Lab 5

In this lab, we will use MapReduce to analyze a dataset of 3-4 million tweets collected over the past weekend. This dataset specifically has English-language tweets from within the United States. You will need to have your AWS account set up (completed in lab 1) to complete this lab.

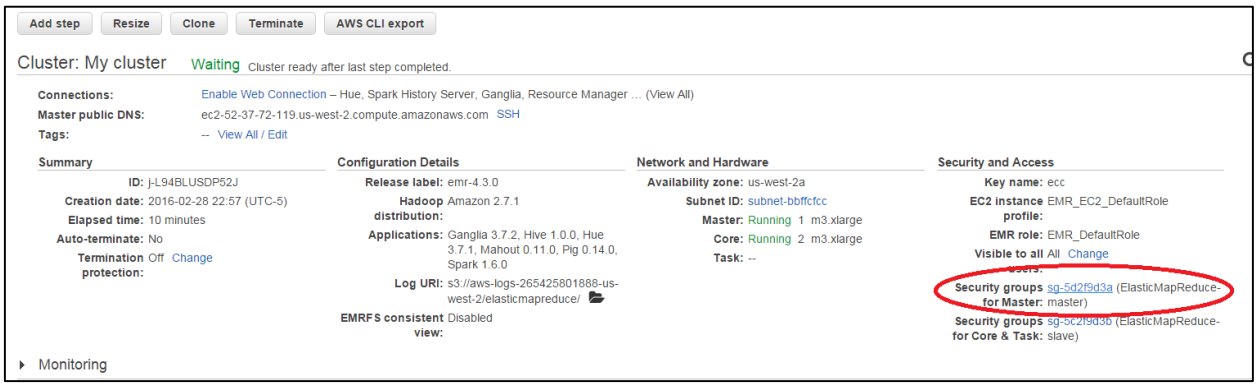
First, download the files **WordCount.java, top.py, and twitterdata.txt**. WordCount.java and top.py can be found in the Lab5 resources folder on NYU classes. The twitterdata.txt data can be found at http://s3.amazonaws.com/bigdataclassecc/twitterdata.txt

Upload the three files to your Amazon S3 bucket.

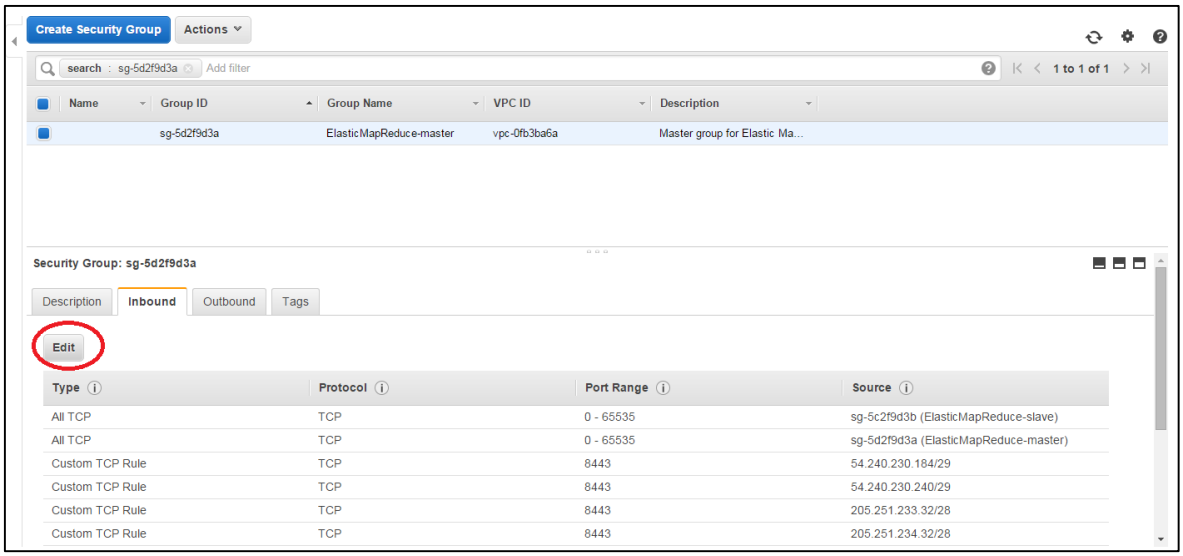
Task1: Setup EMR Cluster and run example JAR

Launch an EMR cluster using the public/private key that you generated in Lab1. Wait for the cluster to start up.

