AWS EMR Instructions

These instructions walk you through the process of creating an initial Amazon EMR (Elastic Map Reduce) cluster

Note, the EMR cluster you set up using these instructions is not meant for a production (secure) environment, and do not cover configuration options in depth. It is meant to help you set up a cluster for class purposes as quickly as possible.

Charges accumulate for cluster you create at the per-second rate for Amazon EMR pricing. The cost will be minimal because the cluster should run for less than a couple of hours after the cluster is provisioned. So, it is important that you decommission the cluster as instructed below after you are done with an assignment.

**Step 1: Prerequisites**

Before you begin setting up your Amazon EMR cluster, make sure that you have completed assignment #1, have an AWS account and understand the basics of working with S3 buckets and associated data objects.

## Step 2: AWS Management Console

## When you log in to AWS you are presented with the AWS Management Console page. Wherever you are on the site, you can always return to the management console page by clicking on the AWS logo at the top left.

## 

## Step 3: Finding Services

## We will be making use of several AWS services including

## EC2 – provides computing capability in the form of virtual machines (servers)

## S3 – for object storage

## EMR – Elastic Map Reduce, the Hadoop cluster as a service

## When you are on the AWS Management Console page (which we can always get to by clicking the AWS logo), you can find the main page for a service by doing one of the following

1. Type the name of the service whose web page you want to reach into the “Search” text box and press Enter/Return
2. If you typed in the name of or used a service recently you might be able to find its name by clicking on “Recently visited” services and then clicking on the name of the desired service
3. If you don’t recall the name of the service, then click on “Services” to get a list and click on the service of interest.

So, in the following steps when you are requested to find some service, you can do the above.

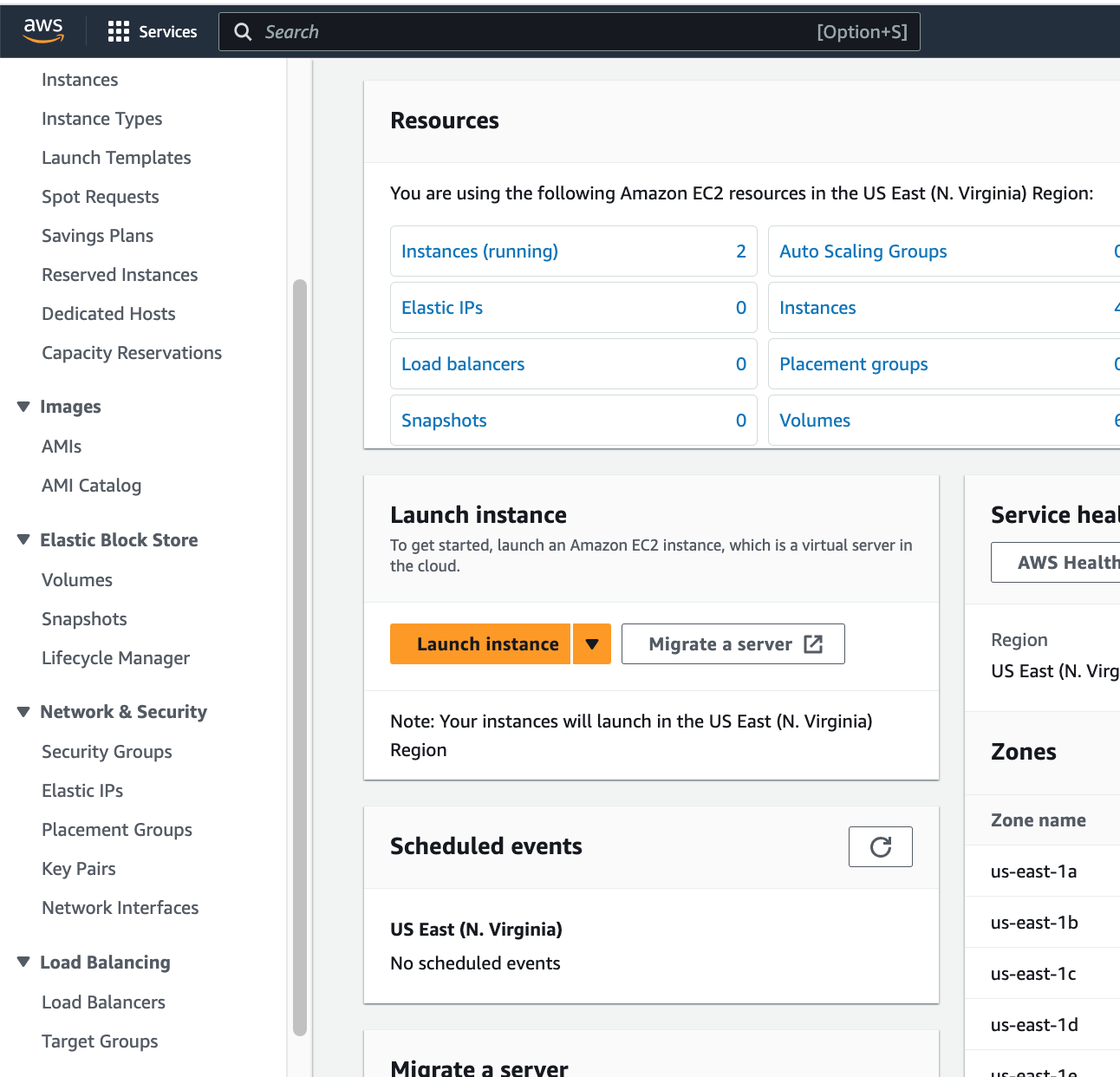
## Step 4: Create an Amazon EC2 Key Pair

You must have an Amazon Elastic Compute Cloud (Amazon EC2) key pair to connect to the nodes in your EMR cluster over a secure channel using the Secure Shell (SSH) protocol. We will understand more about SSH below.

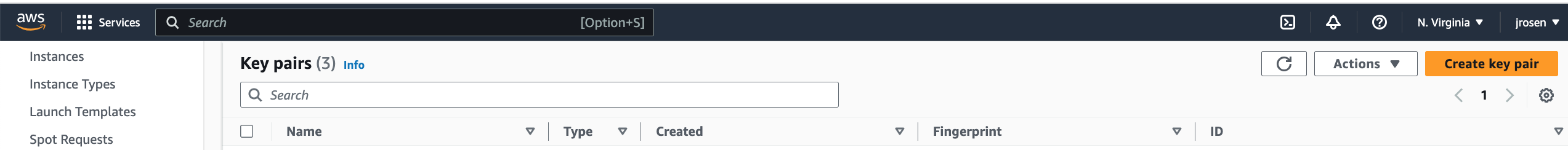
1. Find the EC2 service page
2. In the navigation pane, under **NETWORK & SECURITY**, choose **Key Pairs**.

**Note**

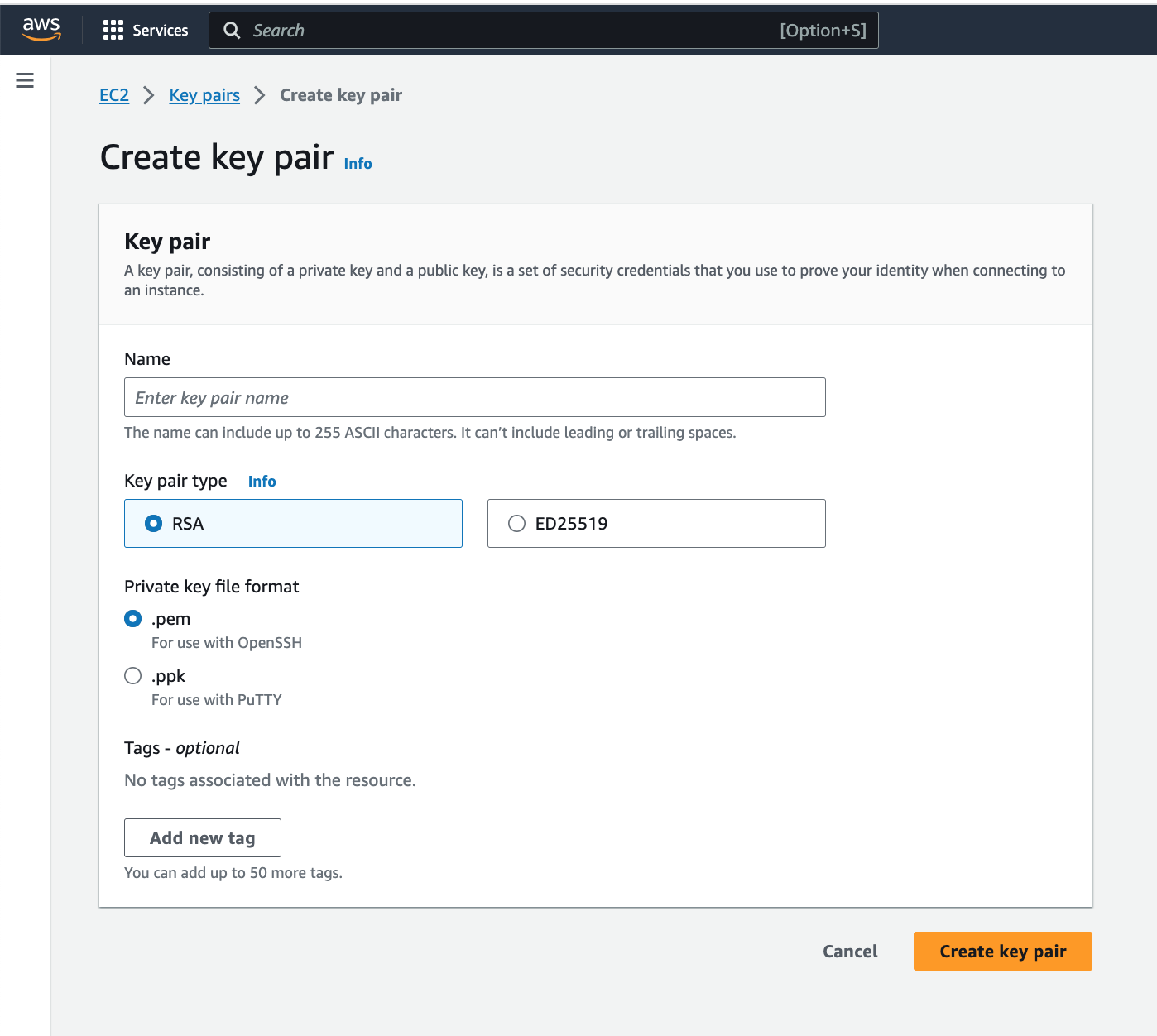
The navigation pane is on the left side of the Amazon EC2 console. If you do not see the pane, it might be minimized; choose the arrow (or three parallel bars) to expand the pane.



