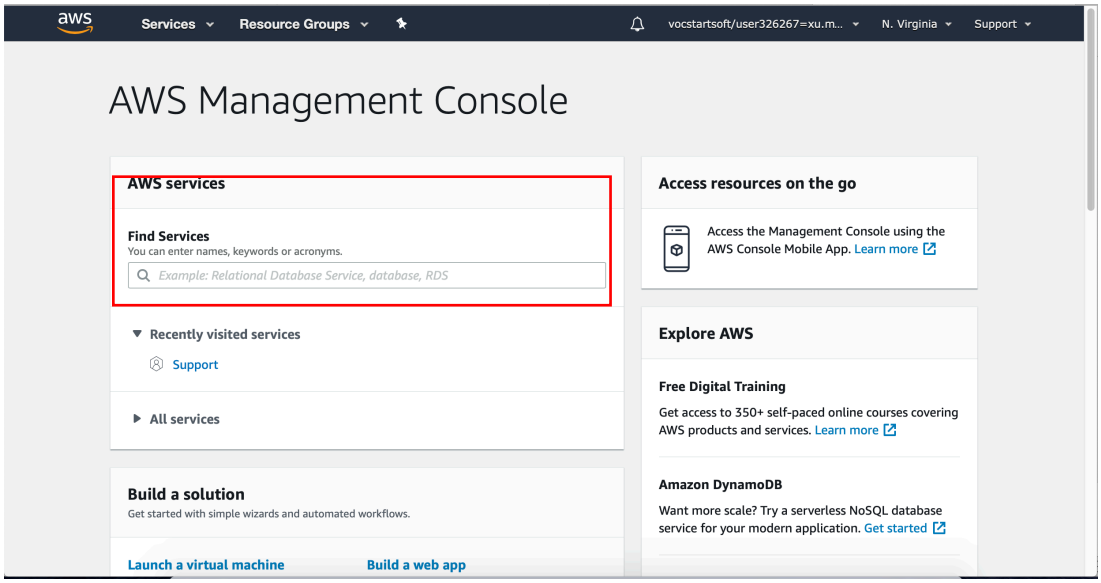
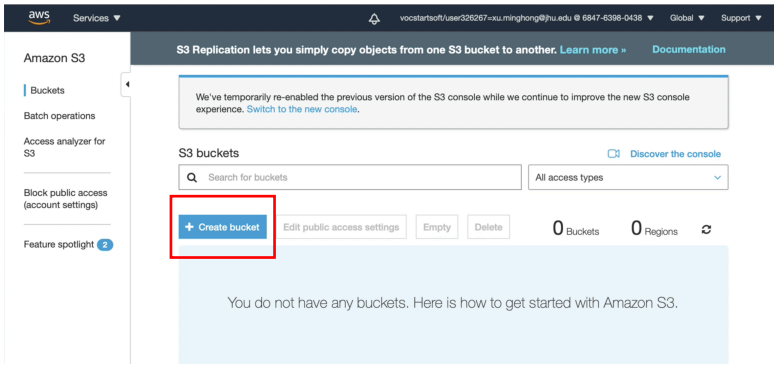
BU.330.740 Large Scale Computing with Hadoop
Lab 2. Set Up a Hadoop Cluster on Amazon AWS

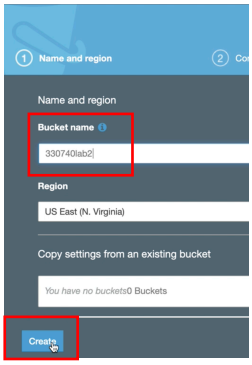
Learning Goal: practice using AWS to create a big data processing environment

1. Download scripts.zip from Blackboard, unzip and store them in a local folder. Login into your AWS Educate account, then go to **AWS Account** -> **AWS Management Console** In **Find Services**, search for S3 and enter/return.