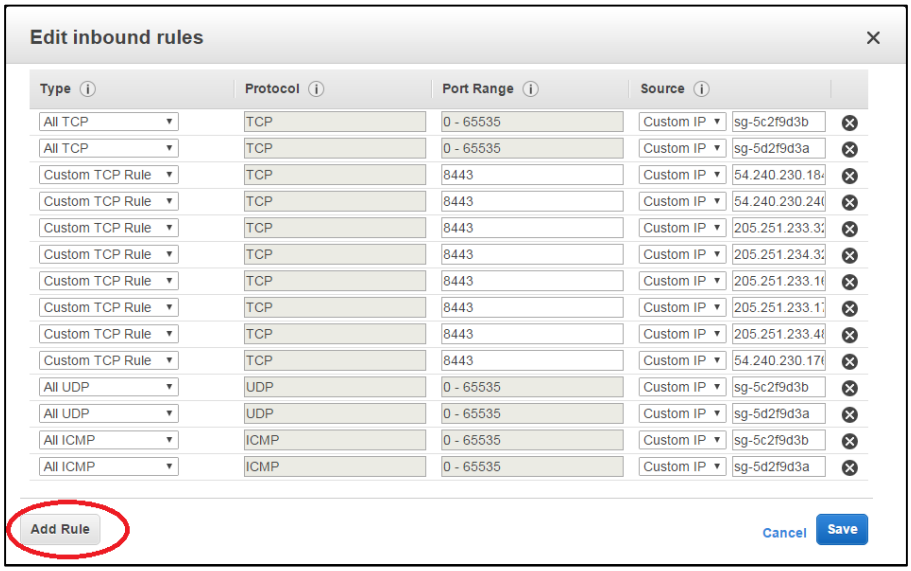
In order to connect to your master node via SSH, you will need to first modify your security group. To do this, go to the cluster you just created, and click on the blue link following "Security groups for Master".

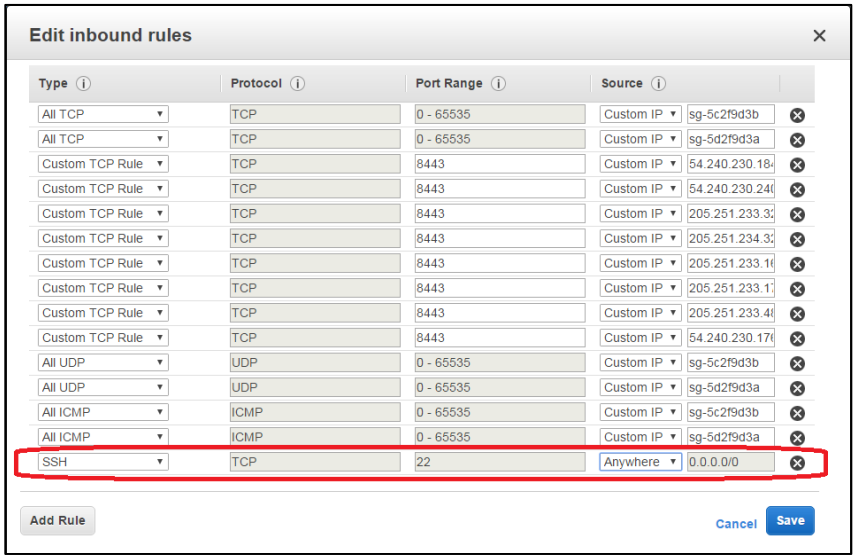


In the bottom pane, select the Inbound tab and click the Edit button.



Click Add Rule, and select SSH for Type and Anywhere for the Source. Click Save.





Now you are ready to connect to the master node using SSH. Follow the instructions here: http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-connect-master-node-ssh.html

Once you have a connection to the master, we will compile the java file and package a jar file.

