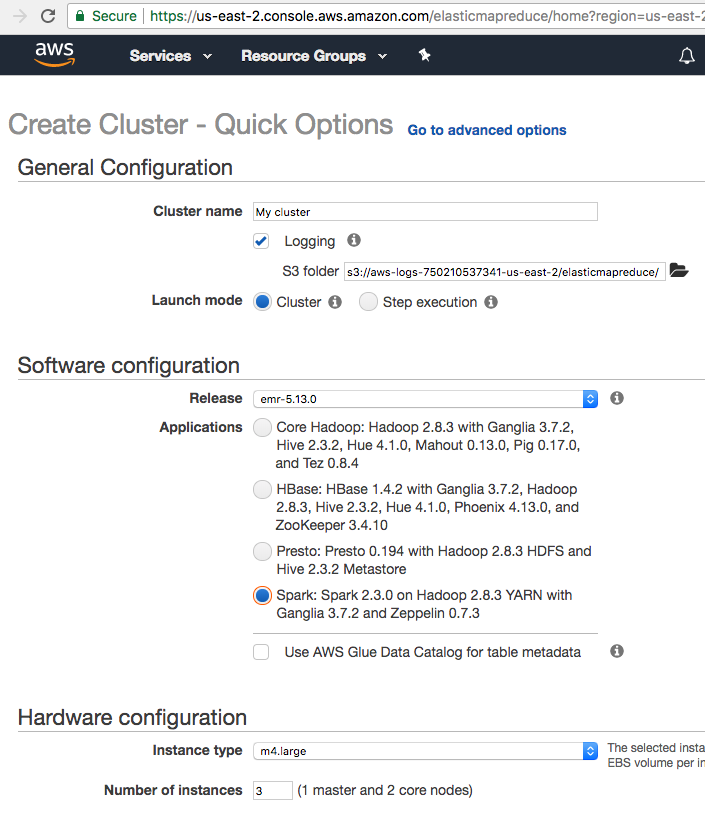
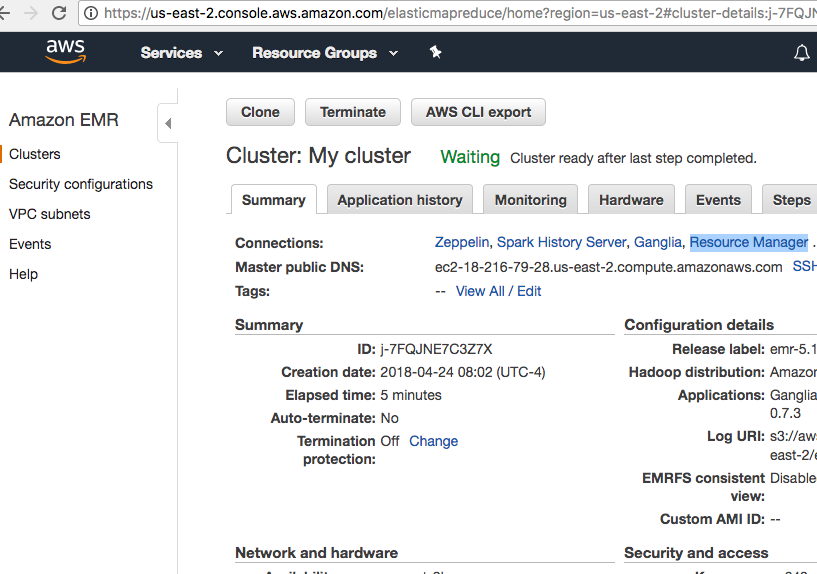
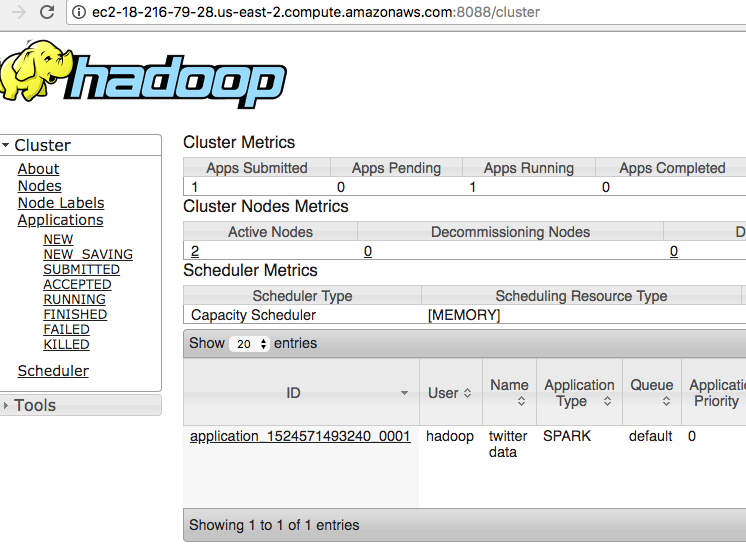
1. Create a spark cluster on EMR