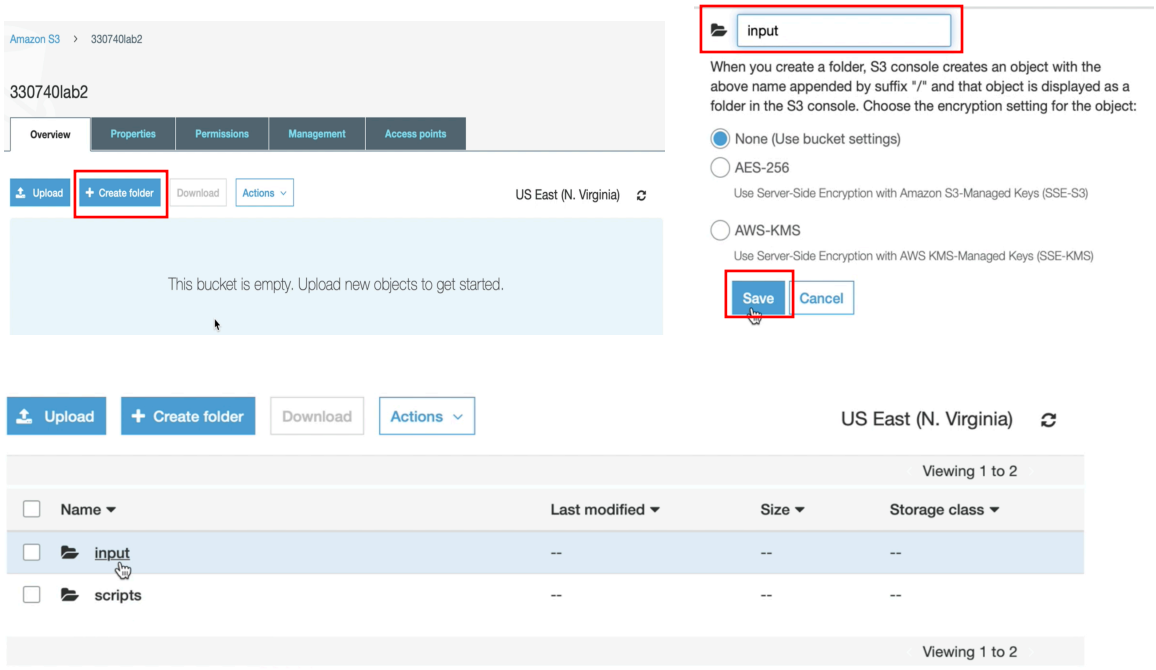


2. At S3 console, create an S3 bucket. You need to name your S3 bucket, and leave all other things unchanged. Because of Hadoop requirements, bucket and folder names that you use with Amazon EMR must contain only letters, numbers, periods (.), and hyphens (-). After you specify the name, do not go to the next steps but click **Create**. Note that your region may be different from mine.



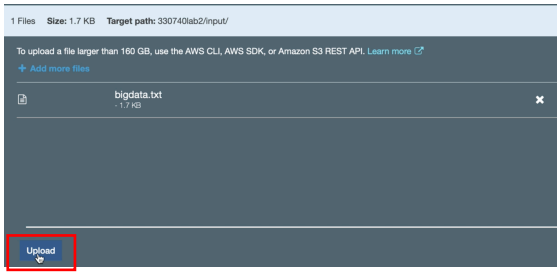


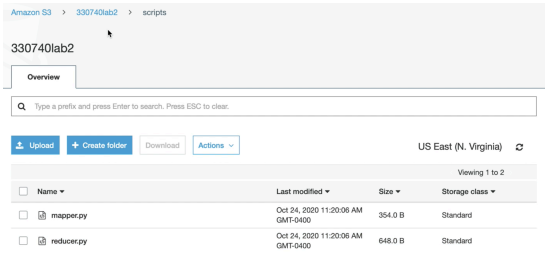
After you create the bucket, choose it from the list and then choose **Create folder**. Create two folders, input and scripts, one by one. After specify the name of the folder, do not change anything else, but choose **Save**.