1. You should them see a pane similar to the following. Choose “Create key pair”



1. Then you should see the following form. For the key pair name, enter a name for the new key pair (something like emr-key-pair), and then choose **Create key pair**. Leave other options as they are, unless you are using Putty, then check ‘ppk.



1. The private key file is automatically downloaded by your browser. The base file name is the name you specified as the name of your key pair, and the file name extension is .pem or “.cer” (or .ppk). Save the private key file in a safe place.

**NOTE: Sometime AWS downloads the key file with a file suffix of “.cer” instead of “.pem” file, even if you request a “.pem” file. This is perfectly ok; the content of the file is the same. In this case, everywhere in the course, where you see instructions or examples refer to files with the “.pem” suffix, just substitute the “.cer” suffix and all will work as intended. Or, if you like you can just rename the file to use the “.pem” suffix, but this is not necessary. Remember, if you can’t find a “.pem” file on your computer check for a “.cer” file.**

In most cases on the Mac the file will download to the directory

/Users/<username>/Downloads

And on the PC the file will most likely download to

/c/Users/<username>/Downloads.

Note, the way I have written the path to the file is formatted for when using the git bash utility.

**Important**

This is the only chance for you to save the private key file. You'll need to provide the name of your key pair when you launch an instance and the corresponding private key each time you connect to the instance. But, if you lose track of it, you can create another by repeating the above steps.

1. So, find the directory into which your .pem file has been downloaded and either keep it there or move it to another directory of your choice. You will need to know the path to this file.
2. Using the “terminal” program on the MAC or the “bash” utility on the PC execute the following command to set the permissions of your private key file so that only you can read it. Note, use the appropriate path and file name for your situation.

**chmod 400 <path-to-file>/*emr-key-pair*.pem**

Note, depending on the operating system used for your personal computer, the above may not work. Things might still be ok, but if not reach out to me.

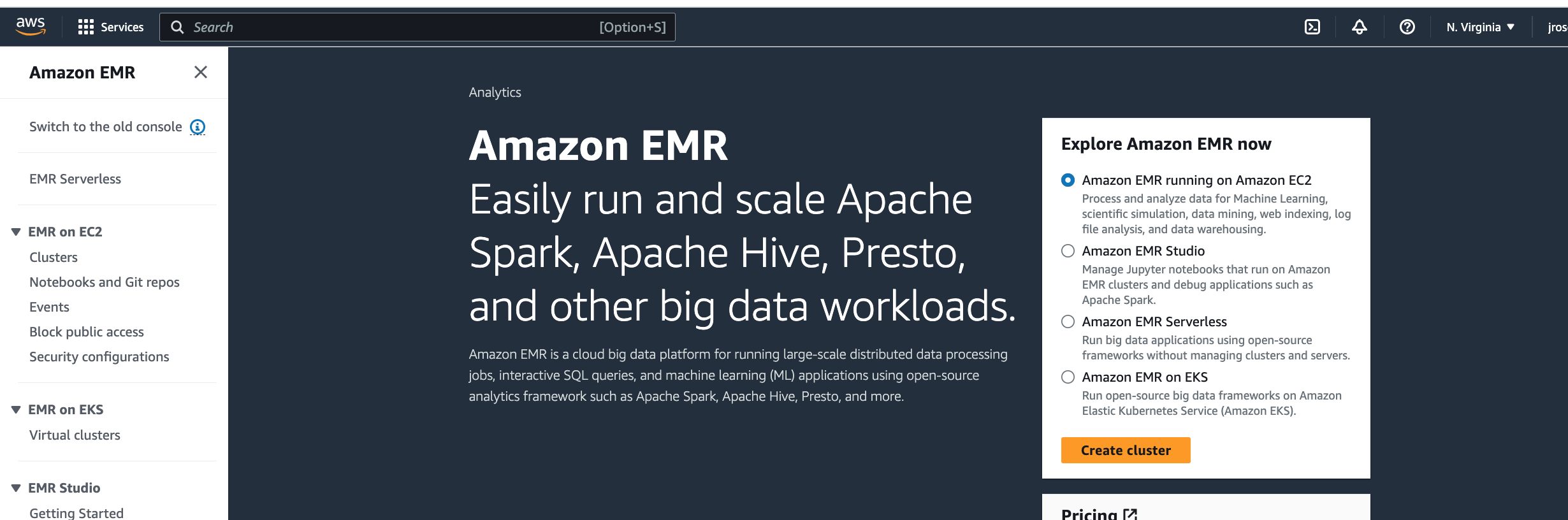
**Step 5: Launch Your Initial Amazon EMR Cluster**

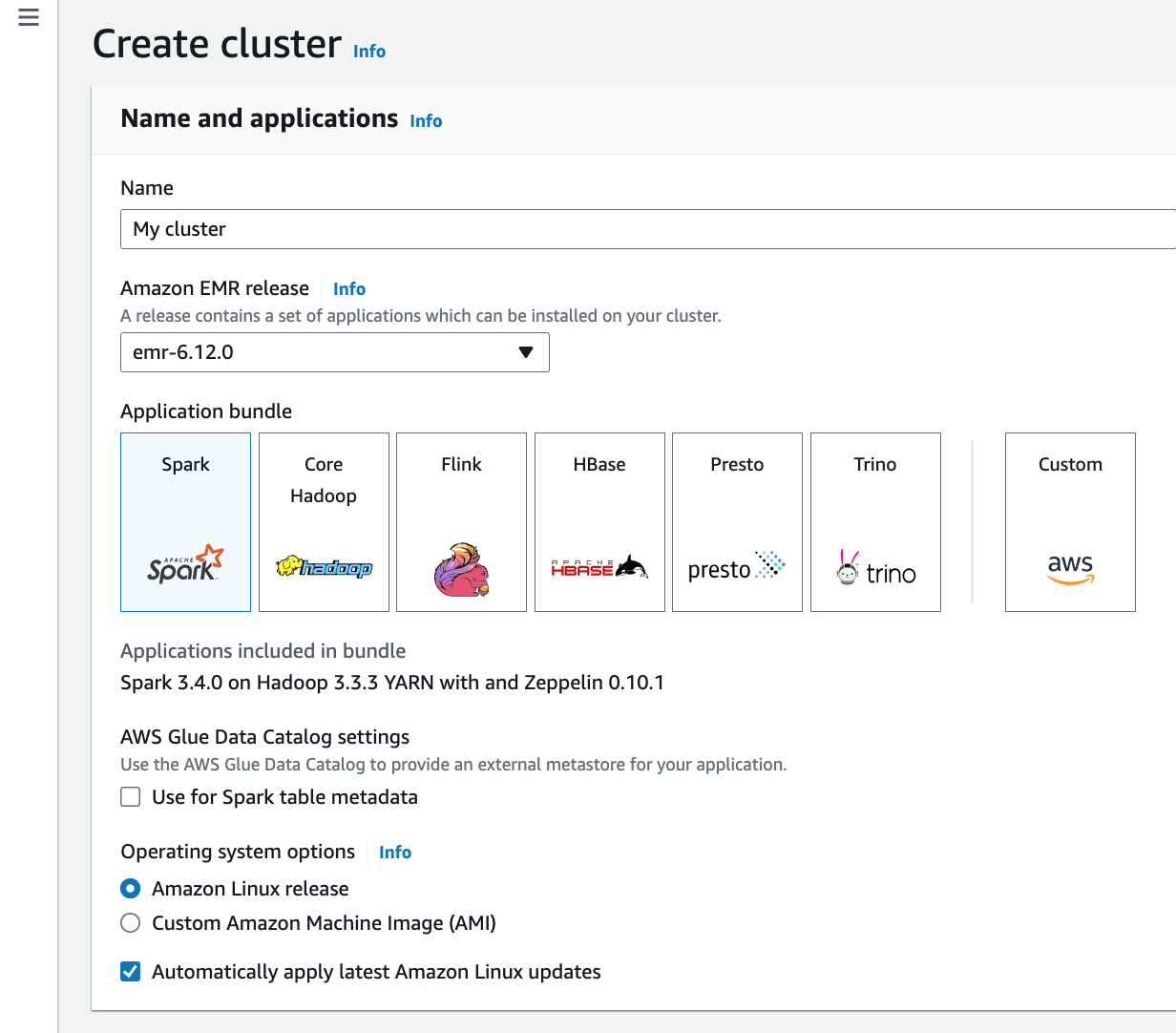
In this step, you launch your initial cluster in the Amazon EMR console.

**To launch the sample Amazon EMR cluster**

1. Find the EMR console page
2. Choose **Create cluster**.

Note, depending on how you got to the EMR service, you might see the following page, but eventually you will see the next one:

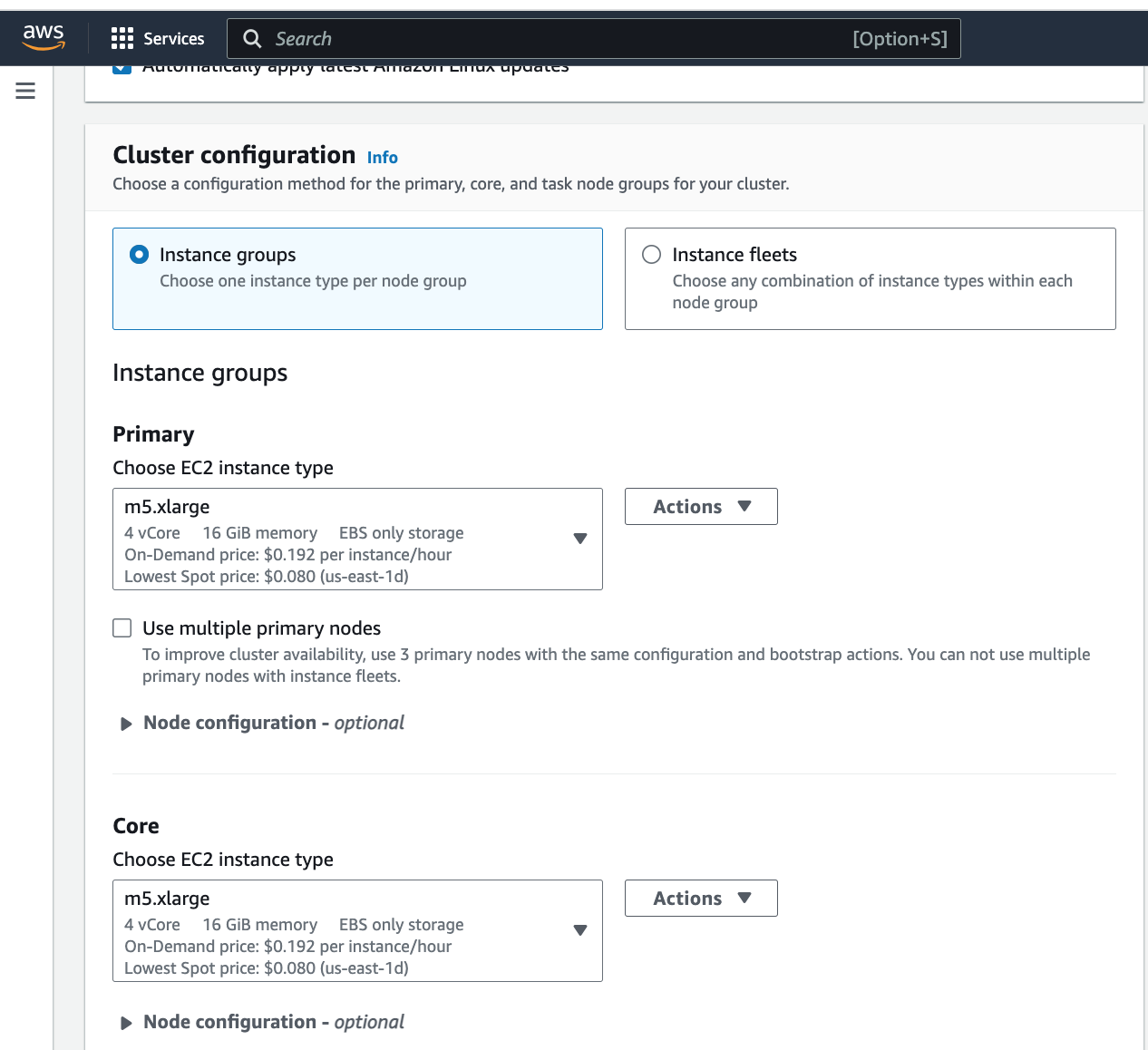




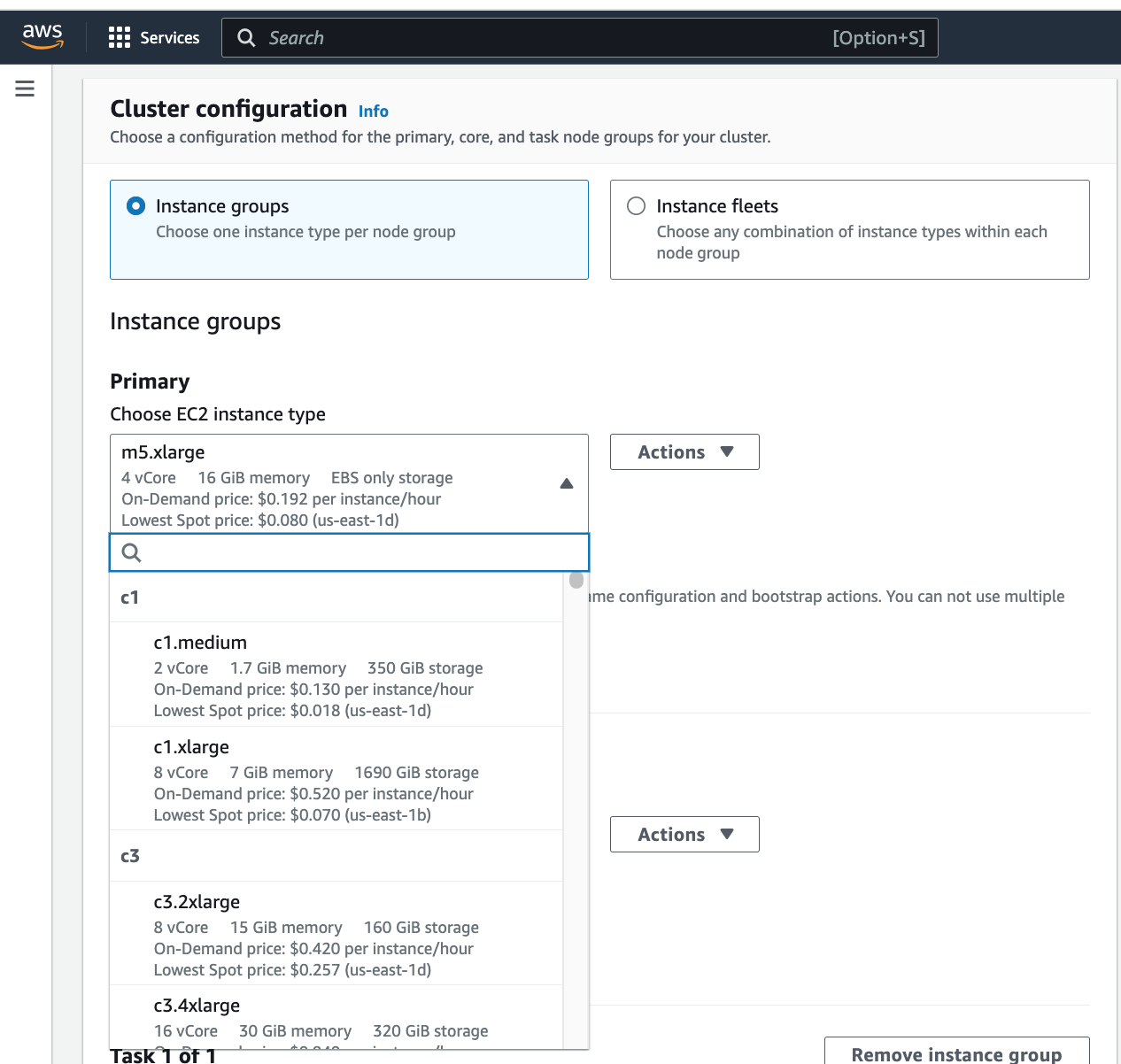
1. In the **Name and applications** section of this page:

* Enter a **Cluster name** that helps you identify the cluster, for example, *My First EMR Cluster*.
* Under **Application bundle,** select “Core Hadoop”

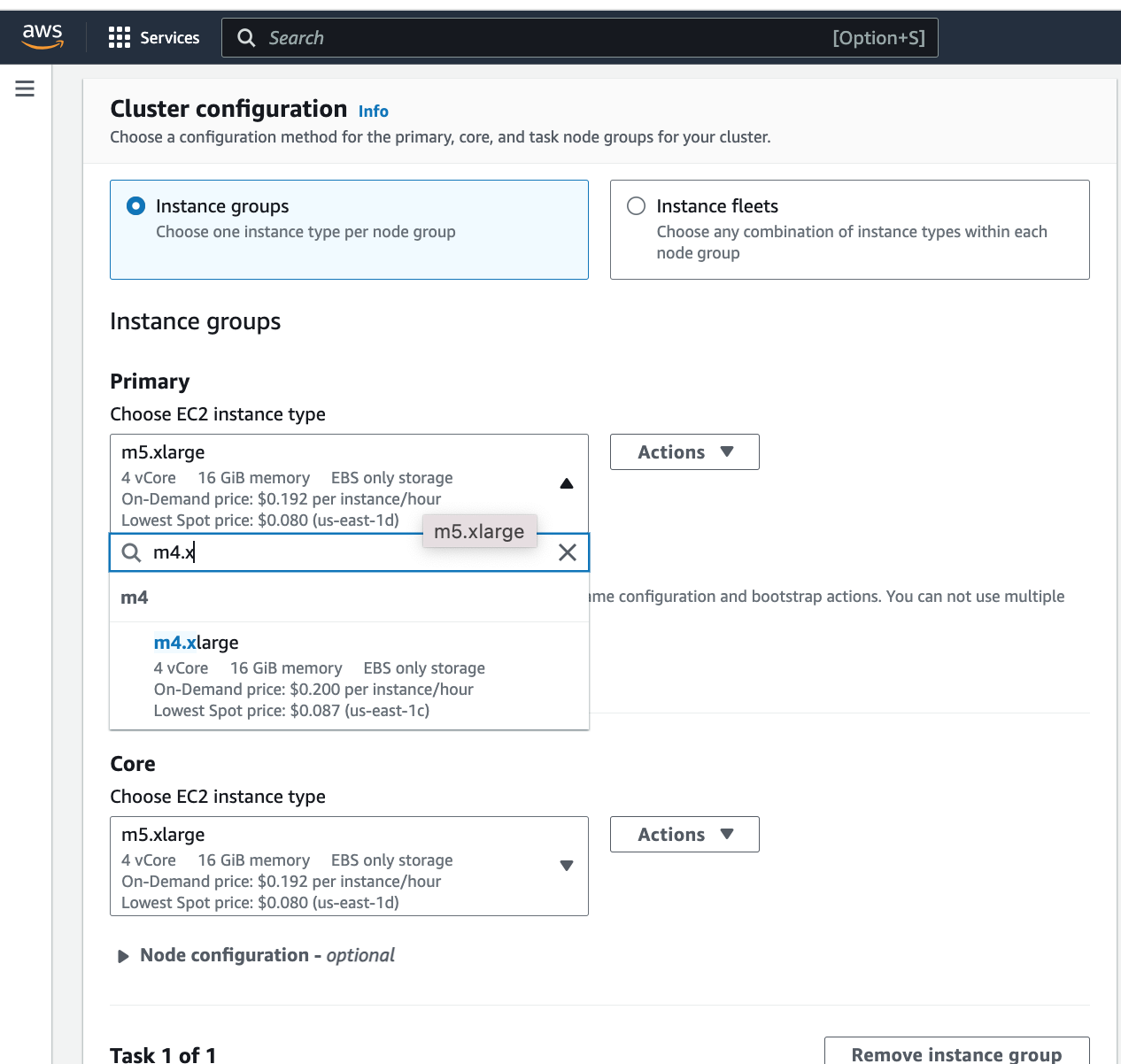
1. Scroll down to the **Cluster configuration** section of this page, you should see the following to start:



We are going to choose a different instance type than the default for the Primary and Core instance types. Underneath “Primary”, click the downward pointing arrow next to the default instance type (in the above screen shot it is m5.xlarge). You should see something like the following:

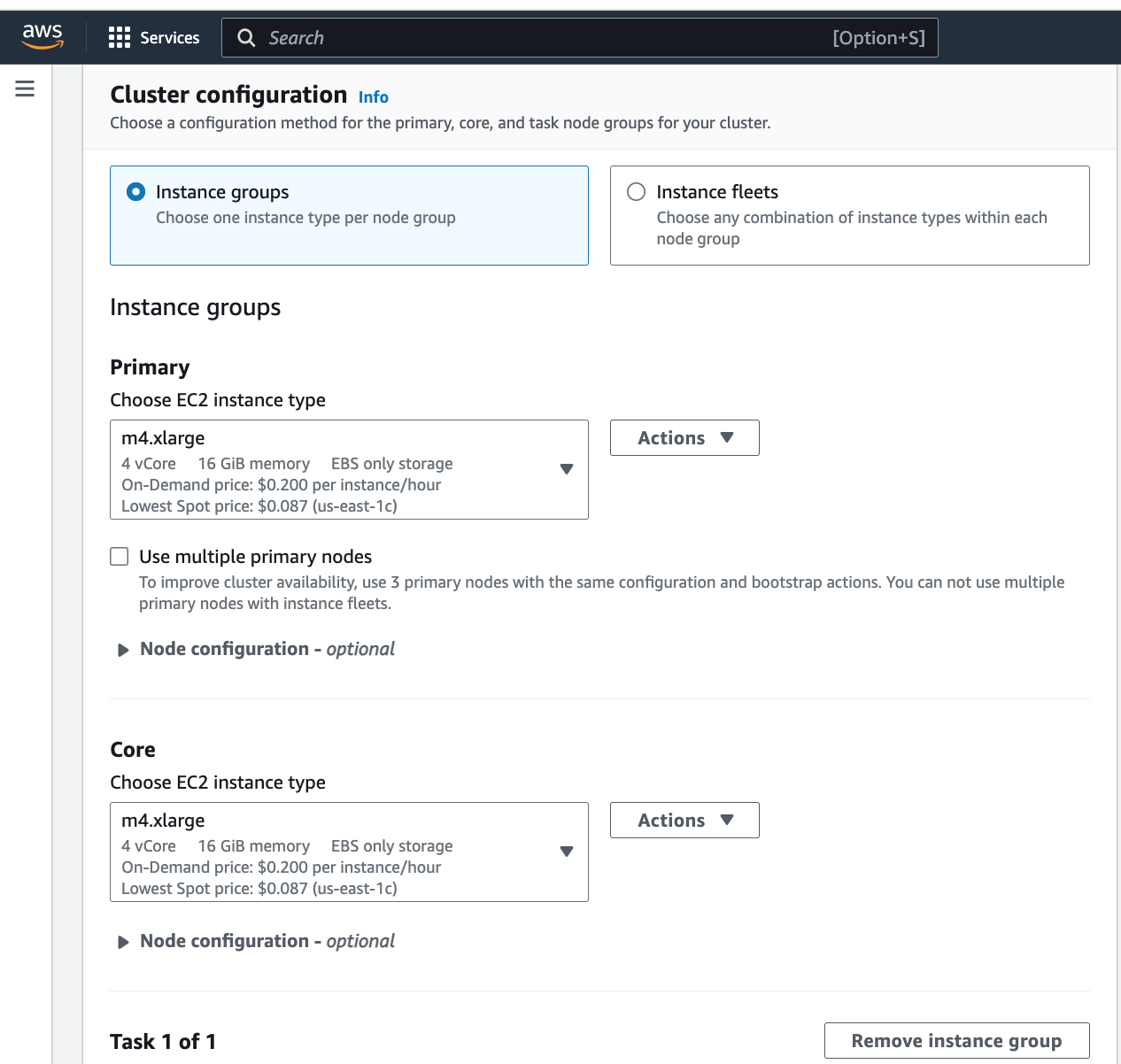


Enter “m4.x” into the search box of the drop down menu and you should see the following:

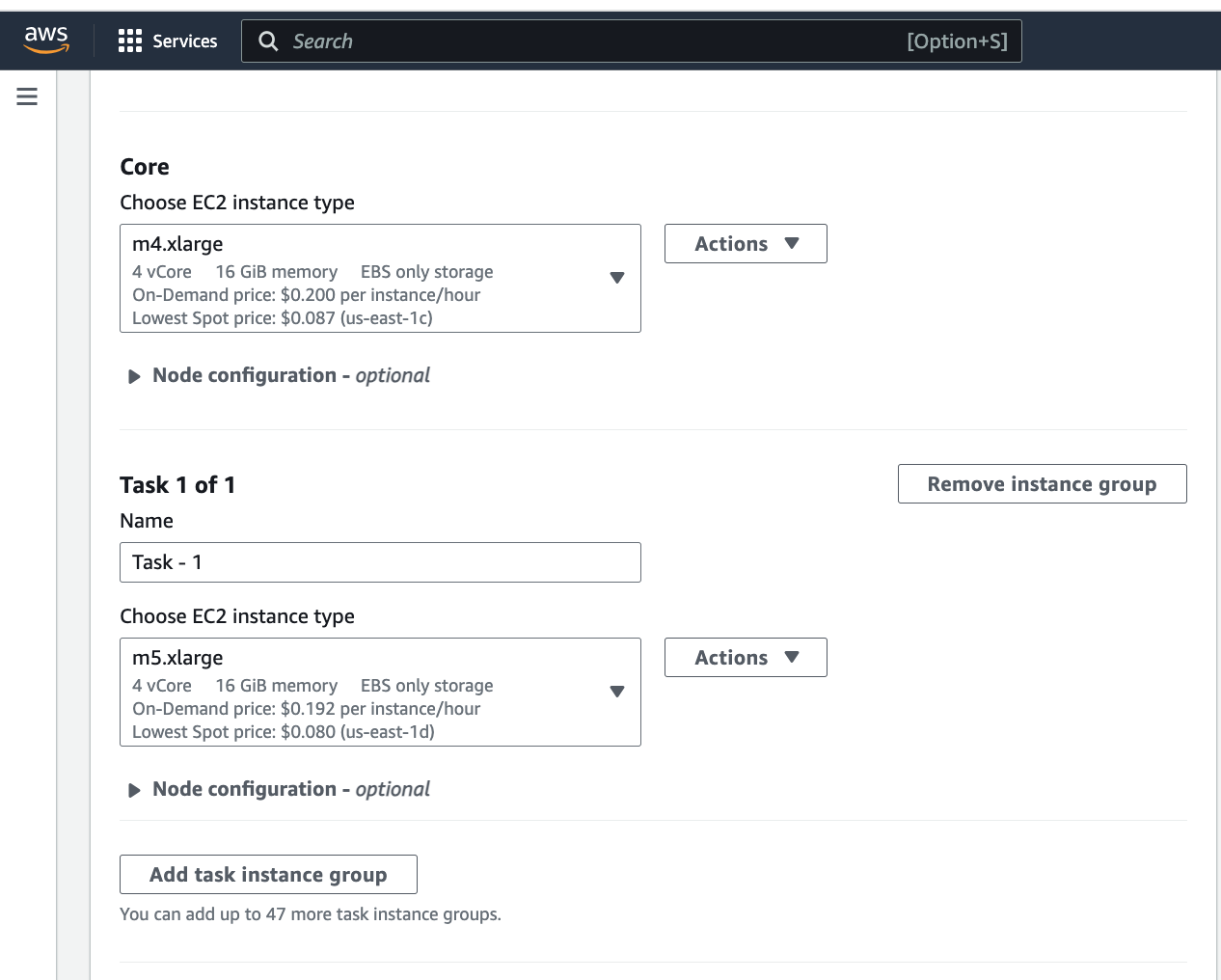


Select the m4.xlarge instance.

Now do the same for the **Core** instance. After this, if all goes well, you should see:

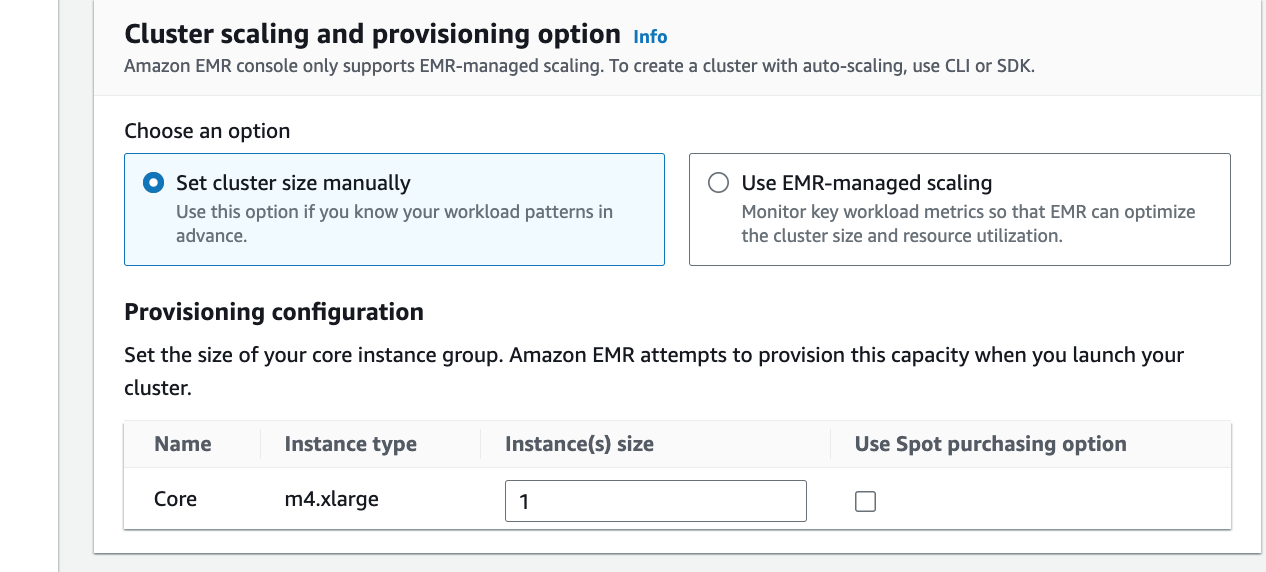


1. Below the Core instance in the **Cluster configuration** section, you should see the following labeled “Task 1 of 1.



Click on “Remove instance group”

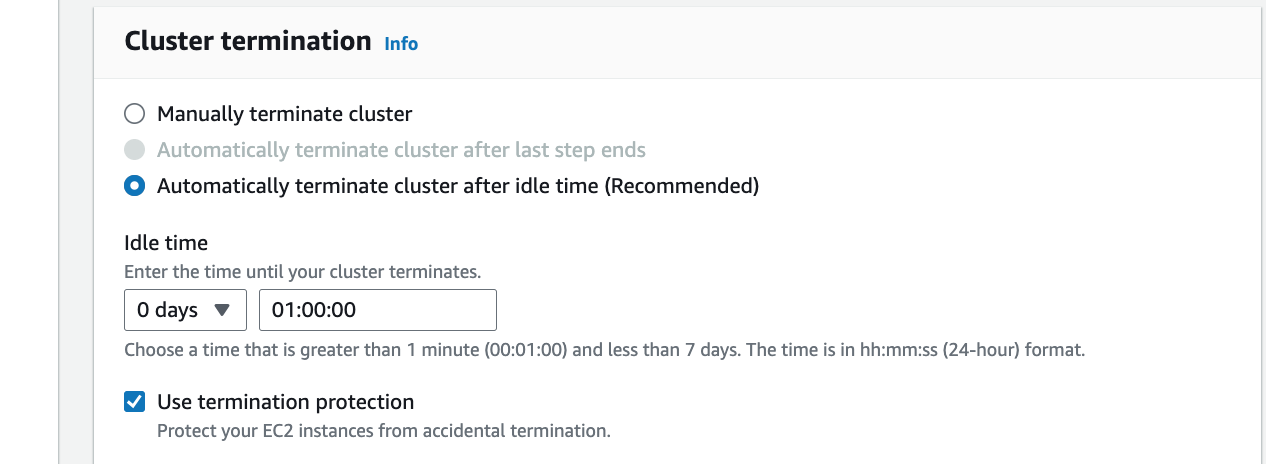
1. Now in the “Cluster scaling and provisioning options section you should see the following:



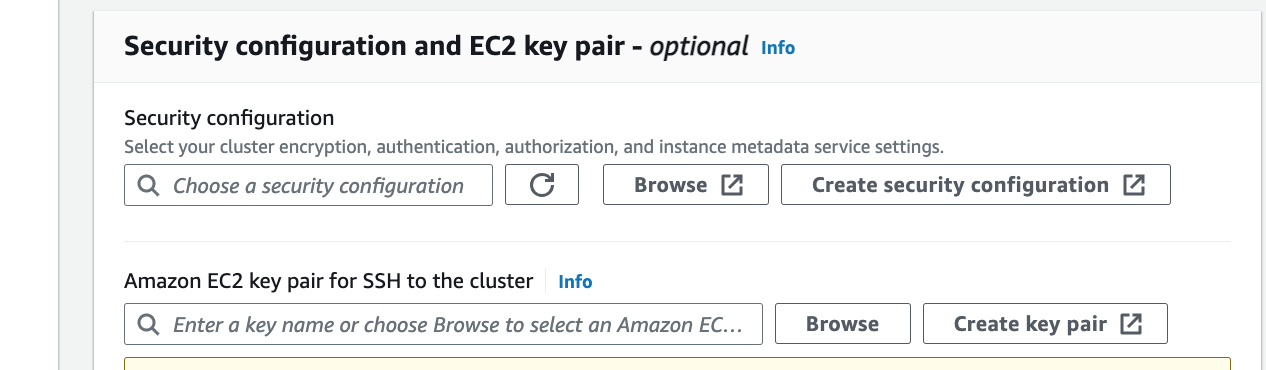
Make sure that the Instance(s) size” is selected to be “1”

1. The next step is very important to ensure you do not leave your cluster running after you finish your assignment, which could lead to unexpected and high AWS charges:

In the **Cluster termination** section, select the “Idle time”, to some number of hours. By default, this is one hour (01:00:00) which might be too short for you to complete your assignment. You migh want to select between 2 and 4 hours (02:00:00 to 04:00:00). After this time your cluster will auto-terminate.

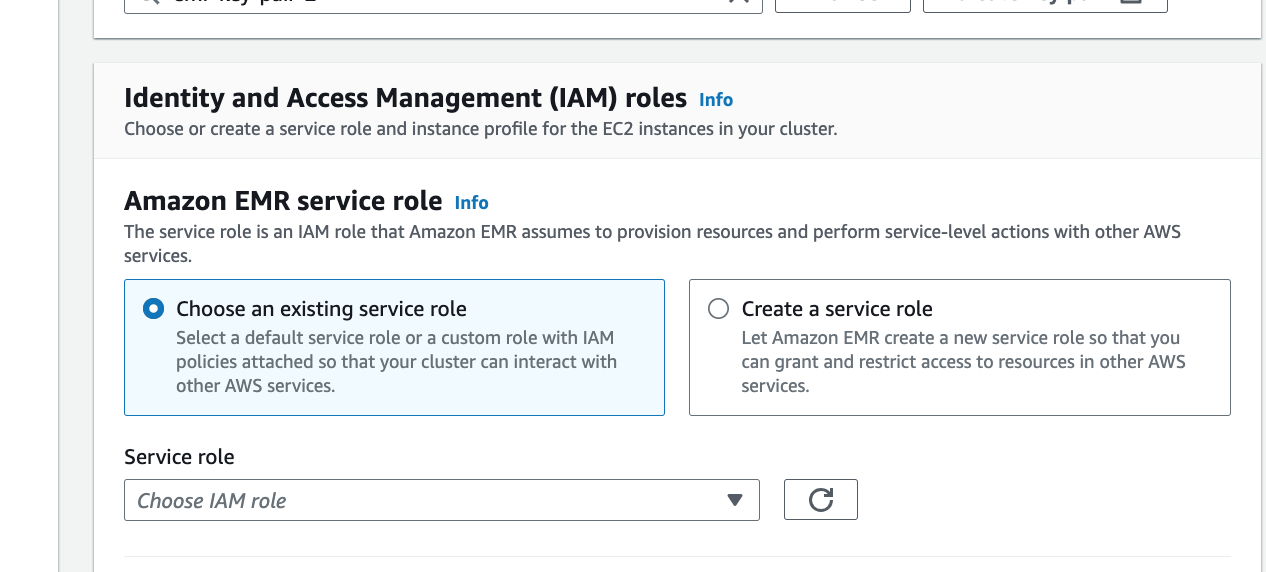


1. Now you need to associate your EMR cluster with the security key you created. Scroll down to the **Security configuration and EC2 key pair** section, where you should see the following:



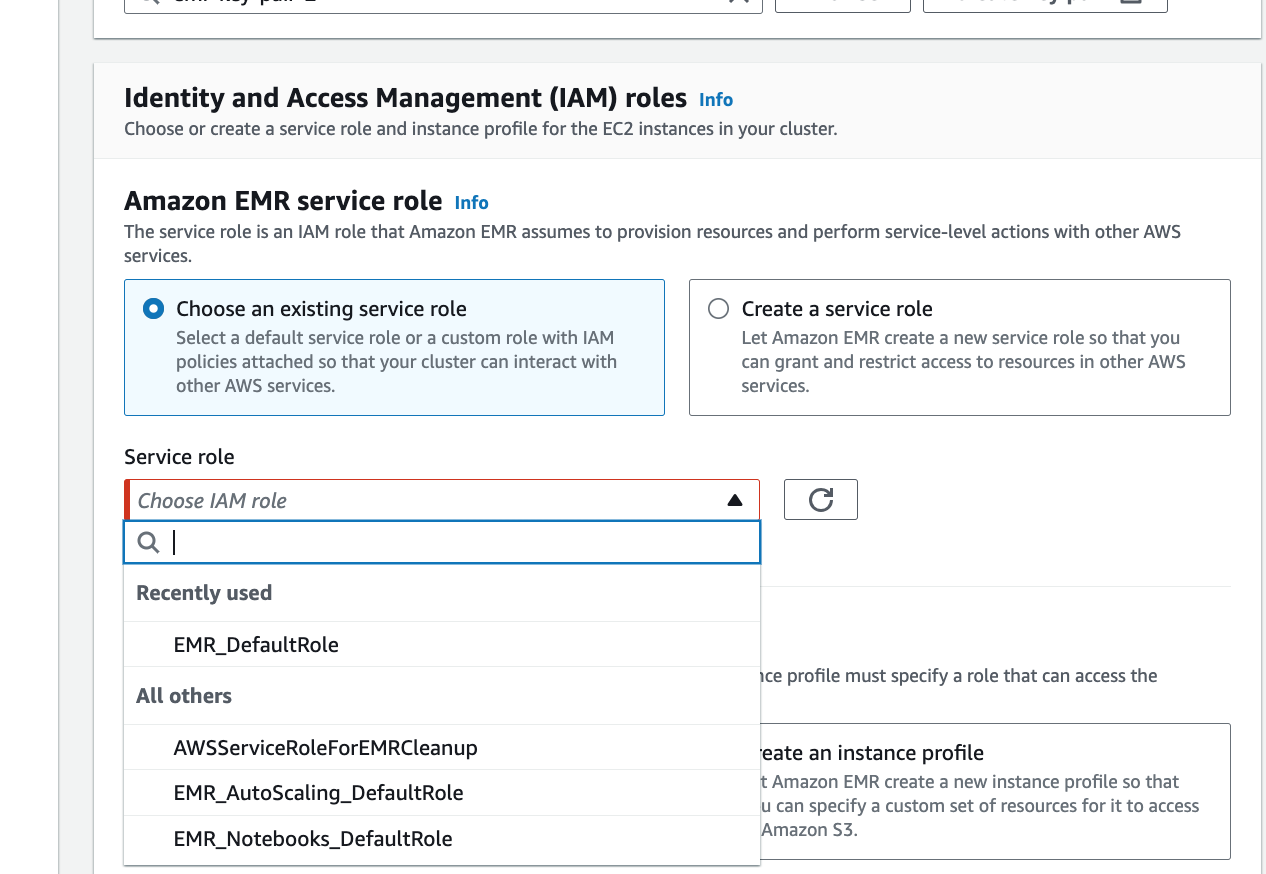
Go to the “Amazon EC2 key pair for SSH to the cluster”. Click on “Browse” and then select the key pair you created previously.

1. Now scroll to the **Amazon EMR service roll** section where you should see the following:



Under “Service role” click on the downward arrow to select an IAM role You should see the following note, your drop down menu may appear slightly different).

