AWS Setup Guidelines

For CSE6242 HW3, updated version of the guidelines by Diana Maclean

Important steps are highlighted in yellow.

What we will accomplish?

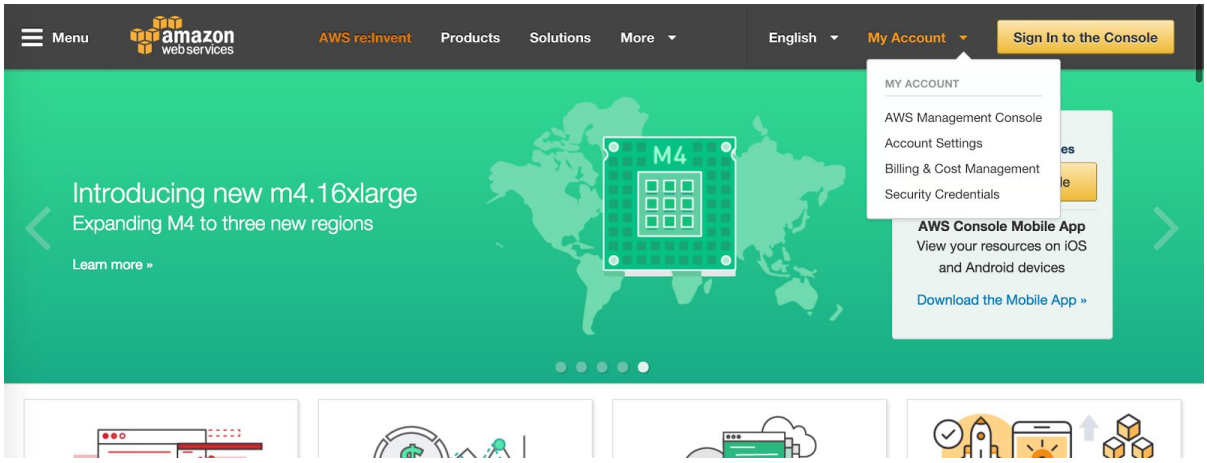
This guideline helps you get set up with the Amazon Web Services (AWS, a "cloud" platform) where you will run large-scale analysis on big data. Here are you will learn to do

1. Create an AWS account (to get access to EC2, Elastic MapReduce and S3 storage).
2. Create storage buckets on S3 (to save outputs and logs of MapReduce jobs).
3. Create a key pair (required for running MapReduce jobs on EC2).
4. Get Access Keys (also required for running jobs on EC2).
5. Redeem your free credit (worth $100).
6. Set up a CloudWatch Usage Alert
7. Familiarize yourself with S3, EC2 and EMR (by doing a sample MapReduce run).
8. Debugging

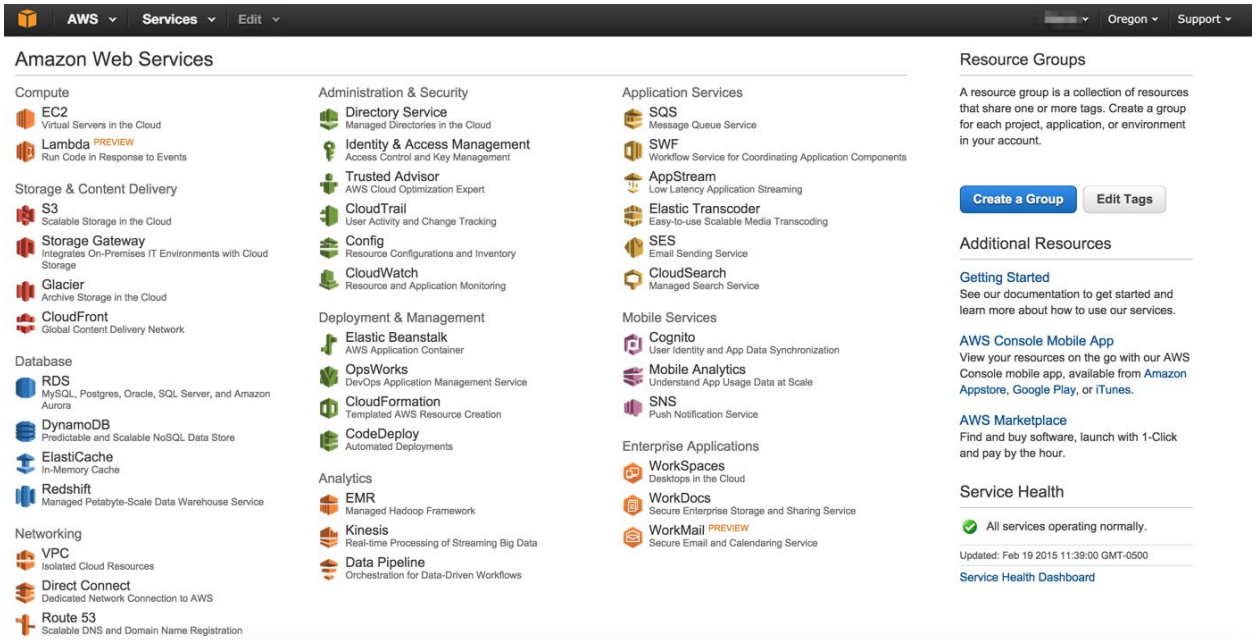
1. Create an AWS account

- Go to http://aws.amazon.com and sign up for an account, if you do not have one already.
- For now, please enter the required details, including payment details (you will need a valid credit card or debit card to sign up). Please follow Step 5 to redeem the $100 credits.
- Validate your account with the identity verification through your phone.

Once your account has been created and your payment method verified, you should have access to the AWS Management Console.



You AWS Management Console should look like this:

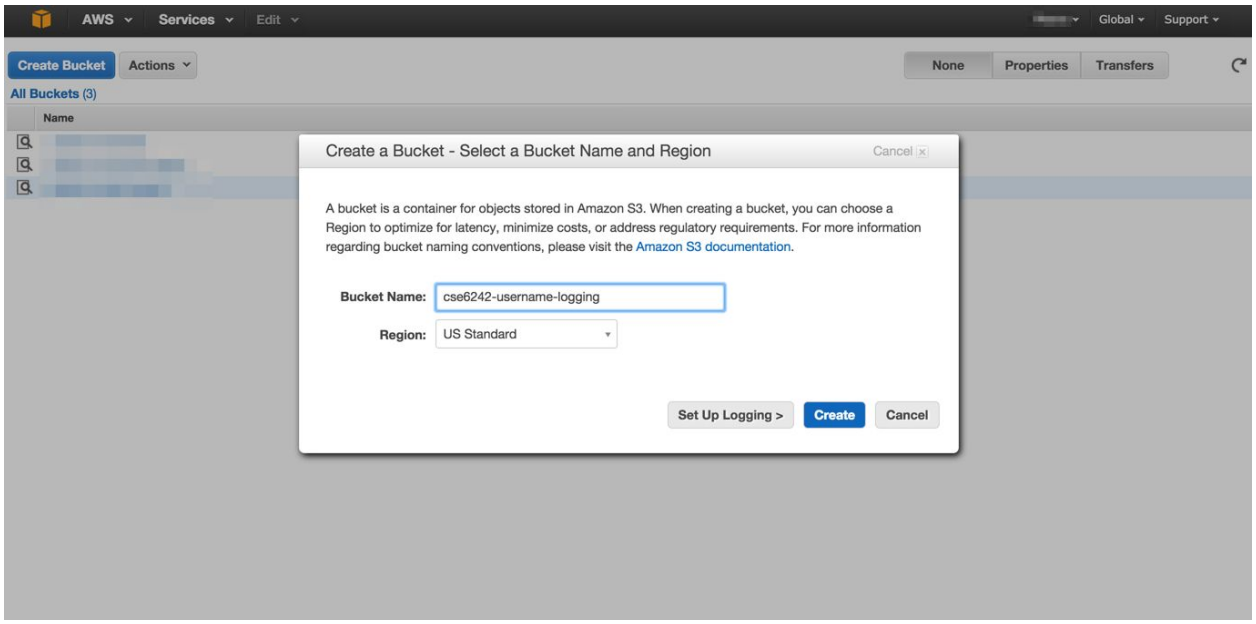


2. Create storage buckets on S3

In the AWS Management Console click on “S3” under Storage & Content Delivery. We needS3 for two reasons:
(1) an EMR workflow requires the input data to be on S3;
(2) EMR workflow output is always saved to S3.