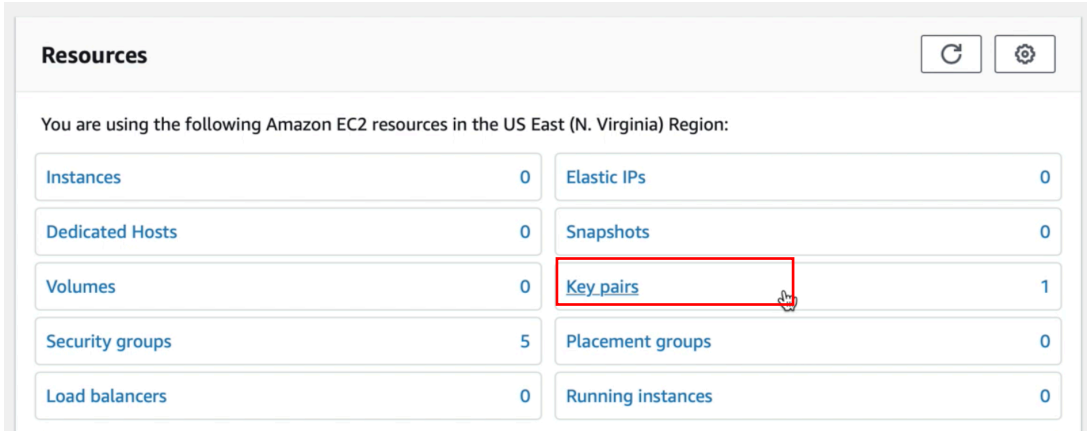
Then click input folder, choose 

in the upper left corner, add bigata.txt to the folder. Do not change other settings, but click **Upload** directly. Go back to your bucket, click scripts folder, and upload mapper.py and reducer.py to the folder.

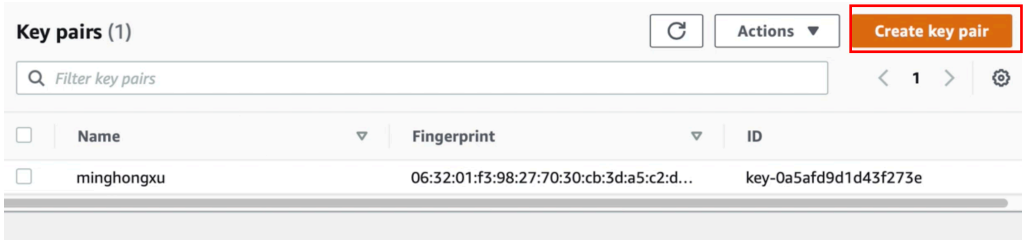




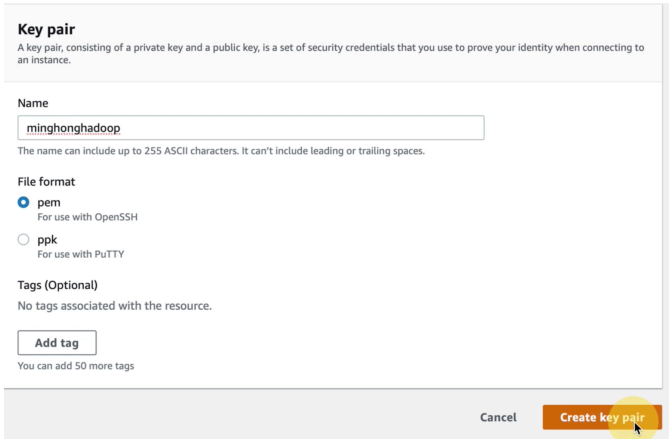
3. Create an EC2 Key Pair. Go back to **AWS Management Console**, search for EC2 to open the EC2 console. In the navigation pane, choose **Key Pairs**.



Choose **Create key pair**.

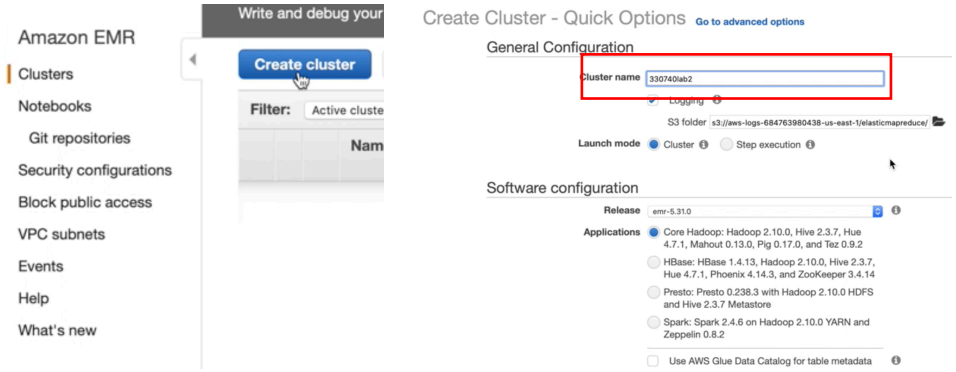


For **Name**, enter a descriptive name for the key pair. For **File format**, choose **pem**. Choose **Create key pair**.



