Creating a Spark EMR Cluster

1. Upload files to S3

1. Services → S3
2. Create bucket
3. Give the bucket a name, and make it **PUBLIC**
4. Click in the name of your bucket and select Upload
5. Upload the files needed for this lab

2. Launch an Amazon EMR Cluster

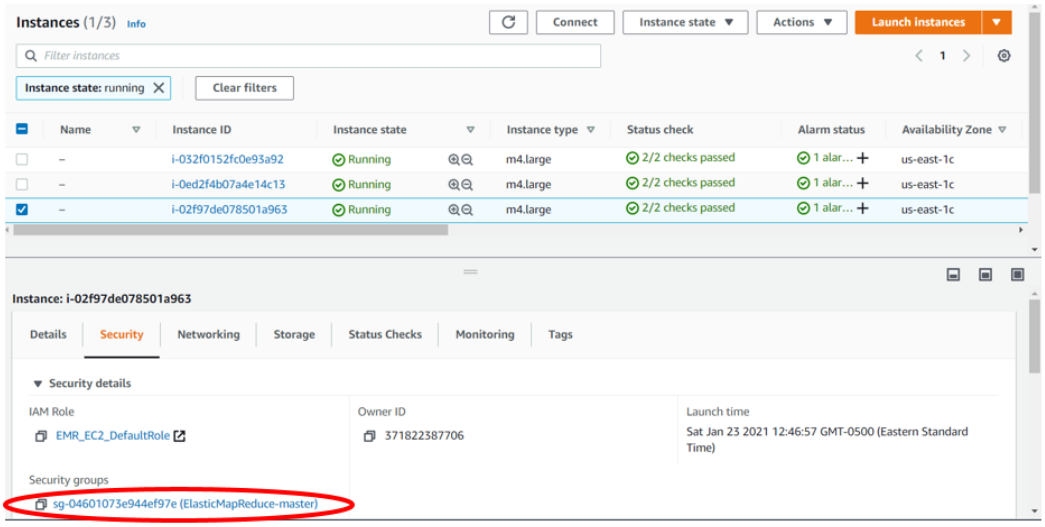
1. Services → EMR
2. Click **Create cluster**
3. In the **Name and applications** section, configure the following:
    a. Cluster Name: give your cluster a name, such as, SparkCluster
    b. Application bundle: select **Spark Interactive**
4. In the **Cluster configuration** section, configure:
        i. Under Uniform instance groups, select: Primary: **m4.large**, Core: **m4.large**, Task -1: **m4.large**
5. In the **Security configuration and EC2 key pair** section, configure:
    a. Amazon EC2 key pair for SSH to the cluster: select the name of the key pair file you created
6. In the **Identity and Access Management (IAM) roles** section, configure:
    a. Amazon EMR service role: under Service role, select EMR_DefaultRole
    b. EC2 instance profile for Amazon EMR: under Instance profile, select EMR_EC2_DefaultRole
7. Click **Create cluster** to launch your EMR Cluster.

The cluster will take approximately **fifteen minutes to launch**. The cluster will go through Starting, Bootstrapping, and Running states before the status changes to WAITING. Your cluster will be ready once the status changes to **WAITING**.

3. Allow your computer to connect to you EMR Cluster (Only if you have a new computer than the Hadoop lab)

1. Services → EC2
2. Click on **Instances (running)**
3. Find your Master:
    a. Click in any of the checkboxes of the instances you see running. You should only check ONE checkbox at the time.
    b. Once you have clicked in the checkbox, in the bottom portion, click on the Security tab.
    c. Under the Security Groups, you will see a description that says either (ElasticMapReduce-slave) or (ElasticMapReduce-master). If the instance you are checking says (ElasticMapReduce-slave), continue the process of checking instance

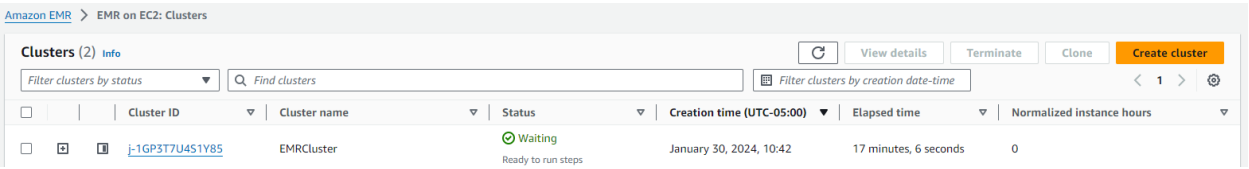
by instance until you find the one that says: (ElasticMapReduce-master). Here is an example of what you will see once you find the master instance.



d. Once you have found the master instance, click where it says ElasticMapReduce-master
e. Click on Edit inbound rules
f. Scroll down to the bottom of the screen, and click on Add rule.
g. Select: **SSH for Type** and **Anywhere-IPv4 for Source**. You need to modify **BOTH** values. Leave the rest of the options in the default.
h. Select Save Rules

4. Connect to your Leader (Master)

1. Services → EMR. You will see your cluster, similarly to this:

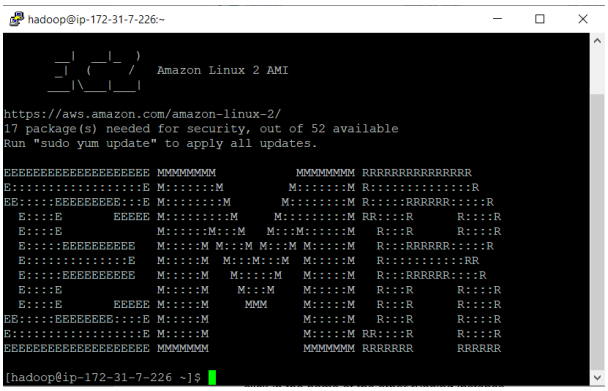


2. Click on the Cluster ID
3. Click on Connect to the Primary Node Using SSH, click on Windows or Mac/Linux (depending on your operating system) and follow those instructions to connect to your EMR cluster. Ignore the warning message (Click Accept) when connecting to the master node.

- **<u>If you are on</u> <u>Mac</u>**, before you type ssh -i… you have to do an additional step:
    1. Using the terminal, navigate to the folder where you have the .pem file (for example: cd desktop), and then type: chmod 400 your_pem_file.pem *[make sure you change the your_pem_file with your PEM file name]*

2. Now, you can type the command you see suggested. Make sure you replace the ~/your_pem_file.pem with the location and filename of your .pem file. For example, mine is: ssh -i /Users/andrea/Desktop/EMR_Mac.pem…

4. You have successfully connected to the Master node once you see this screen:



- Type pyspark. Congratulations! You are in the Spark environment now.