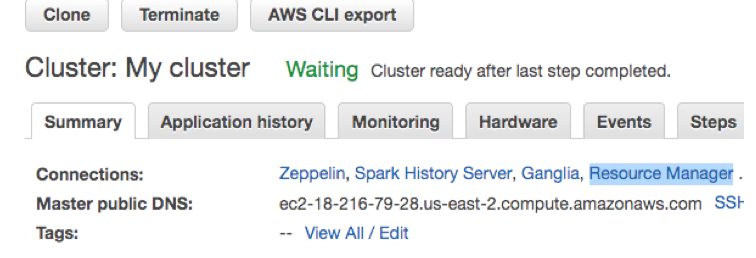
please choose your own .pem file, then launch the cluster. And you will see the following image:



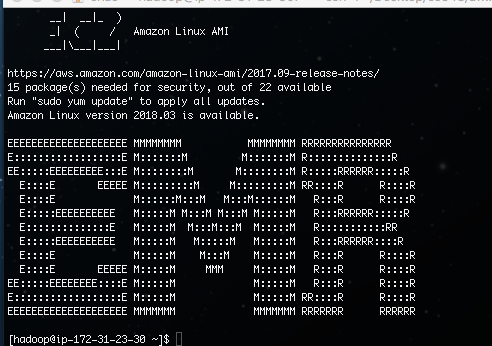
wait until you see ‘waiting’, then click ‘Resource Manager’ so you can see the job monitor. All your submitted jobs can be viewed here.



1. Open a shell, run SSH command. For command details, you can click the “SSH” link below the “resource manager” as showing below:



Once you successfully login in, you will see the following:



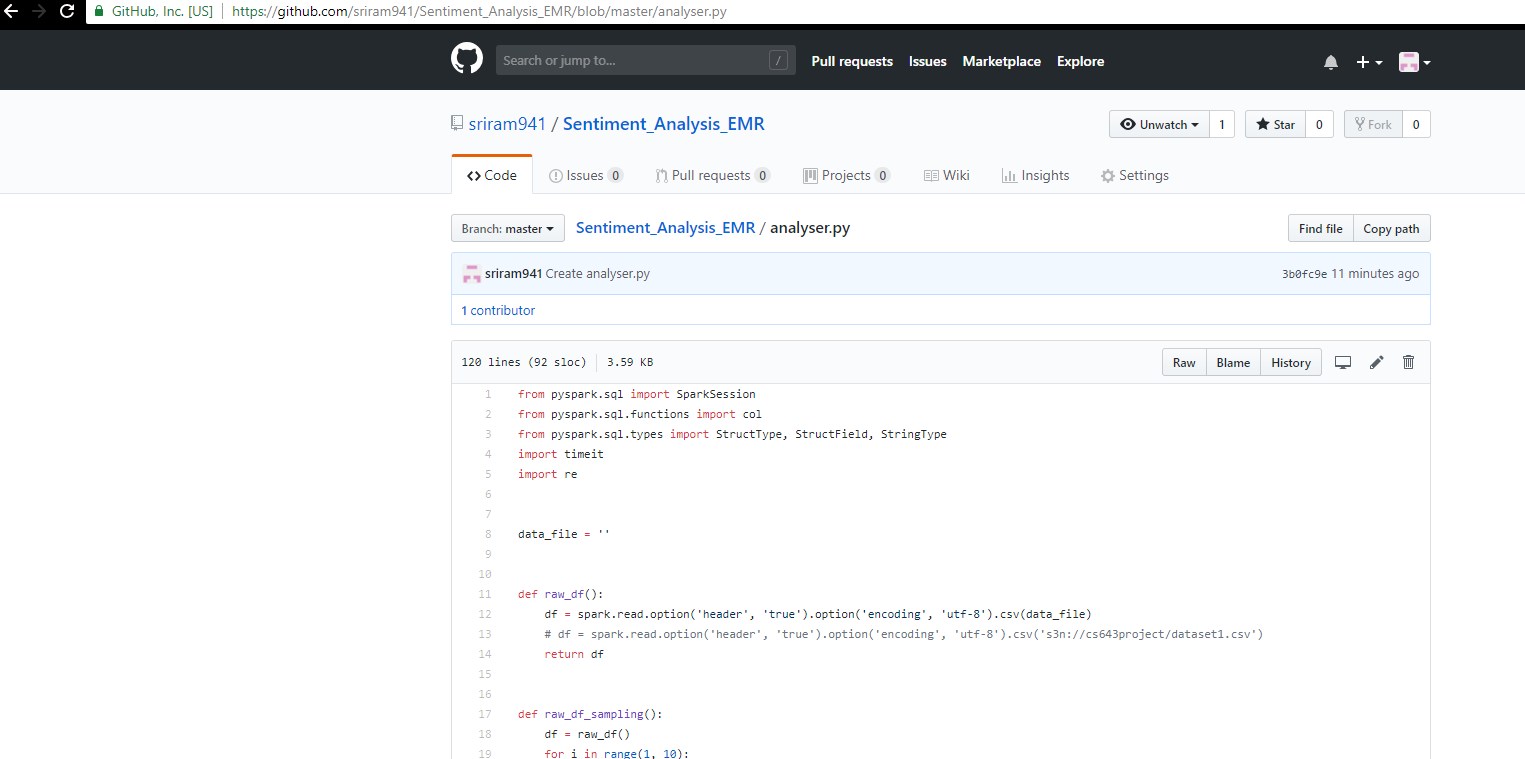
1. Run the jobs:

Please run these command to install git and download the codes from git directly:

sudo yum upgrade

sudo yum install git

git clone https://github.com/sriram941/Sentiment\_Analysis\_EMR.git



Then go to the ‘sentiment analysis’ folder and run the command:

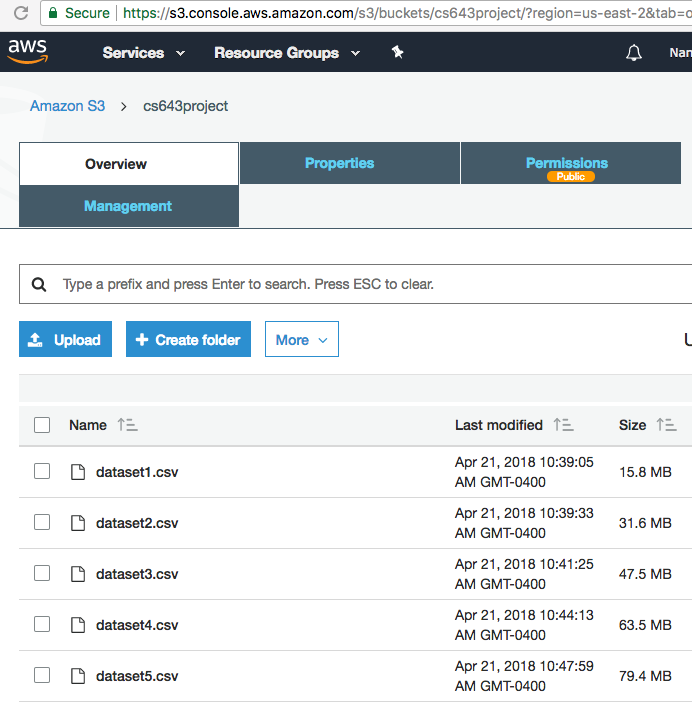
spark-submit analyzer.py --master yarn --executor-memory 12G --num-executors 8