The private key file is automatically downloaded by your browser. Save the private key file in a safe place.

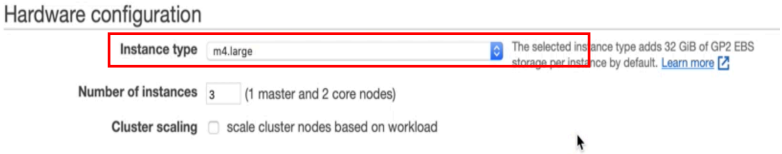
4. Launch the EMR Cluster. Go back to **AWS Management Console**, search for EMR to open the EMR console. Choose **Create cluster**.

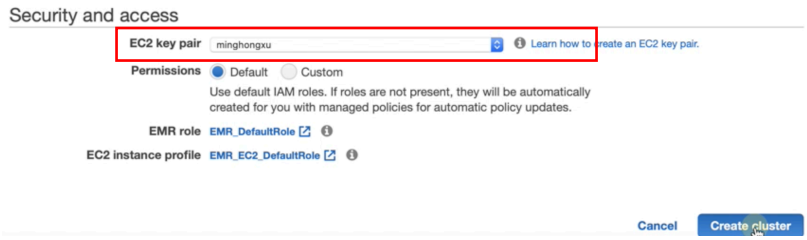


On the **Create Cluster - Quick Options** page, accept the default values except for the following fields:

a. Enter a **Cluster name** that helps you identify the cluster.
b. Under Hardware configuration, choose m4.large as the **Instance type**.
c. Under **Security and access**, choose the **EC2 key pair** that you created.

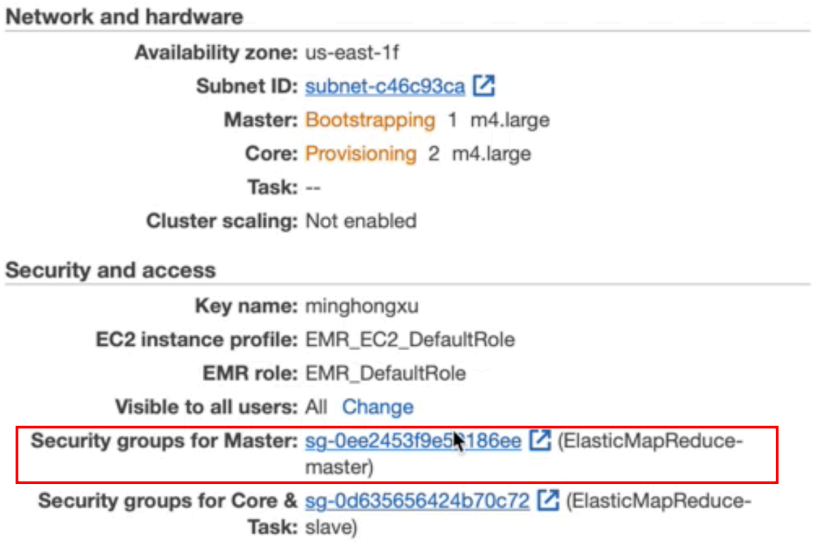
Choose **Create cluster**.



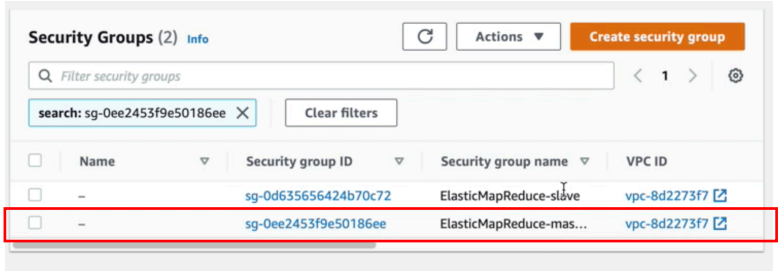


Now the cluster is starting, which means AWS will provision your cluster (find available nodes to fulfill your job request).