Data (or objects) in S3 are stored in what we call “buckets”. You can think of buckets as folders. You will need to create some buckets of your own to (1) store your EMR output and (2) store your log files if you wish to debug your EMR runs. Once you have signed up, we will begin by creating the log bucket first.

i. In the S3 console, click on “Create Bucket”.

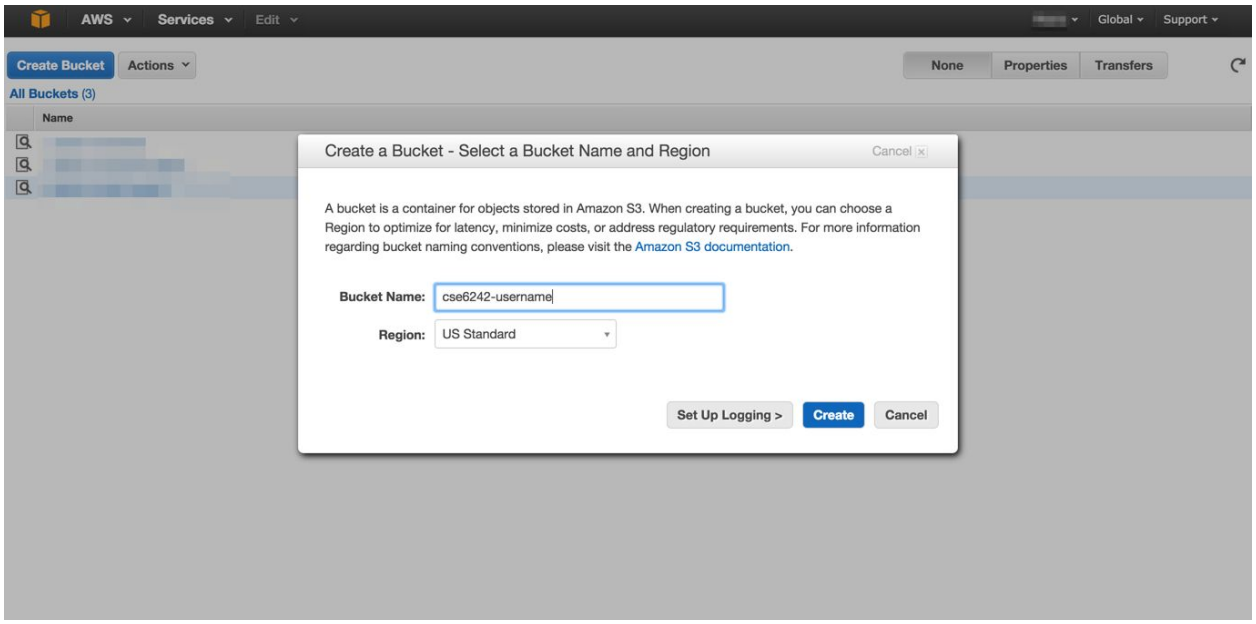


ii. All S3 buckets need to have unique names. You could name the logging bucket cse6242-<gt-username>-logging. Important: Please select "US Standard" in the Region dropdown. Click on "Create" (not on "Set Up Logging >>").

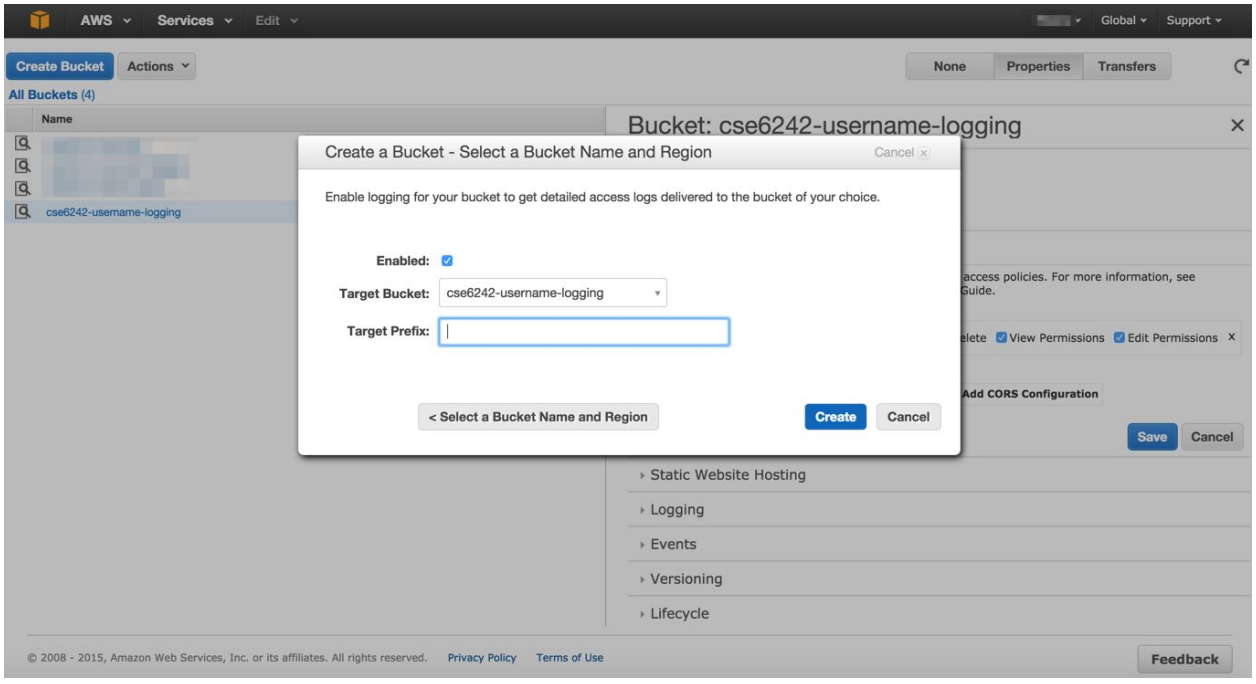
US Standard is important, because if you have buckets in other regions, data transfer charges would apply.

iii. Your new bucket will appear in the S3 console. Clicking on it will show you that it is empty.

iv. Now we will create our main bucket. Go back to the main screen (clicking on "All Buckets"). Again, click on "Create Bucket". Call this one cse6242-<gt-username>. Again, pick "US Standard" for the Region dropdown. Since we will link this bucket to our logging bucket, the regions for the two buckets should be the same. We will link our logging bucket to the one we are creating now, so click on "Set Up Logging >".