**Note: if you do not see the selection “EMR\_DefaultRole” skip from this step to step 10.**

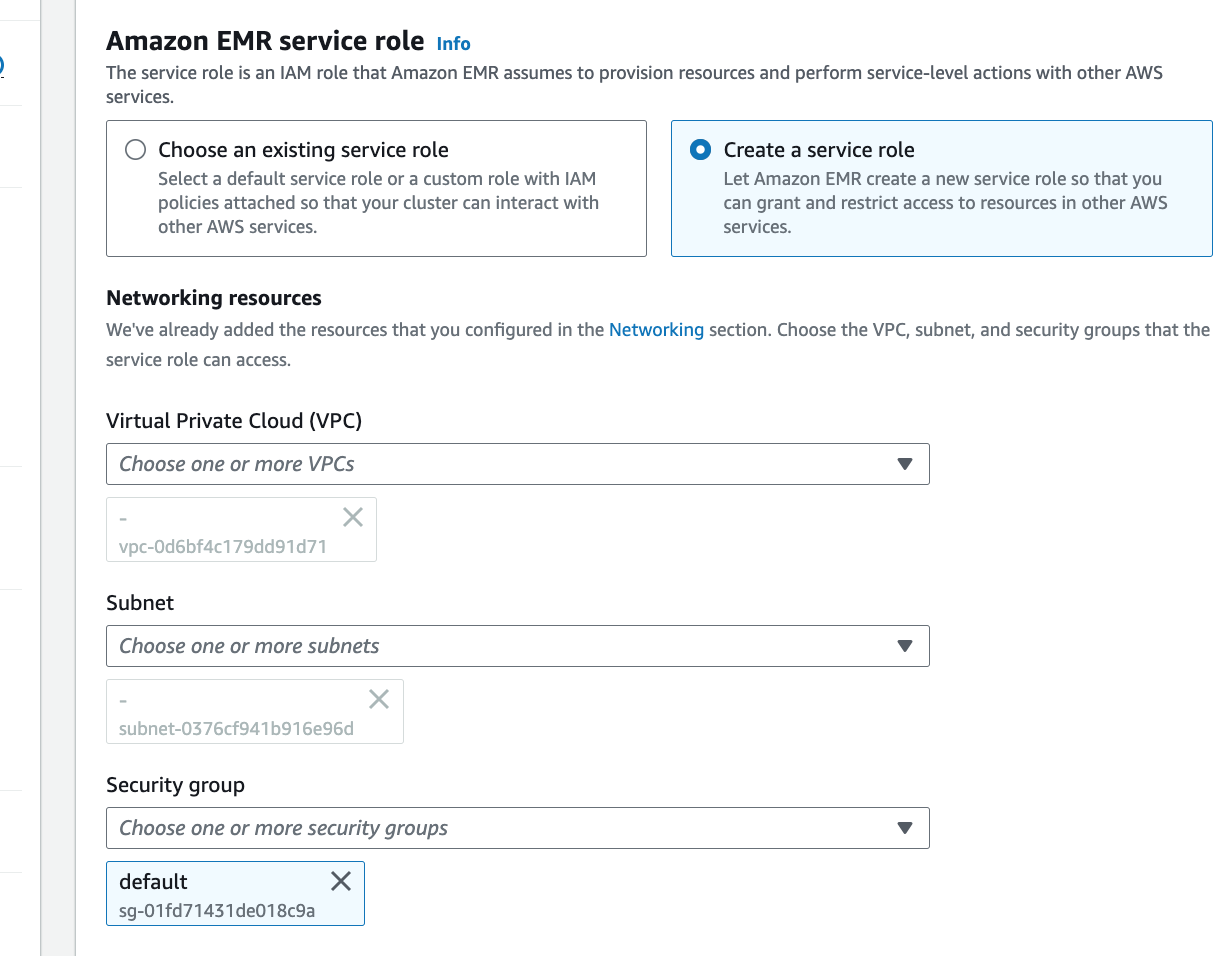


From the dropdown menu select “EMR\_DefaultRole”.

Now skip to step 11.

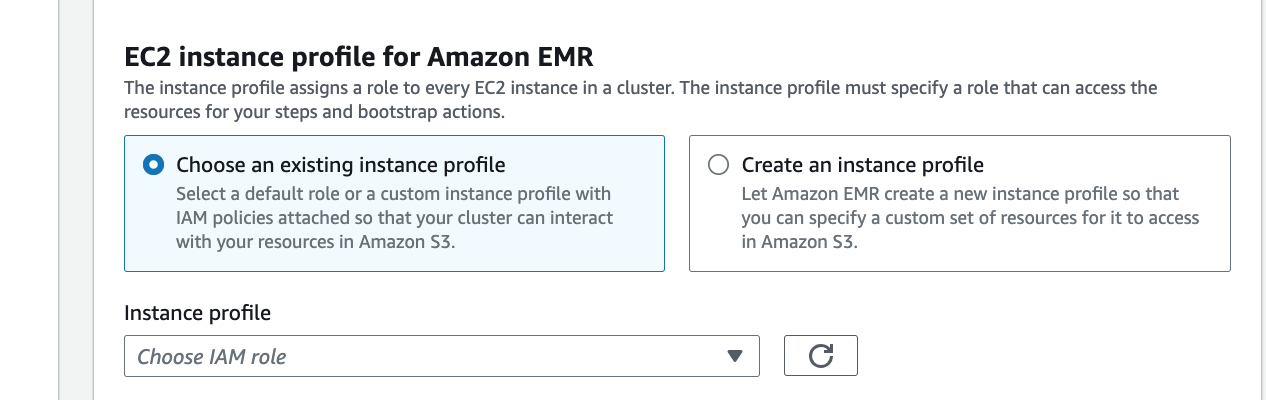
1. In the **Amazon EMR service roll** section do the following.

Select “Create a service roll as follows:



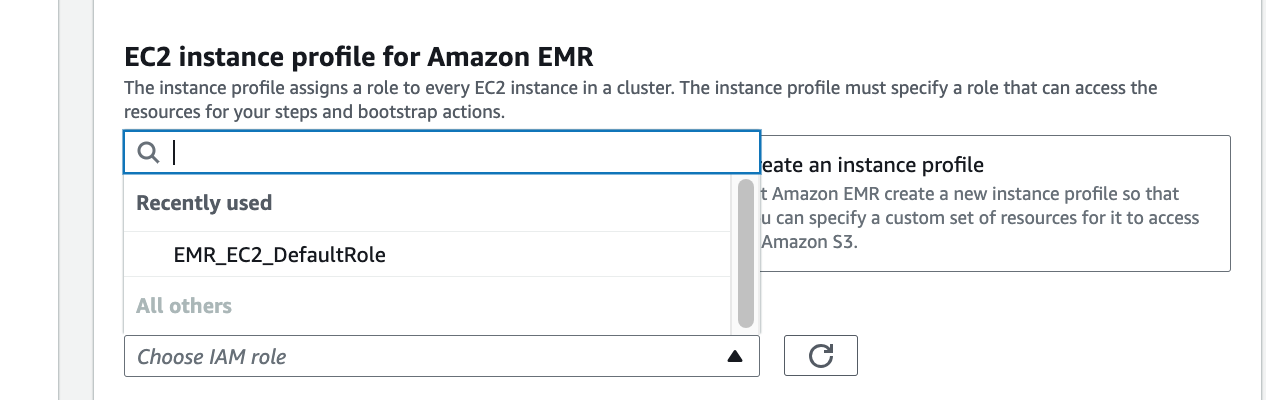
Leave the selections for “Virtual Private Cloud (VPC)”, “Subnet”, and Security group” as is and continue to the next step.

1. Now scroll down to **EC2 instance profile for Amazon EMR** where you should see the following:



Under “Instance profile” click on the downward arrow to select an IAM role You should see the following note, your drop down menu may appear slightly different):

**Note: if you do not see the selection “EMR\_EC2\_DefaultRole” skip from this step to step 12.**

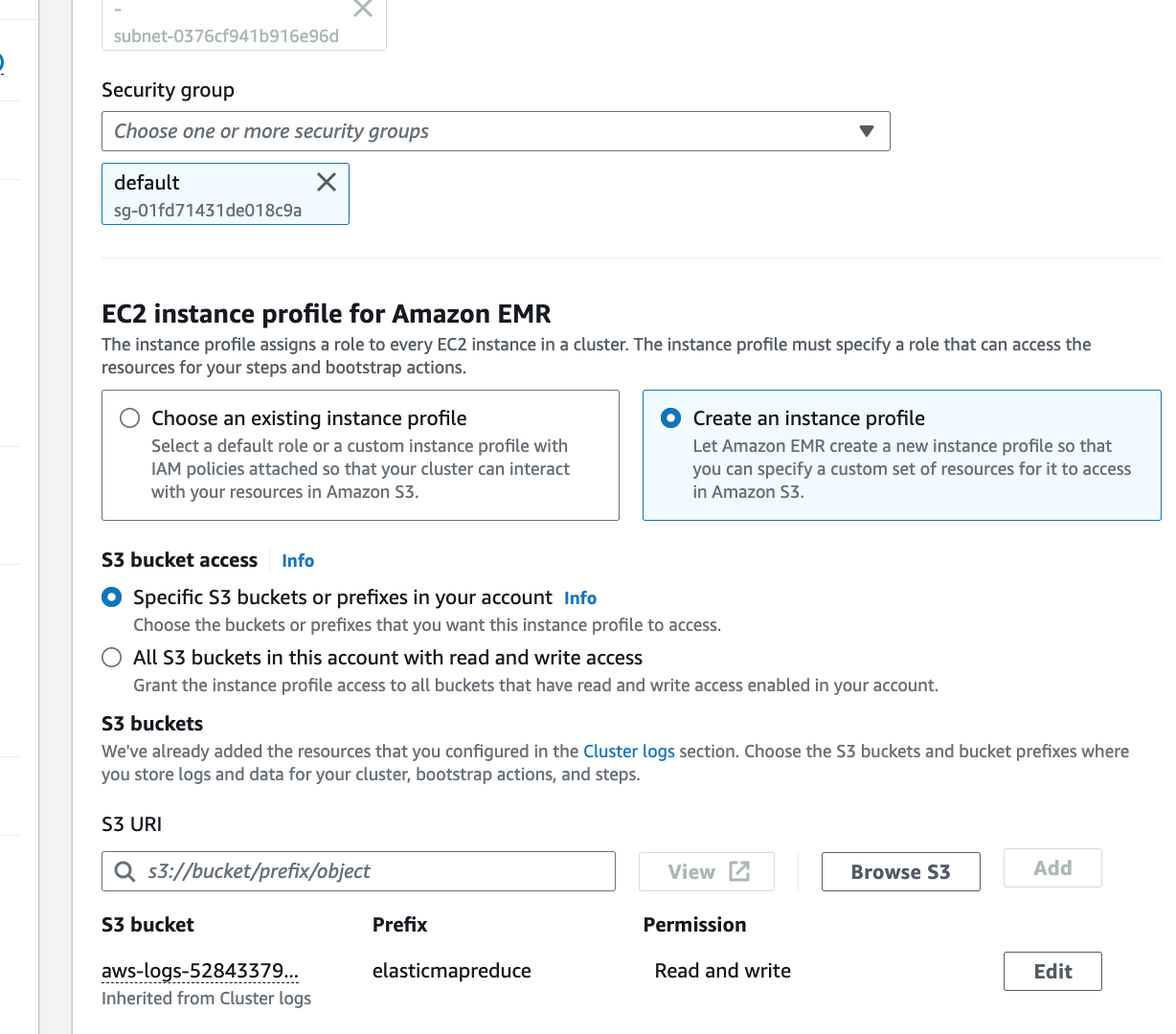


From the dropdown menu select “EMR\_EC2\_DefaultRole”.