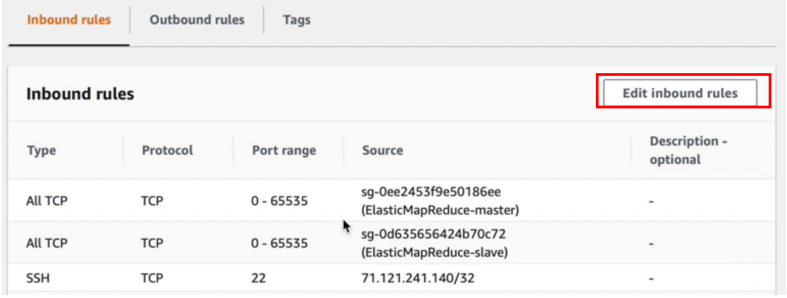
5. While you are waiting for the cluster to be ready, change the security group rules to allow SSH cnnections to the cluster from your own IP address. Under **Security and access** choose the **Security groups for Master** link.



Choose **ElasticMapReduce-master** first from the list.



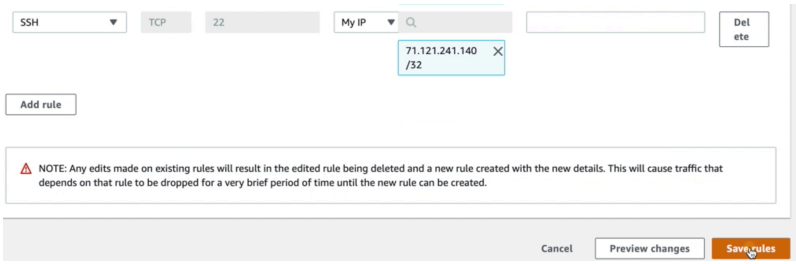
Scroll down to choose **Edit inbound rules**.



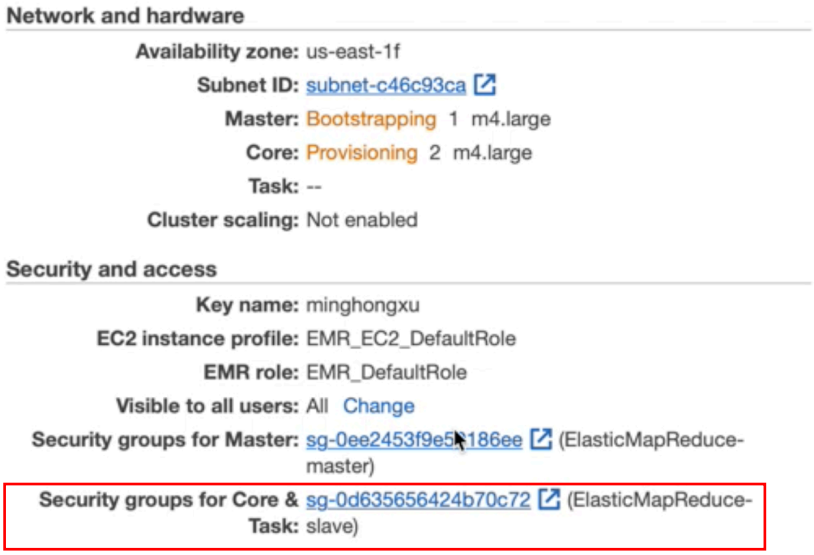
Check for an inbound rule that allows public access with the following settings. If it exists, choose **Delete** to remove it.