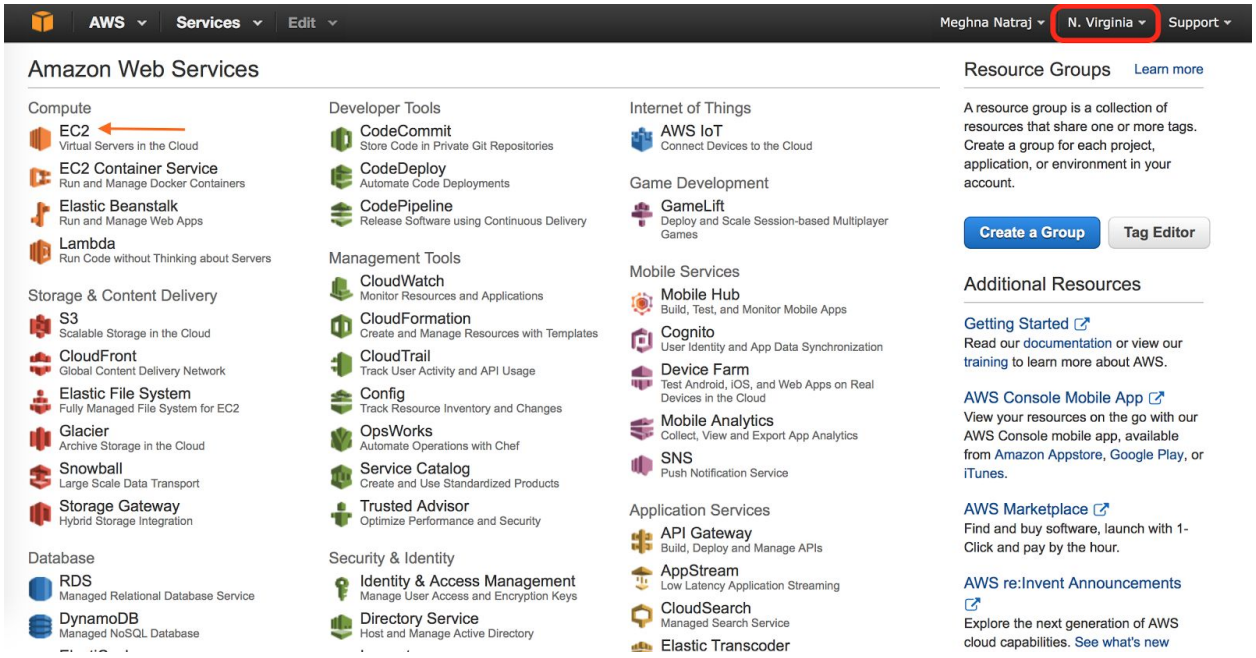
v. Click on “Enabled” to enable logging, and start typing in the name of your logging bucket. It should appear in the drop down menu, select it. Clear the “Target Prefix” field and click “Create”.



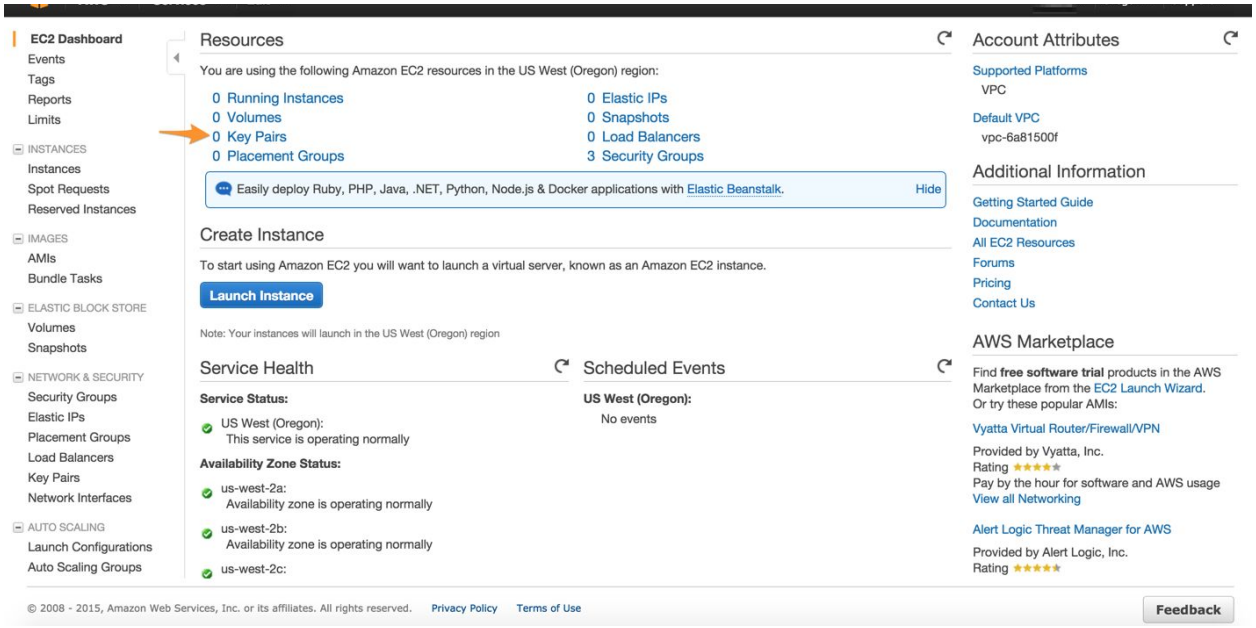
We are done creating buckets at this point.

3. Create a key pair

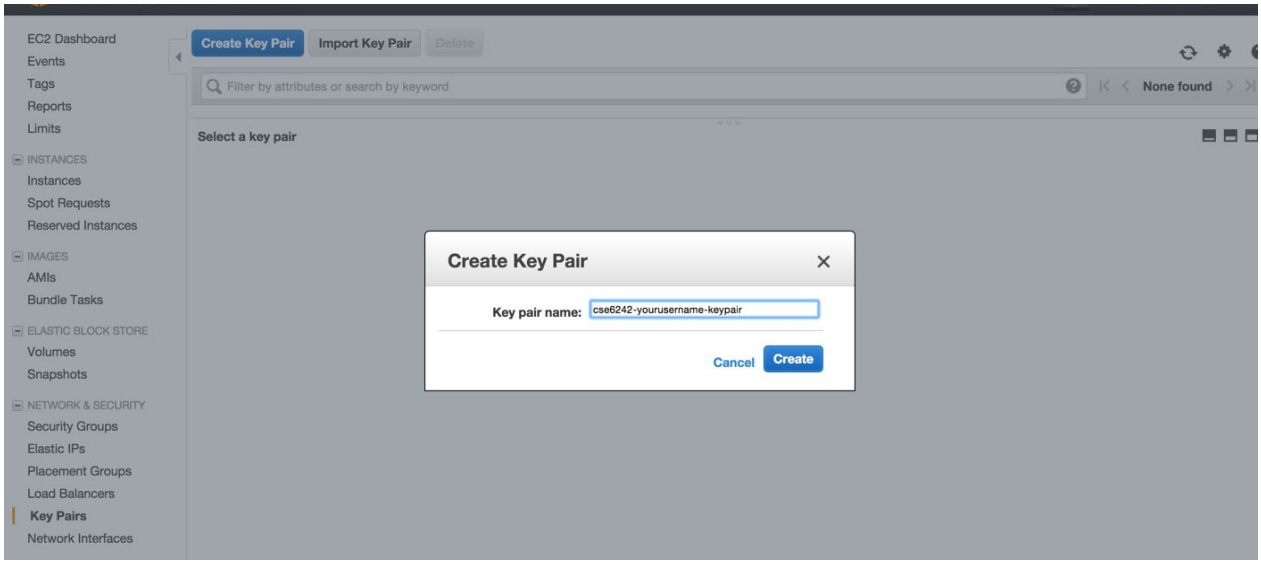
Select the region on the top right as US East (N. Virginia) since the data bucket is stored in this region. When you run jobs on EMR, you will need to have a valid public/private key pair. To create your first key pair, click on “EC2” under Compute in the AWS Management Console.



You should see a link stating “0 Key Pairs” under Resources. Click on this.



You will be given an option to “Create Key Pair”. Name your key pair as you wish. Upon providing a name and clicking on “Create”, your private key (a .pem file) will automatically download. **Save it in a safe place where you will be able to find it again (IMPORTANT, do not lose this file)**.