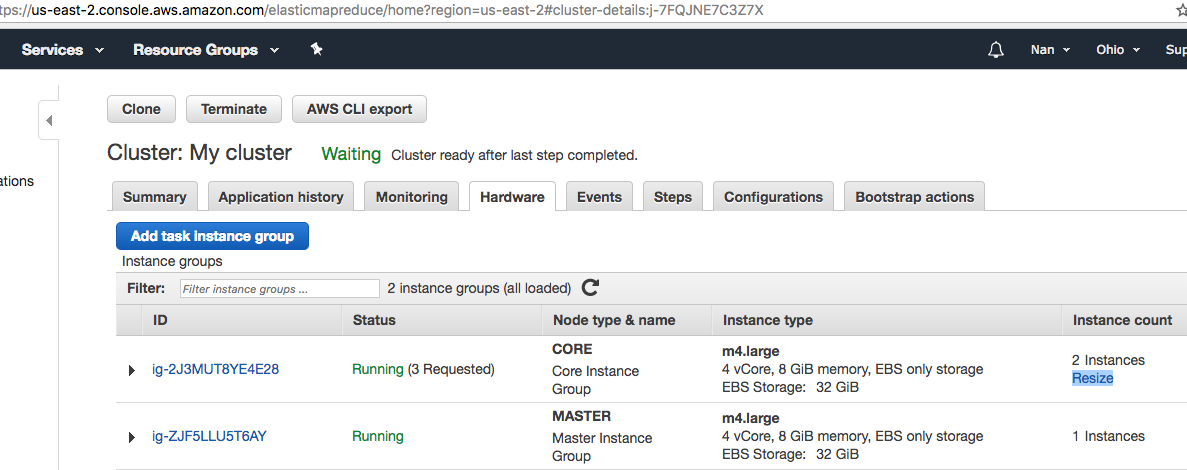
(You can find the available memory and cores in the job monitor page showing previously)

1. Add new node for performance comparison:

This can be easily done under the “hardware” tab by clicking the ‘resize’

PS: The dataset is stored on S3 which is public accessible. And you don’t need to configure it because it’s incorporated in the codes already.





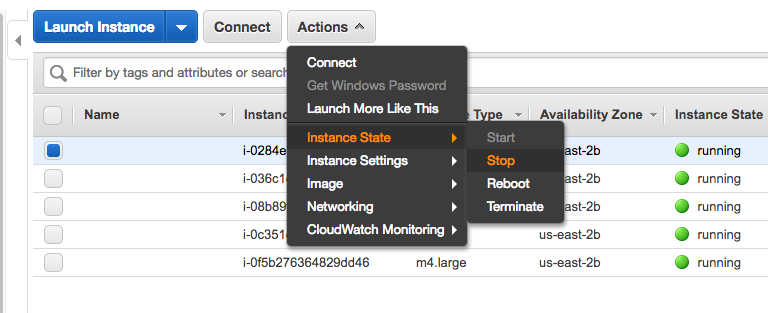
Once more memory and cores are available, please change the values in the command as below:

spark-submit analyzer.py --master yarn --executor-memory 18G --num-executors 12

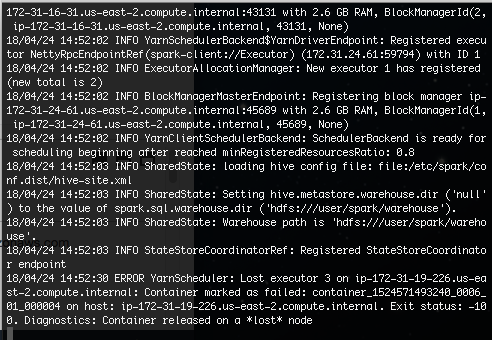
spark-submit analyzer.py --master yarn --executor-memory 24G --num-executors 16

1. Fault tolerance examination

Rerun any of above spark-submit command, then go to EC2 cluster tab and stop any of your worker node (**NOT** your master node)



Then you will see some message as below:



then record the results from the shell since the codes will display the results on your shell monitor.