- **Type:** SSH
- **Port:** 22
- **Source:** Custom 0.0.0.0/0

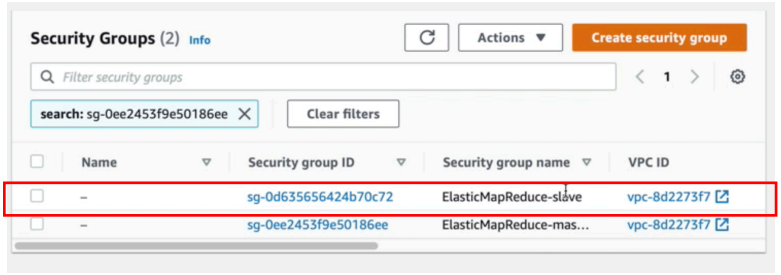
Scroll to the bottom of the list of rules and choose **Add Rule**. For **Type**, select **SSH**. This automatically enters **TCP** for **Protocol** and **22** for **Port Range**. For source, select **My IP**. Choose **Save**.



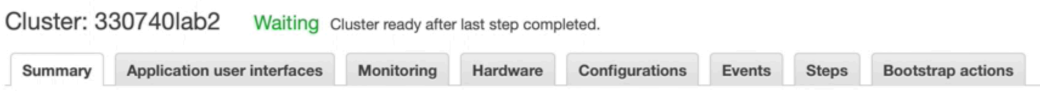
Go back to your EMR cluster and click **Security groups for Core & Task** link.



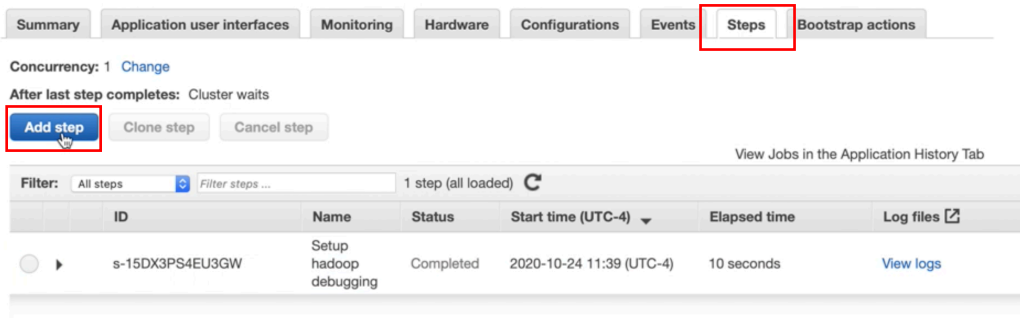
Choose **ElasticMapReduce-slave** from the list and repeat the steps above to allow SSH client access to core and task nodes from your own IP address.



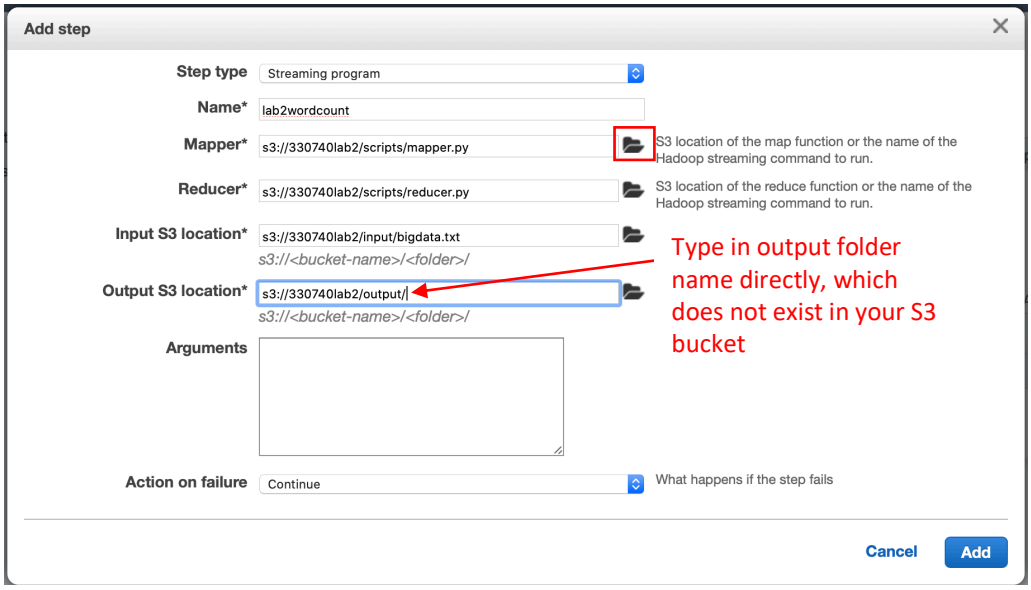
6. Under **Network and hardware**, find the **Master** and **Core** instance status. The status goes from **Provisioning** to **Bootstrapping** to **Waiting** during the cluster creation process.