If you need to access your public key, you will be able to find it in the same place where you found your account credentials. Amazon keeps no record of your private key, and if you lose it, you will need to generate a new set.

If your computer runs **Windows**, use the steps in the following link to convert your .pem file to a .ppk file for use with PuTTY.
Read the section titled Converting Your Private Key Using PuTTYgen in the link below:
http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html

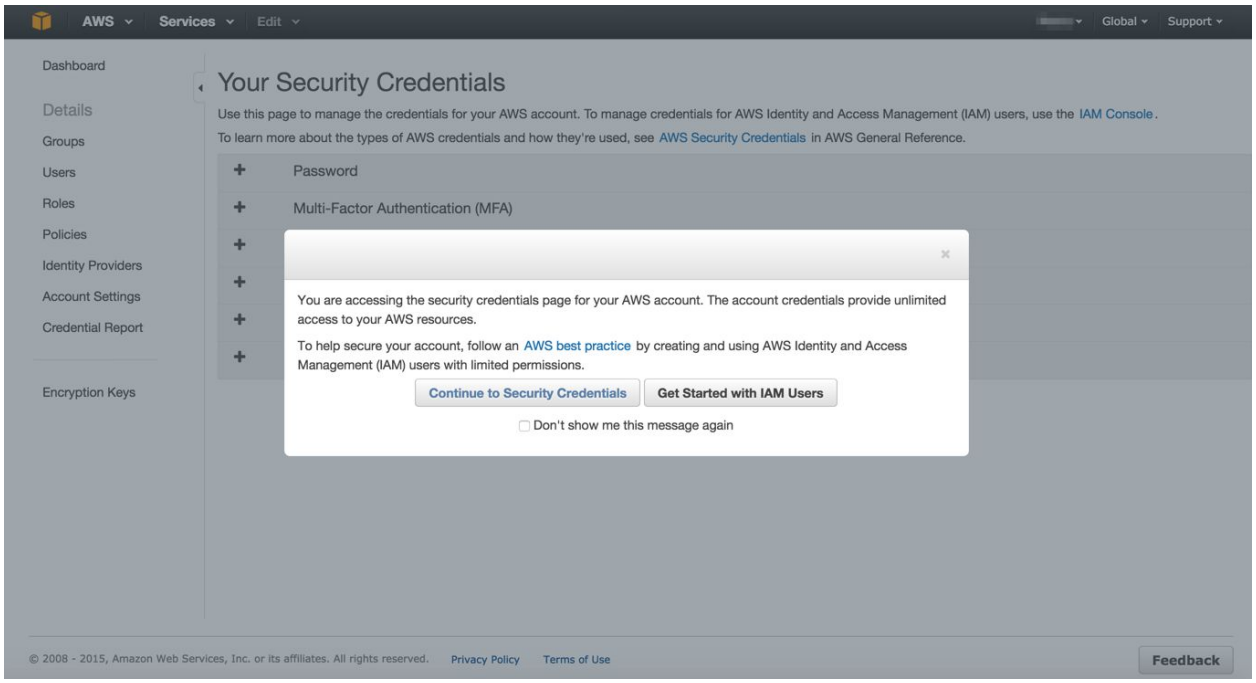
Note: If you use the AWS Management Console, you would typically not be required to access your private key. However, you will be asked to name your access key pair and the private key each time you run an EMR job.

If you wish to log into the master node running your MapReduce job, you will need your .pem file (you will need this in case you wish to run an interactive HIVE/PIG job flow). To log on to the master node (you can find the address of the master node from the MapReduce dashboard), you will need to do the following: (do not copy paste the command from this pdf as your command may fail due to the presence of special characters)

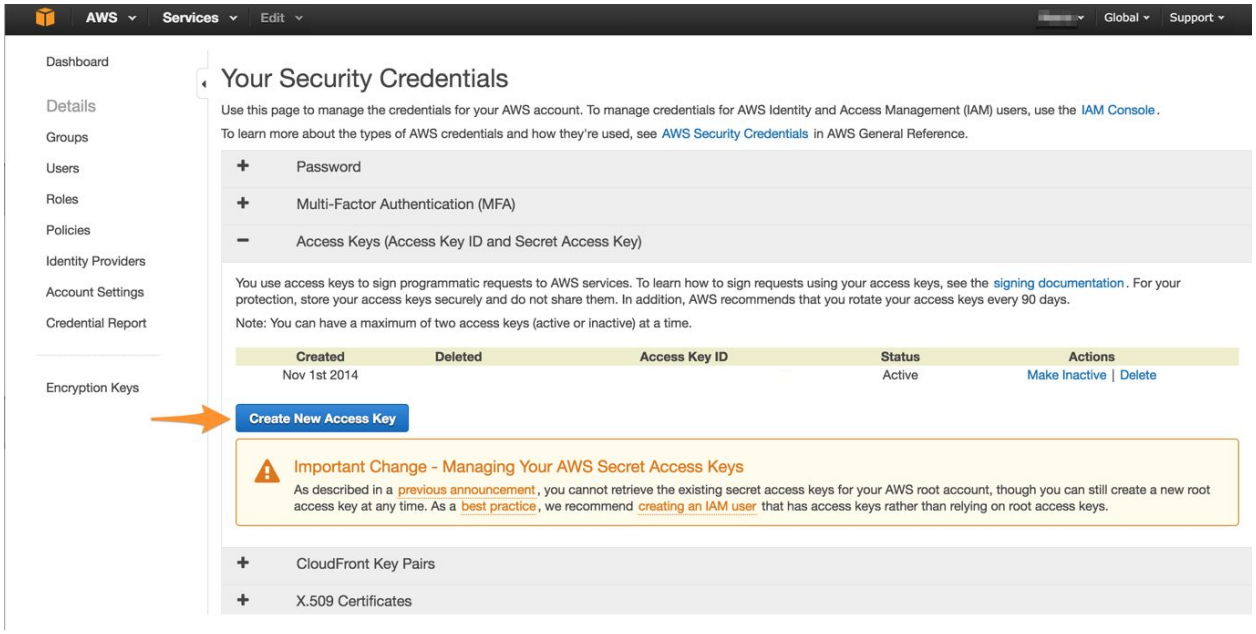
$ ssh hadoop@<master-node-address> -i <path-to-pem-file>/<pem-file-name>.pem

4. Get Access Keys (new site)

Click on “Security Credentials” under your username (top right). Click on “Continue …”

