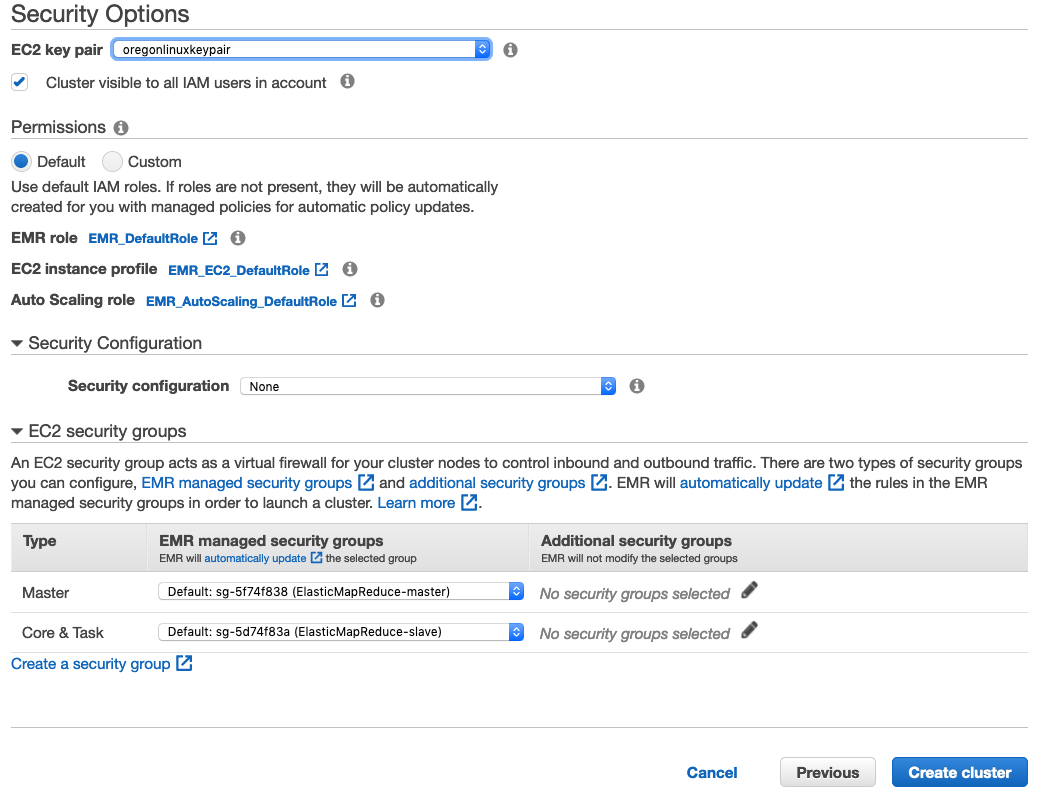


1. Click ‘Next’

**Step 4: Security**

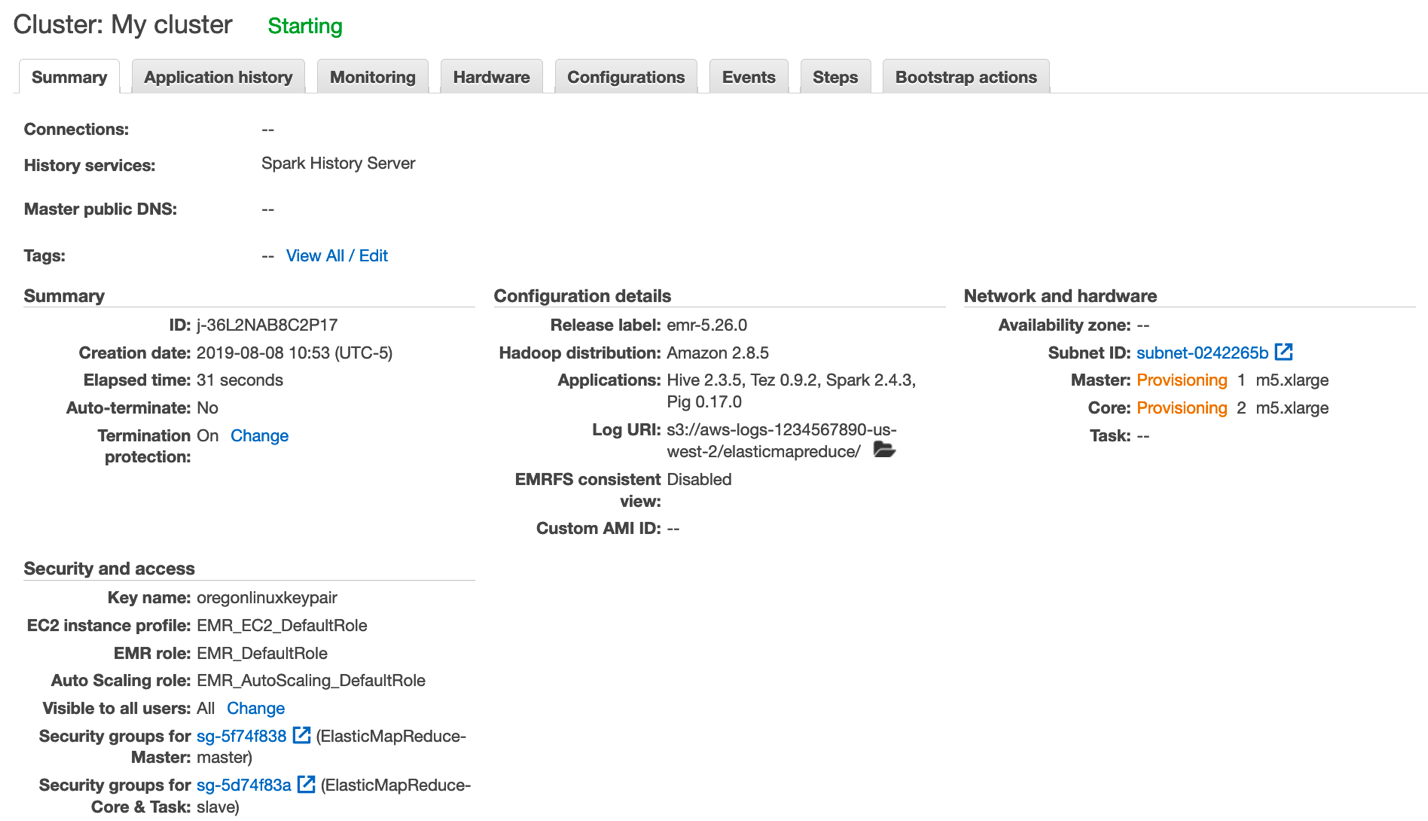
1. Ensure the following settings are correct then click ‘Create cluster’

|  |  |
| --- | --- |
| page5image18689824EC2 key pair | Choose keypair created in step 2 |
| Cluster visible page5image18690240 | Leave checked |
| Permissions page5image18690448 | Leave as default |
| page5image18704128Authentication and encryption | Leave as default |
| EC2 Security Groups page5image33053056 | Leave as default |