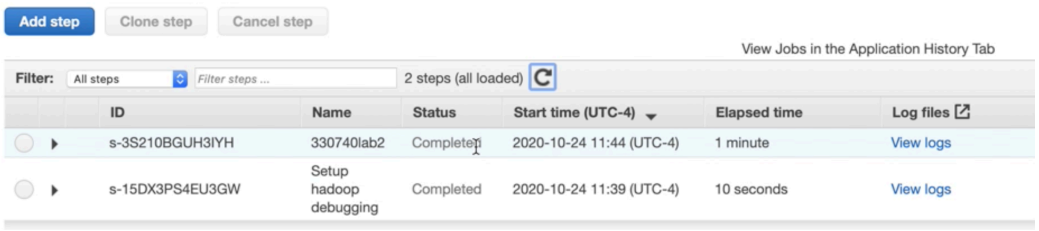
7. After the cluster is ready, click **Steps**, and then **Add step**.



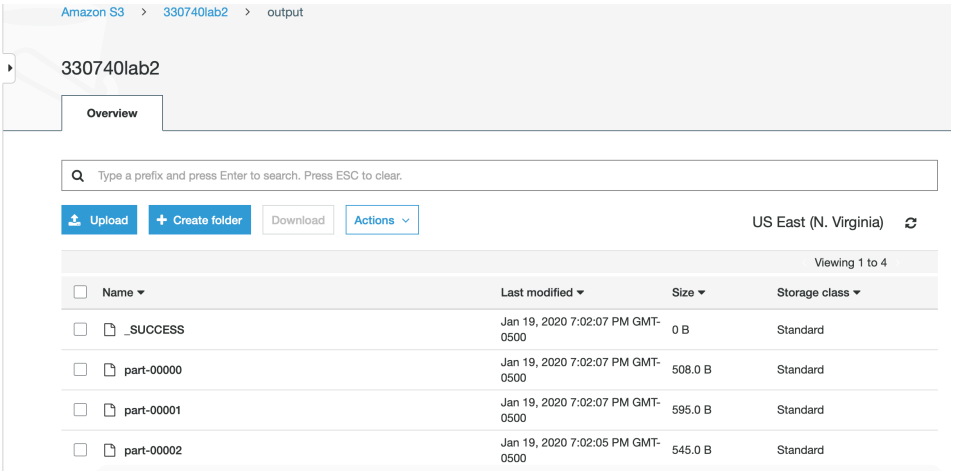
8. Choose **Streaming program** for **Step type**. **Name** your Streaming program; point **Mapper** to your mapper.py and **Reducer** to your reducer.py on your S3 instance; point **Input S3 location** to bigdata.txt; designate **Output S3 location** to a folder called output on your S3 instance. **Note that this output folder should not be created before**, and Hadoop will create it for you. You can use the folder icon (shown in red rectangle) to have access to folders on your own S3 instance. Click **Add** to execute the step.



9. Wait while the step is going from **Pending** to **Running** to **Complete**. You can always use icon on the pages to refresh the status!



10. After it's completed, you can check and download results from your S3 bucket -> output folder. Download these files to your local folder. And you can also check log files in your EMR cluster. Being able to read log files is a skill to acquire, especially if your task has a "failure" or "error" condition. You will get more practice as we go.



11. **DO NOT FORGET TO CLEAN UP RESOURCSES!!** Otherwise, you will incur unnecessary charges to your account.
    a. Terminate the cluster
    b. Delete your S3 bucket
    c. Delete the log files bucket (we are not coming back to this instance)

    Now that you are done with lab 2. Take some time to navigate your AWS Management Console. Double check that your S3 is empty, your EMR cluster is terminated, and no running instance in your EC2. And log out your AWS account every time you are done.

Reference:
https://aws.amazon.com/getting-started/projects/analyze-big-data/