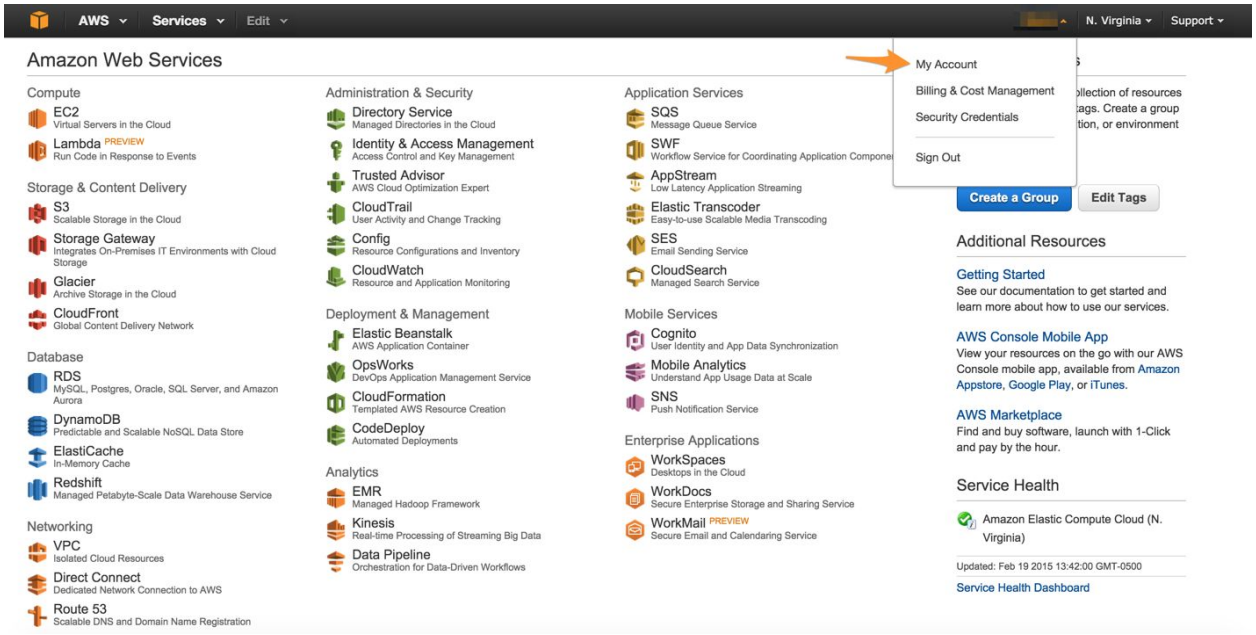


Click on the **Create a new Access Key** link (under Access Keys), and download the Access Key file (**do not lose this file**). Now you are ready to run a MapReduce job.

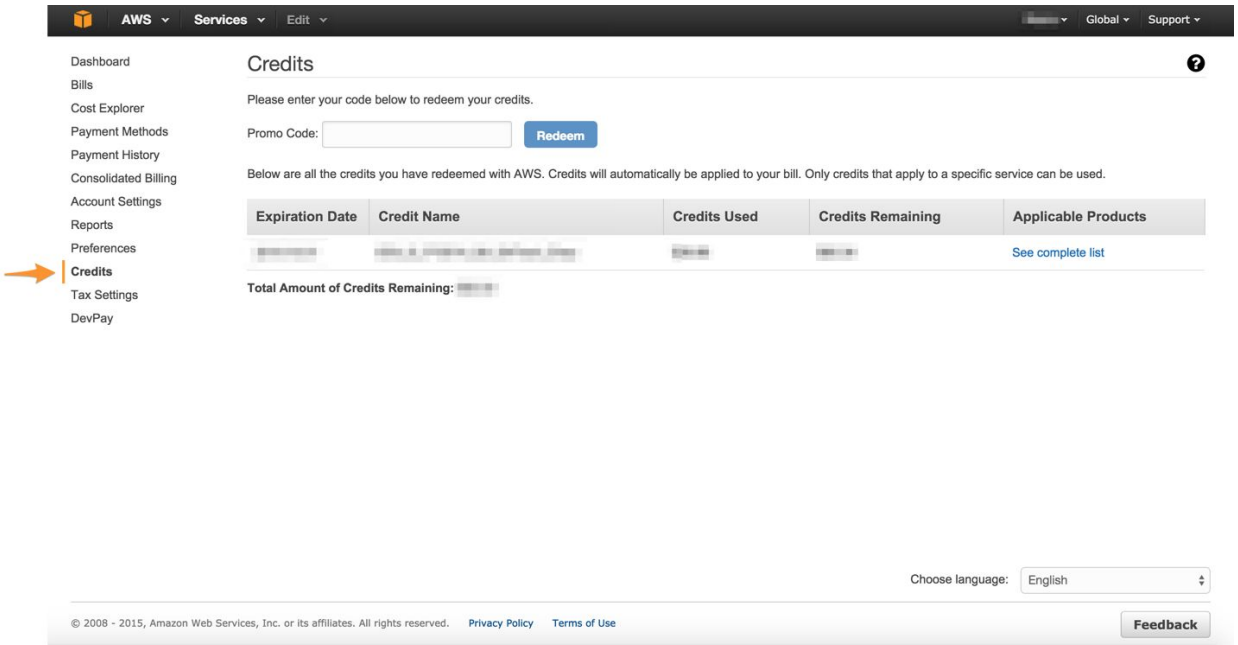


5. Redeem your free credit

In order to add the credit to your account, you will need your unique Credit Code obtained after applying for the AWS Educate program for Students (follow steps listed at the start of HW3) . Once you have your code, go to your account page ([http://aws.amazon.com/account)](http://aws.amazon.com/account)



Click on “Credits”. Enter the Code into the Promo Code text box, and click Redeem.



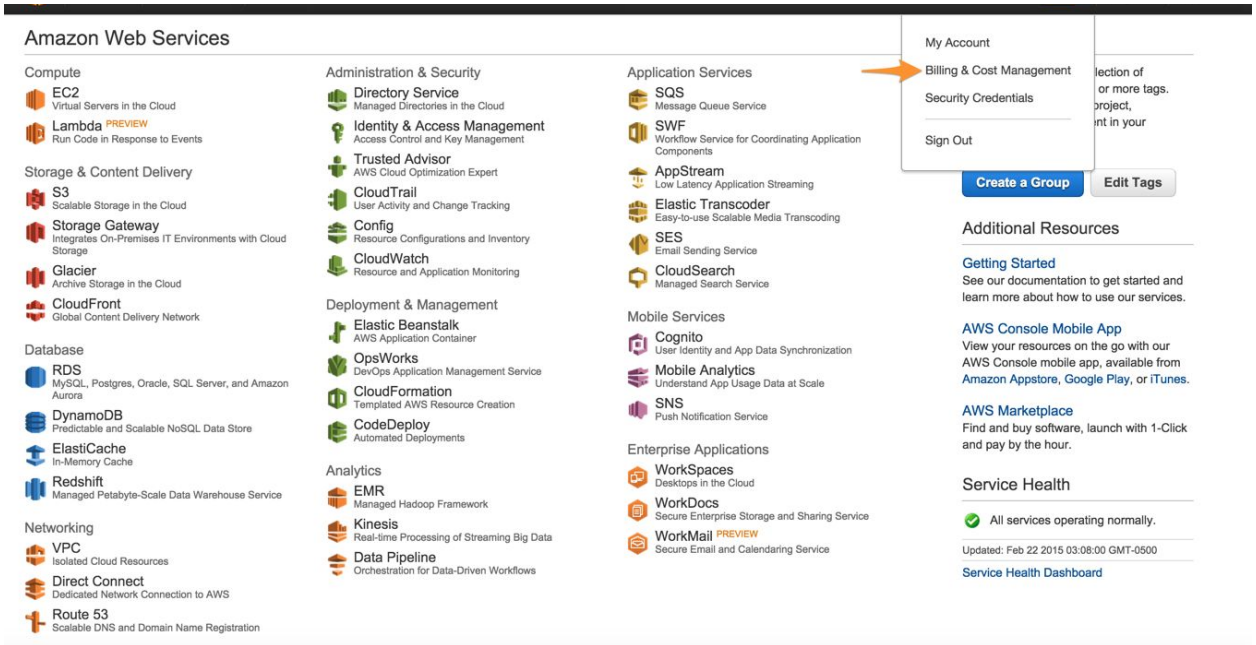
Please contact the CSE6242 instructors immediately if this does not work. You can check the credit remaining by clicking on the "Account Activity" link from your account page or by returning to this page. Sometimes this can take a while to update, so don't be surprised if recent changes are not immediately apparent. We will set up a monitor in the next step which is triggered when you utilize half of the credit.