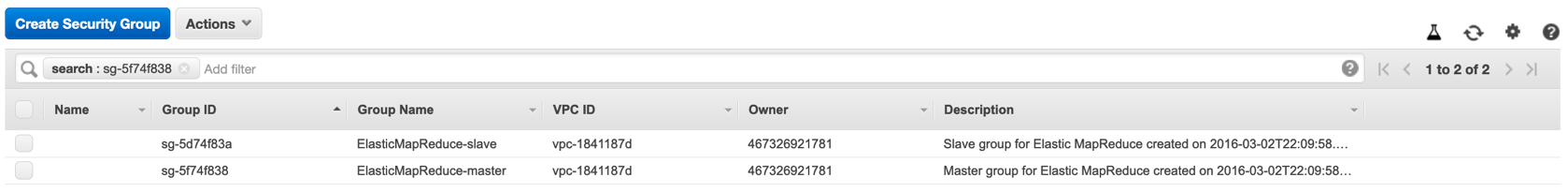


4. Update Security Groups

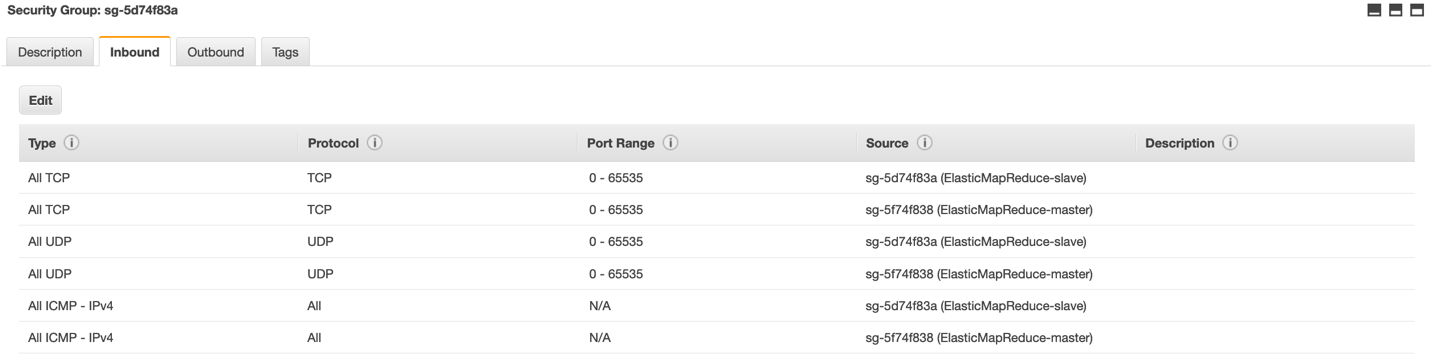
1. In the “Summary” tab, scroll down to the “Security and access” section and click on the security group shown for ‘Security Group for Master’



1. Click on the security group for ‘ElasticMapReduce-master’



1. Click on the ‘Inbound’ tab
2. Click the ‘Edit’ button



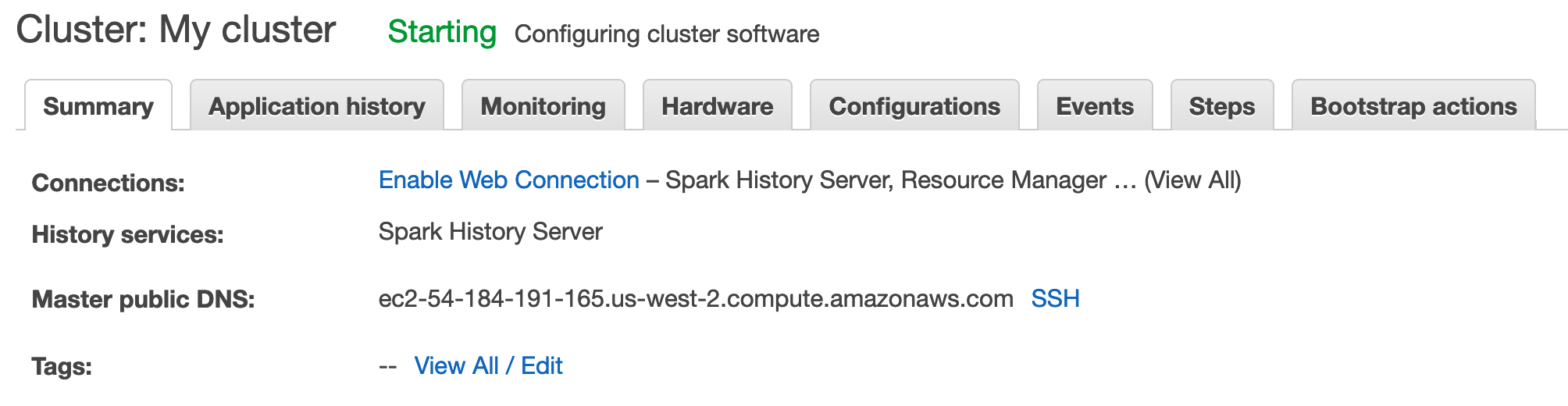
1. Click the ‘Add Rule’ button
2. Add a rule that allows SSH from your IP Address