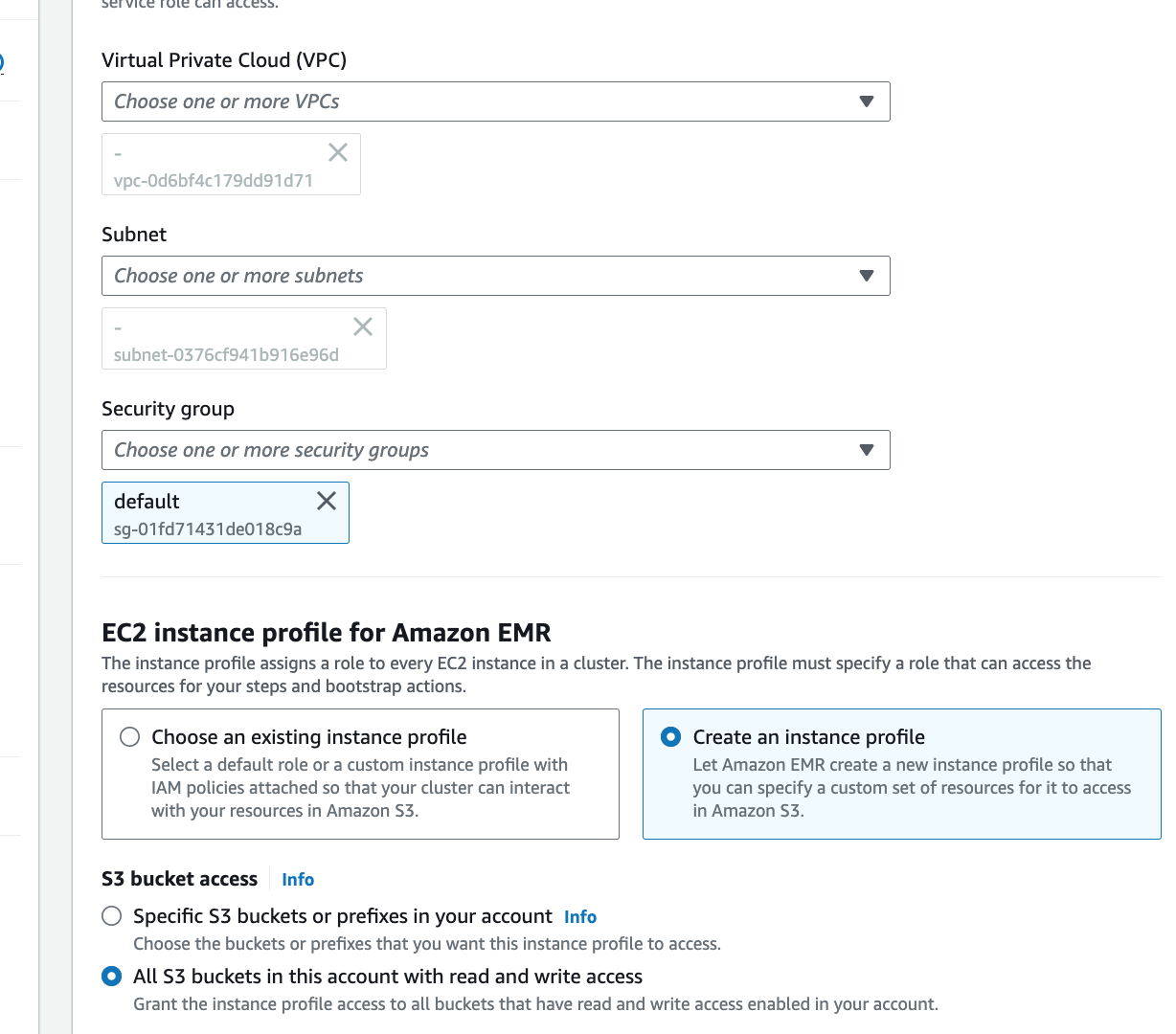
Skip to step 13.

1. **In EC2 instance profile for Amazon EMR** section you should do the following:

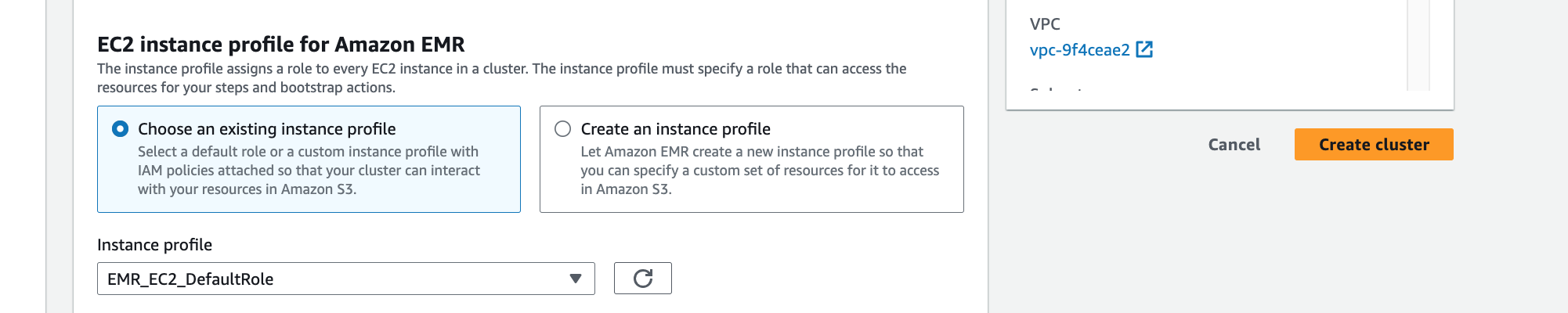
Select “Create an instance profile” as shown below.

or prefixes in you

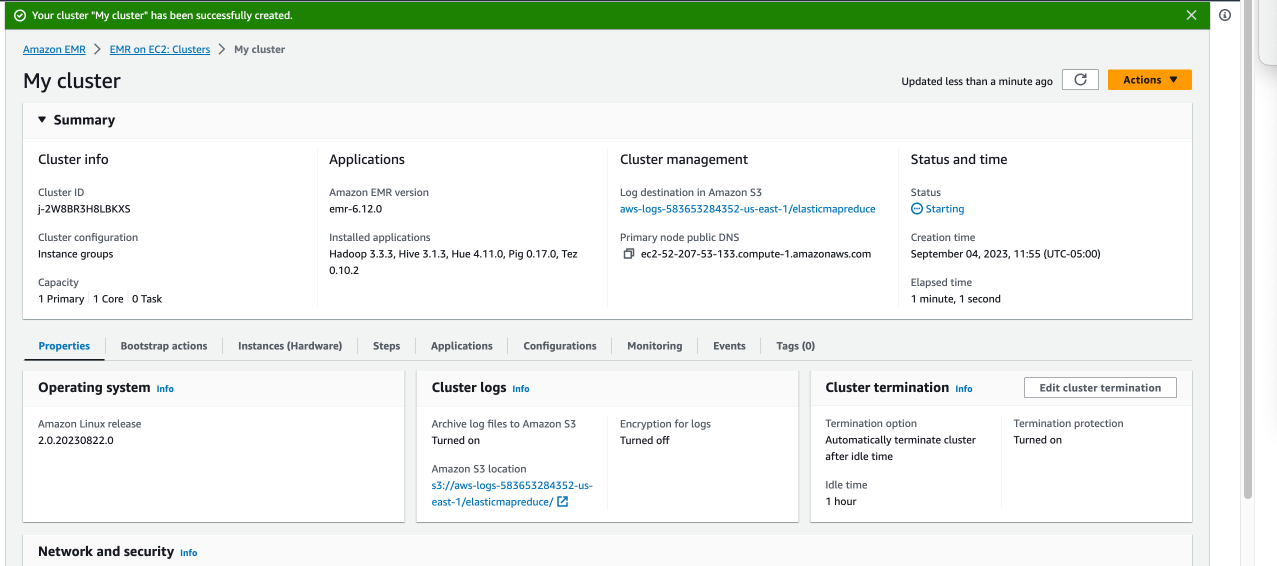
Now select “All S3 buckets in the account with read and write access” as shown below:



1. Now select “Create cluster” :

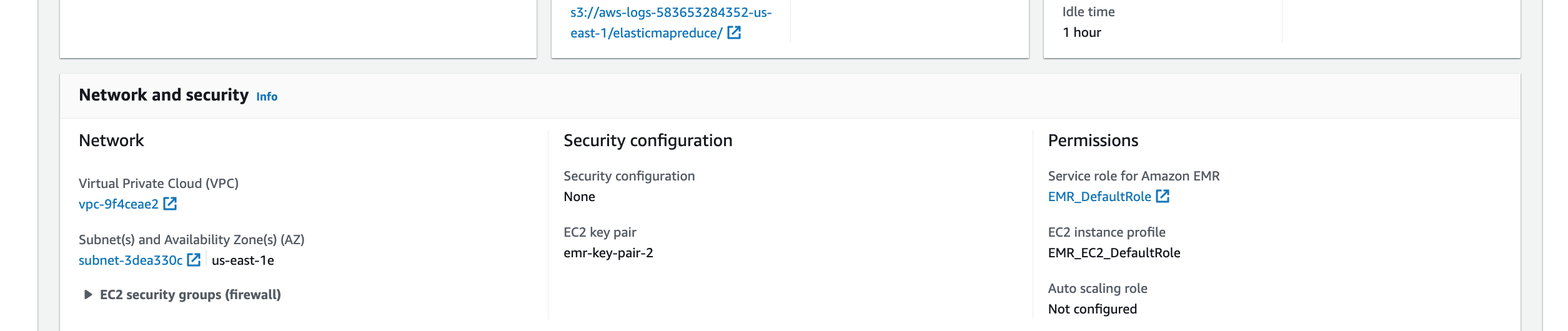


1. Now the cluster startup page should appear and look something like the following:

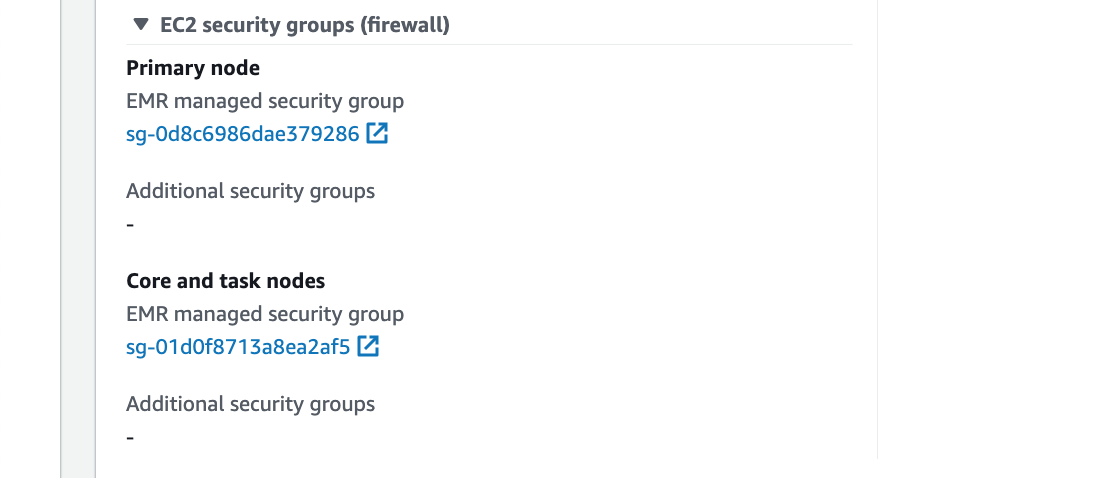


Note your cluster is ready for use when, instead of “Starting” it says “Waiting” This could sometimes take 10+ minutes, so don’t worry. As cluster creation tasks finish, items on the status page update. You may need to choose the refresh icon (circular arrow) on the right or refresh your browser to receive updates.