6. Set up a CloudWatch Usage Alert

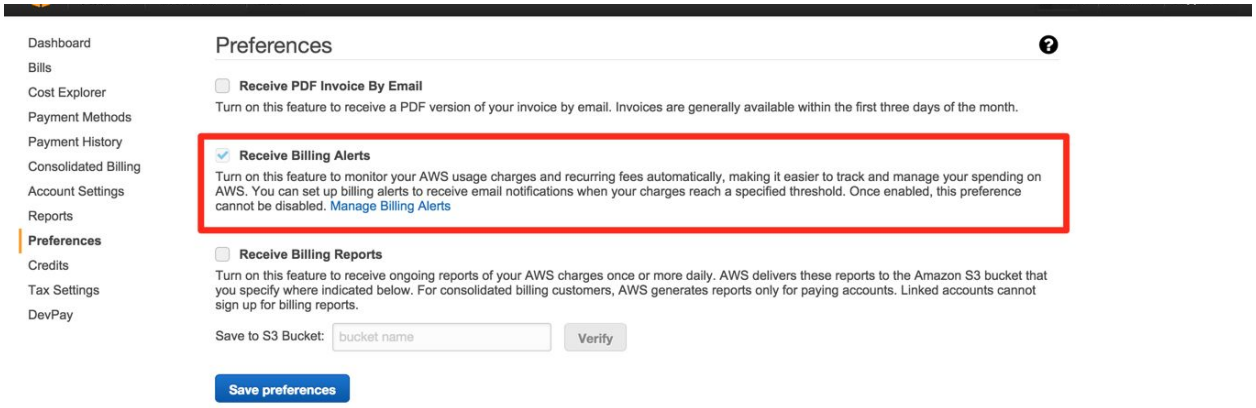
Make sure your region (in the upper right corner of the screen) is set to: US East (US Standard). Test whether this email alert is working before scheduling in practice. That is, out of 100$, when your credit balance goes below say 95$, schedule a test alert and make sure it works. Remember this alert works only once. So once you got an alert for 95$, you schedule the next alert for 70$ and the next one for 60$ and so on.
http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/free-tier-alarms.html.

Now we will turn on alerts.

1. Go to the "Billing and Cost Management" page. (Log In using your AWS credentials if necessary)

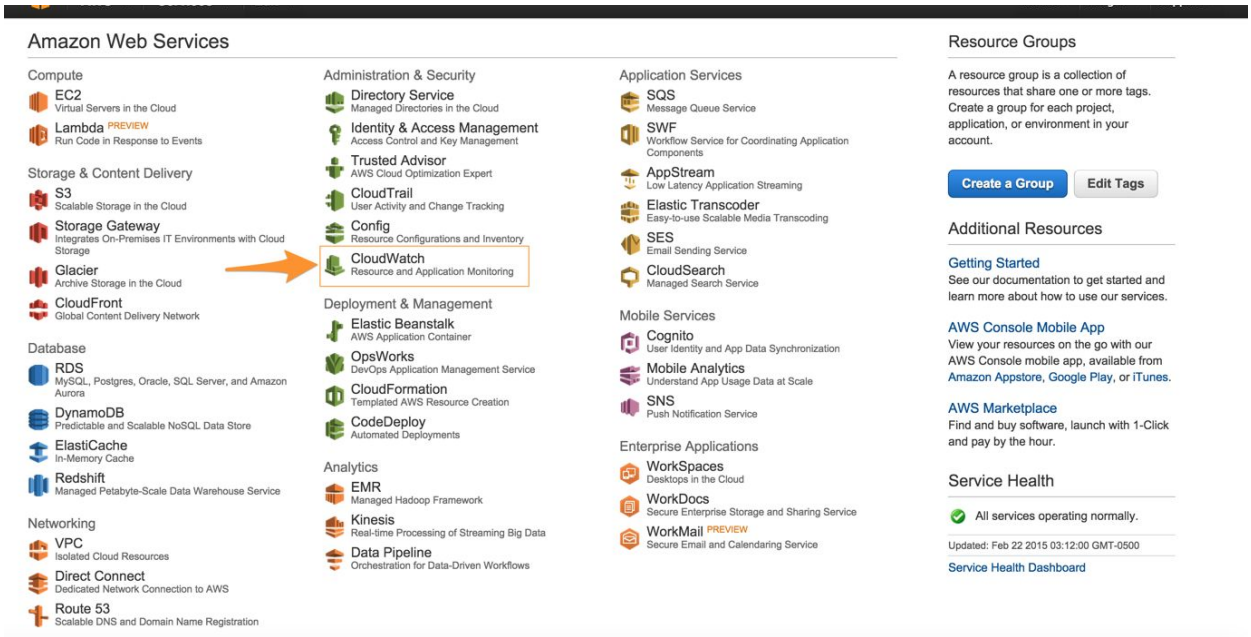


2. Under **Preferences**, check the box labeled **Receive Billing Alerts**

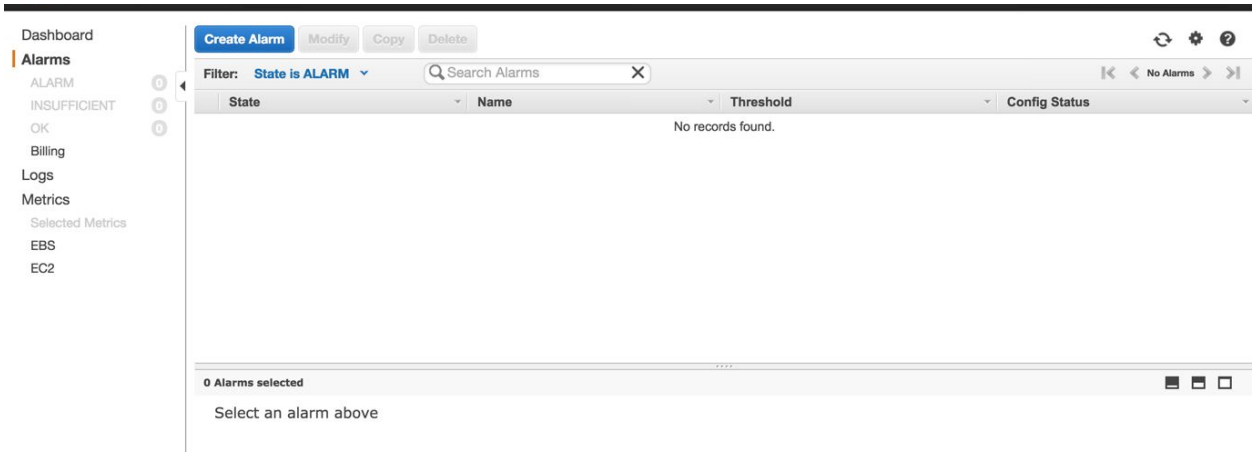


Now we need to create a custom alarm so that it tells you when you have spent money.