1. Click ‘Save’

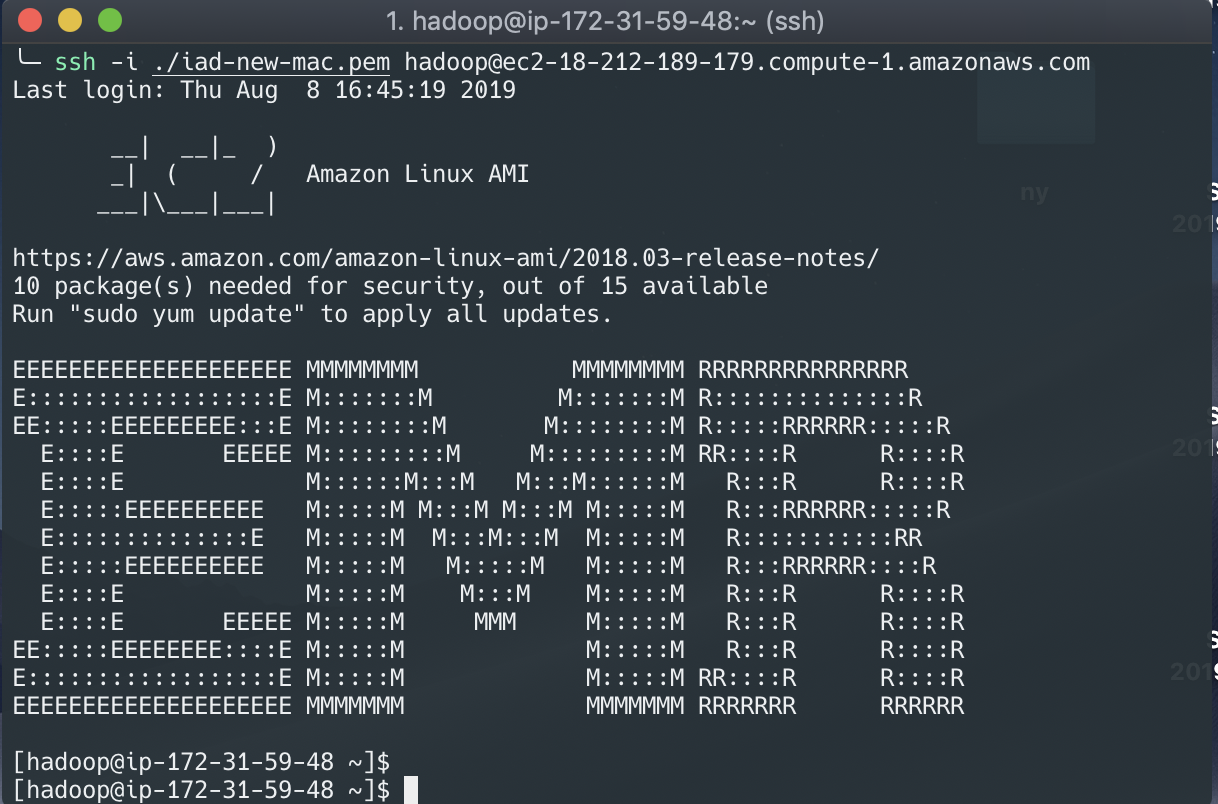
5. SSH into the Cluster

Once the cluster shows as “Waiting – Cluster ready” open up a terminal or SSH client and SSH into the master node. Obtain the SSH connection information from your cluster summary tab:



>> ssh -i <<key-pair>> hadoop@<<emr-master-public-dns-address>>

If logged in successfully you should see the following screen:



**\*\* Note** if you are unable to SSH successfully into the master node ensure your IP is correct in the security group rule you set up for the master. Also ensure the cluster is in a waiting status and isn’t still being created as you cannot SSH into the master node until the cluster is fully launched.