1. After the status of the cluster becomes “Waiting” scroll down to the Network and security section of the **My Cluster** page and you should see something like



1. Select the arrow next to **EC2 security groups (firewall)** to expand this section. As indicated below, under **Primary node**, select the security group link. This opens the EC2 console.



1. Choose the **Inbound rules** tab and then choose **Edit inbound rules**.
2. Check for an inbound rule that allows public access with the following settings. If it exists, choose **Delete** to remove it.
   1. **Type**

SSH

* 1. **Port**

22

* 1. **Source**

Custom 0.0.0.0/0

1. Scroll to the bottom of the list of rules and choose **Add Rule**.
2. For **Type**, select **SSH**. This selection automatically enters **TCP** for **Protocol** and **22** for **Port Range**.
3. For source, select **My IP** to automatically add your IP address as the source address. You can also add a range of **Custom** trusted client IP addresses, or create additional rules for other clients. Many network environments dynamically allocate IP addresses, so you might need to update your IP addresses for trusted clients in the future.
4. Choose **Save rules** at the bottom right of the page.

Note, once you have set up this rule, in most cases when you create a new cluster, it will use the same security group, so you likely will not need to set up this rule again. But it always is good to check.

**Step 6: Connect to the Primary Node Using SSH**

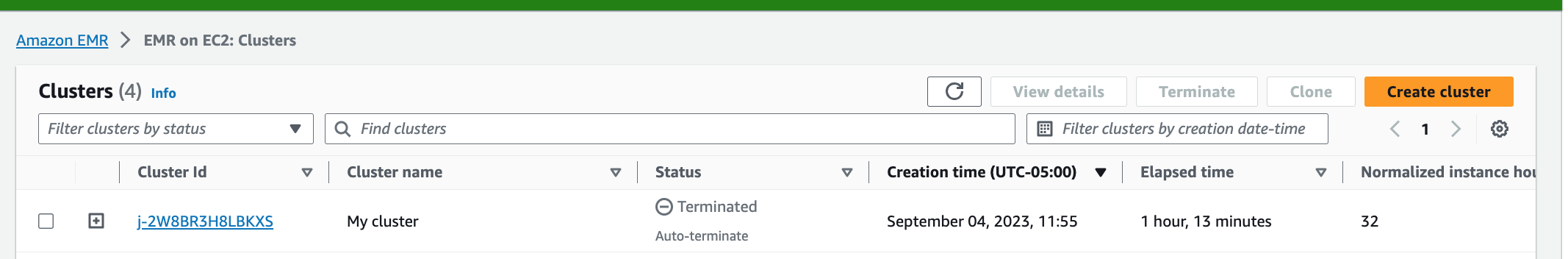
Secure Shell (SSH) is a network protocol you can use to create a secure connection to a remote computer. After you make a connection, the terminal on your local computer behaves as if it is running on the remote computer. Commands you issue locally run on the remote computer, and the command output from the remote computer appears in your terminal window.

When you use SSH with AWS, you are connecting to an EC2 instance, which is a virtual server running in the cloud. When working with Amazon EMR, the most common use of SSH is to connect to the EC2 instance that is acting as the primary node of the cluster.

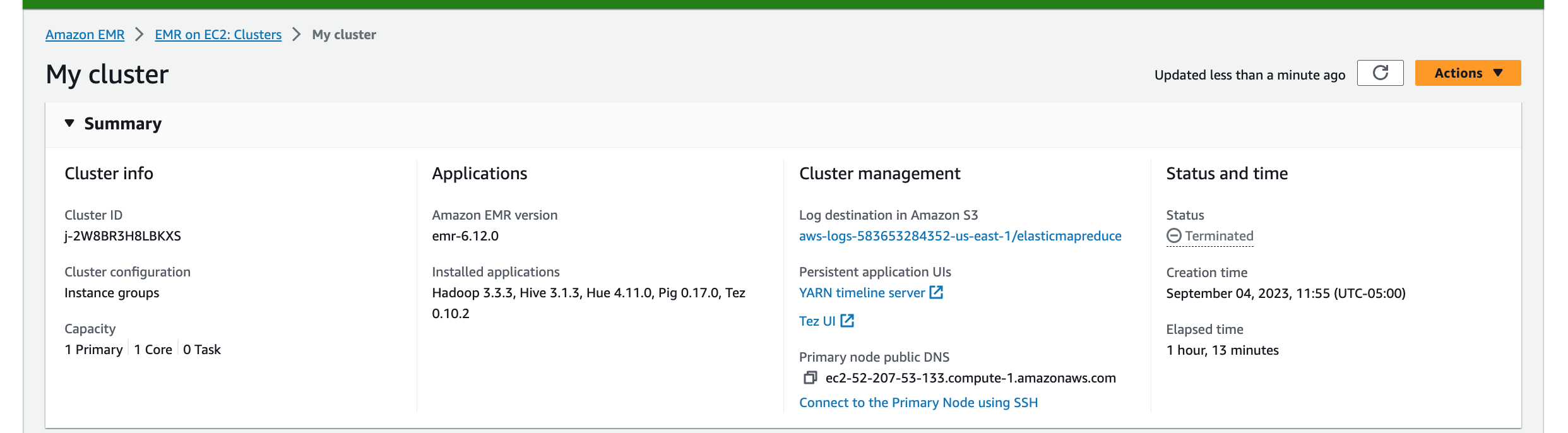
Using SSH to connect to the primary node gives you the ability to monitor and interact with the cluster. You can issue Linux commands on the primary node, run applications such as Hive and Pig interactively, browse directories, read log files, and so on.

To connect to the primary node using SSH, you need the public DNS name of the primary node. In addition, the security group associated with the primary node must have an inbound rule that allows SSH (TCP port 22) traffic from a source that includes the client where the SSH connection originates (something you did above).

1. To retrieve the public DNS name of the primary node using the Amazon EMR console. Find the EMR service page by typing EMR into the **Search** box and selecting EMR. Next click on the Cluster id of your cluster



1. Note the **Primary node public DNS** value that appears on the **My Cluster** page:



1. To connect to the Primary EMR node Using SSH and an Amazon EC2 Private Key, open a terminal window the MAC or use the bash utility on the PC.
2. To establish a connection to the primary node, type the following command.
   1. Replace *ec2-###-##-##-###.compute-1.amazonaws.com* with the primary public DNS name of your cluster
   2. Replace */<path-to-file>/mykeypair.pem* with the path (on your PC/Mac) and file name of your .pem file.

For MACOS or Linux, something like:

ssh -i /path/to/emr-key-pair.pem hadoop@*ec2-###-##-##-###.compute-1.amazonaws.com*

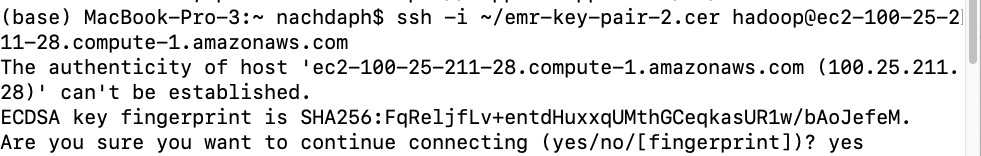
For Windows, something like;

ssh -i c:/path/to/emr-key-pair.pem hadoop@*ec2-###-##-##-###.compute-1.amazonaws.com*

**Important**

You must use the login name hadoop when you connect to the Amazon EMR primary node; otherwise, you may see an error similar to Server refused our key.

1. When you enter this properly you should see something similar to:



1. You might see a waring. The warning states that the authenticity of the host you are connecting to cannot be verified. If needed, type yes to continue.
2. When you are done working on the primary node (as you might be at the end of an assignment), type the following command to close the SSH connection.

exit

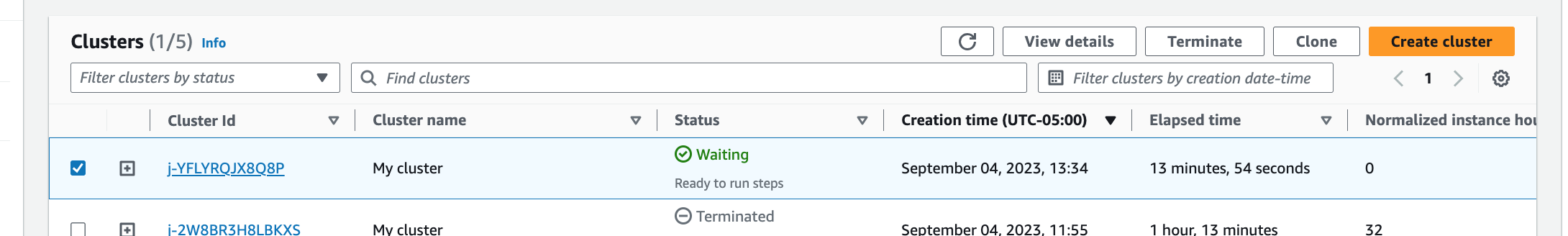
# Step 7: Terminate the Cluster and Delete the Bucket

After you complete your homework assignment or other project work, you may want to terminate your cluster and delete your Amazon S3 bucket to avoid additional charges.

Terminating your cluster terminates the associated Amazon EC2 instances and stops the accrual of Amazon EMR charges. Amazon EMR preserves metadata information about completed clusters for your reference, at no charge, for two months. The console does not provide a way to delete terminated clusters so that they aren't viewable in the console. Terminated clusters are removed from the cluster when the metadata is removed.

**To terminate the cluster**

1. Find the EMR service
2. Choose **Clusters**, then choose your cluster, then select Terminate



**To delete the cluster logging output bucket**

1. Find the S3 service
2. Choose the EMR bucket from the list, so that the whole bucket row is selected.
3. Choose delete bucket, type the name of the bucket, and then click **Confirm**.