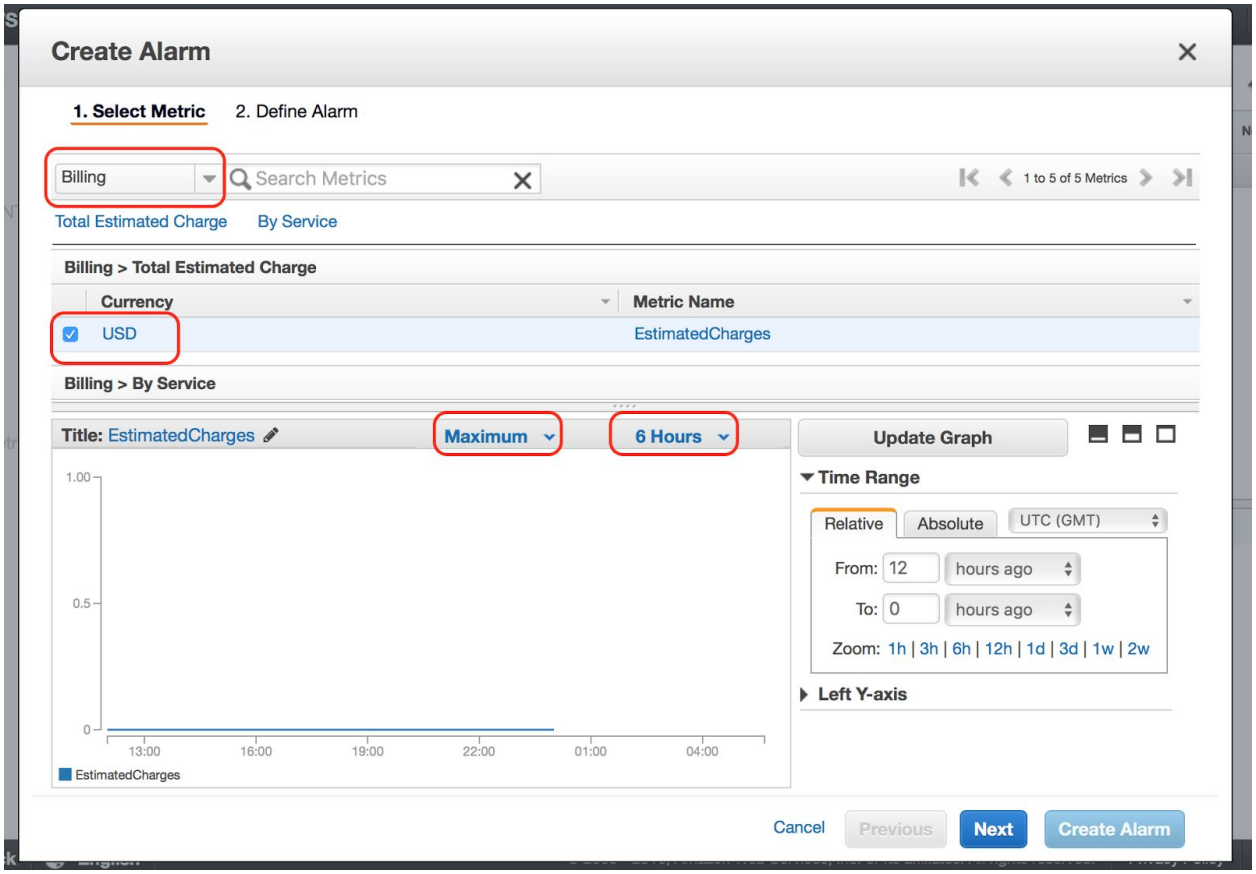
1. Click **CloudWatch** in the AWS Management Console.



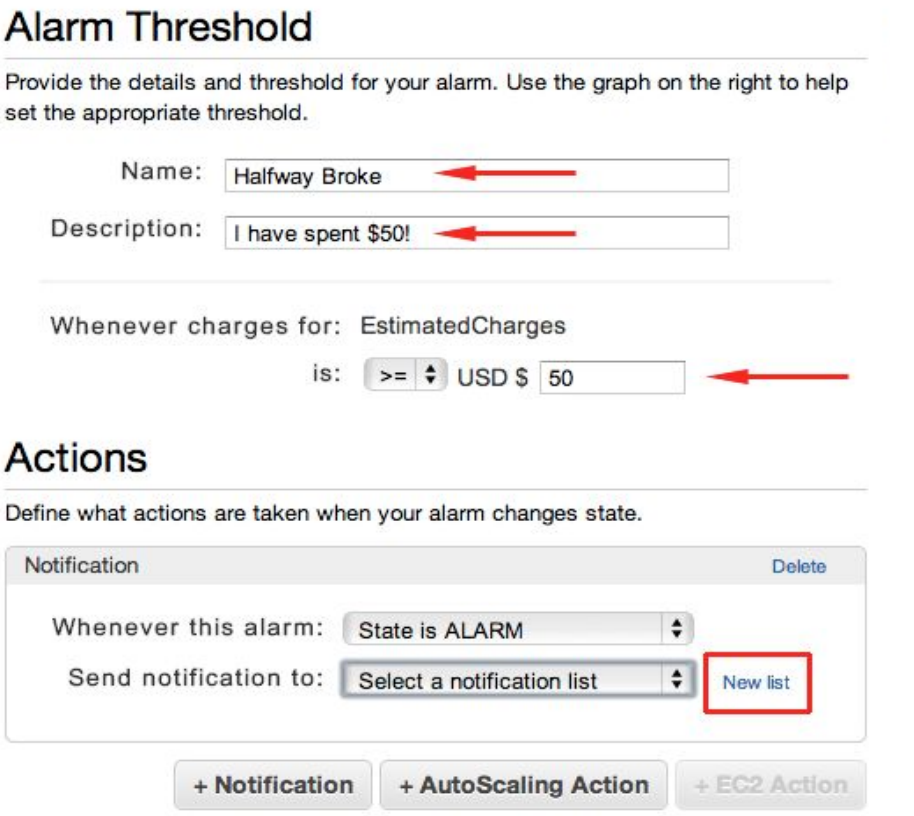
2. In the navigation pane on the left, click **Alarms**, and then in the **Alarms** pane, click **Create Alarm**.



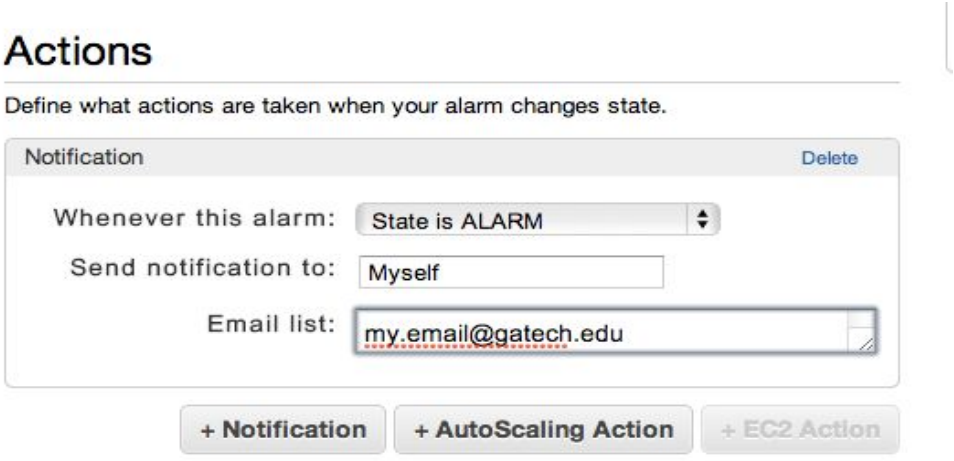
3. In the **CloudWatch Metrics by Category** pane, under **1.Select Metric,** in the dropdown choose **Billing** and check **currency as USD**. Select **"Maximum"** and **"6 Hours"** in the dropdown as shown in the image below. Click **Next**.



4. Fill out the alarm details and click **New List** next to “Send notification to:”:



Enter your name and email.



You have now created an alert that will bother you when you pass $50. Consider making another alert which is activated when you use up $90 so that you do not get charged!