Follow the instructions to build binaries using Amazon EMR: http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-build-binaries.html

The source files are WordCount.java, which can be found under Lab5 Resources on NYU Classes, and twitterdata.txt

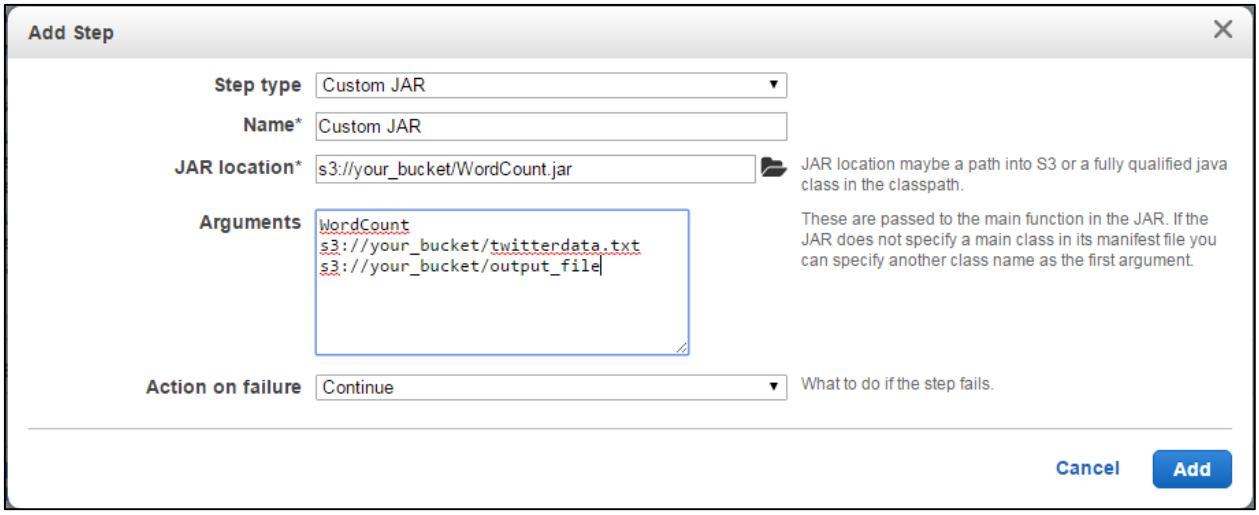
For the building binaries step, you should type

```
javac –cp $(hadoop classpath) WordCount.java
jar cvf WordCount.jar *.class
```

After building the binary, copy the jar file to your S3 bucket by typing

```
hadoop fs –put WordCount.jar s3://your_bucket/WordCount.jar
```

Now, go to EMR console and add a step as follows:



Once the step completes, you can view the output your S3 bucket in the folder your specified.

Task 2: Popular Hashtags

Modify the WordCount.java file such that the output lists hashtags that appear more than 100 times in the data, along with the number of times each distinct hashtag appears.

If you are unfamiliar with Java, you may find the following links helpful:
https://docs.oracle.com/javase/tutorial/java/nutsandbolts/if.html
https://docs.oracle.com/javase/7/docs/api/java/lang/String.html

You will then need to recompile and repackage the binaries on the EMR master, upload the JAR file to your S3 bucket, and add a step as in Task 1.

Task 3: Deliverable: Top 50 Hashtags

Download the python script top.py from NYU Classes and upload it to your S3 bucket.

Move the output file from your WordCount program you created in Task 2 to the EMR master

```
hadoop fs -get s3://your_bucket/output_folder/* output_folder
hadoop fs -get s3://your_bucket/top.py
```

To find the top (most-used) 50 hashtags using the data output by your Task 2, type

```
python top.py 50 output_folder/ top50.txt
```

This outputs the top 50 hashtags into a file called top50.txt. To view the contents of this file, type

```
cat top50.txt
```

Deliverable: Submit your top50.txt file to the Lab5 assignment link in NYU Classes. If you worked with a partner, you may both upload the same file. This must be submitted by Monday, March 7, 2016 at 12:00pm.

IMPORTANT: Remember to terminate your EMR cluster!