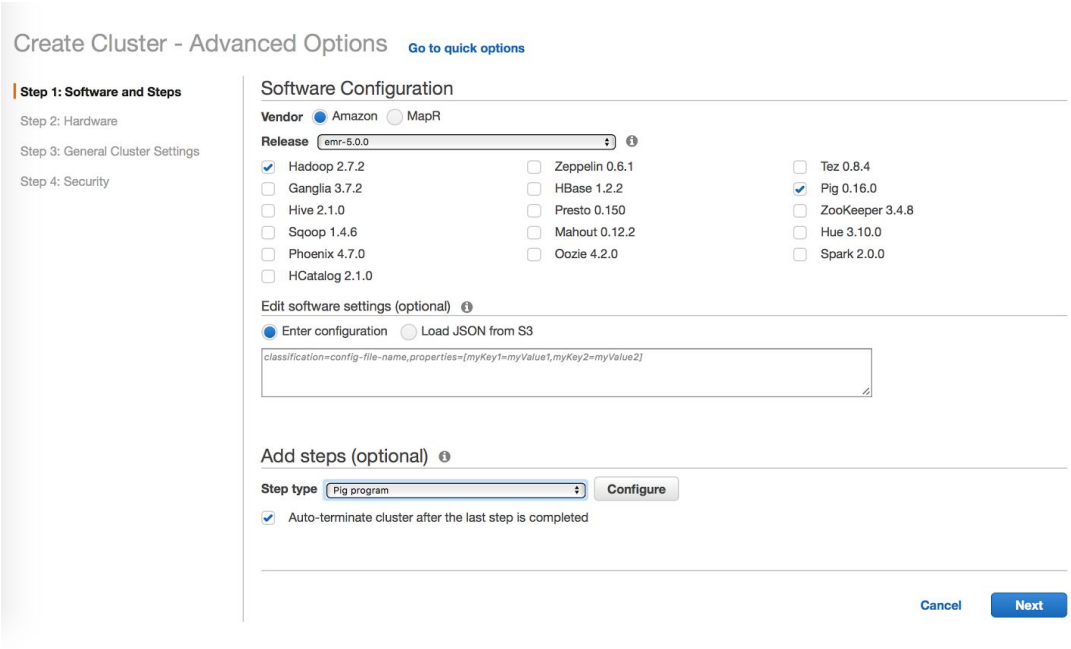
7. Familiarize yourself with S3, EC2 and EMR

We will now run a sample application. We will begin by clicking on the Elastic MapReduce(**EMR**) link in the Analytics section of the AWS Management Console. This will take you to the EMR Job Flows page. Click on the **"Create Cluster" → "Go to the advanced options".** You will be directed to the following steps.

Note:

- Ensure that you first test your code on the smaller dataset. (not larger)
- Each time you run the code, it may take a couple of hours to terminate.
- To test and debug your code step by step, refer to the Debugging section at the end of the document. This is highly recommended if you are not familiar with Pig.

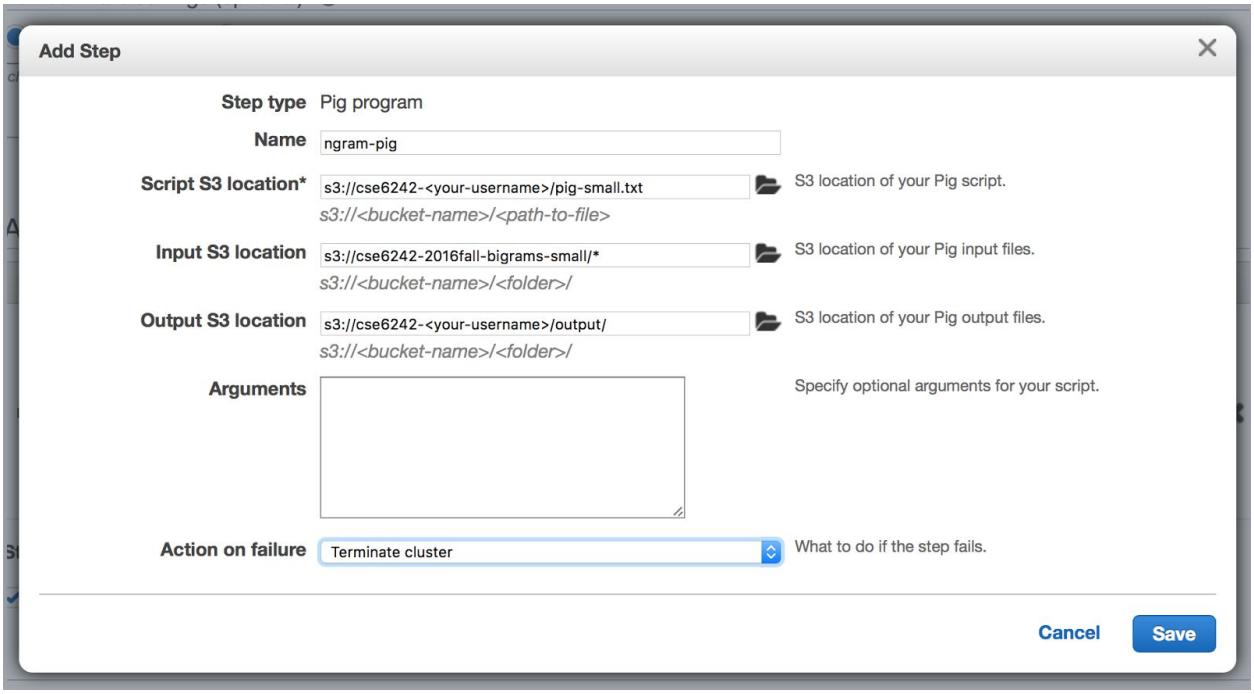
1. Under **Step 1 : Software and Steps** , Select only **"Hadoop"** and **"Pig"** from the options and unselect others in the Software Configuration options menu. In the **"Add Steps"**, choose Step Type **"Pig Program"** and then click **"Configure"**.



1a. Fill the form with details as provided in the box and image below.

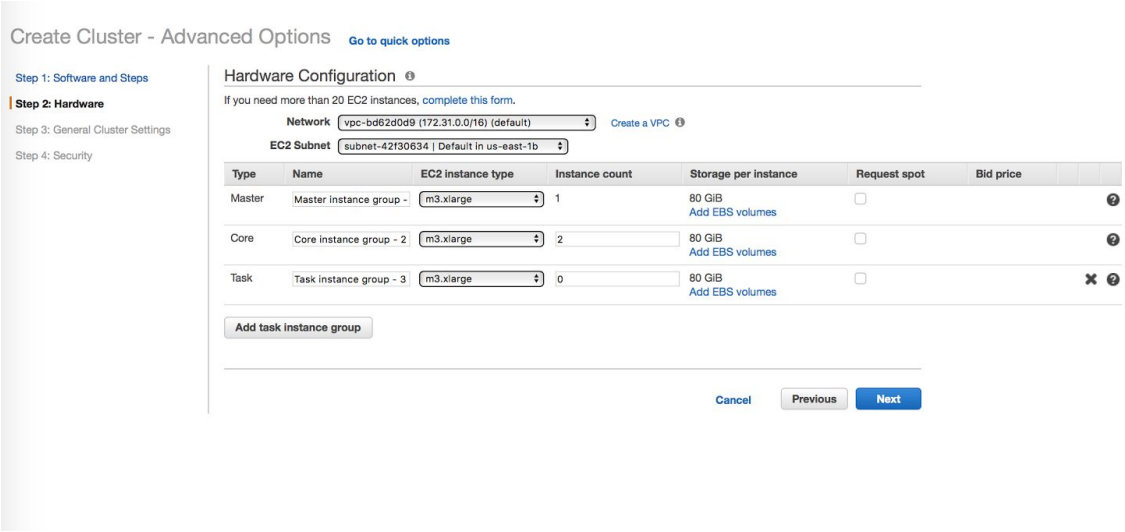
Name : (any name)
Script S3 Location : s3://cse6242-<your-username>/pig.txt (must upload the script here)
Input S3 Location: s3://cse6242-2016fall-bigrams-small/* (or big - as provided in the HW)
Output S3 Location : s3://cse6242-<your-username>/output (must be unique)
Action on failure : Terminate (else you may be charged even if the task fails)

Upload your script to an S3 location and select the location of your script from the list of items available at **"Script S3 Location"**. For the S3 output Location you should specify the bucket and an **additional unique folder for each new run**. It will help with organization. Now, click **Save.**

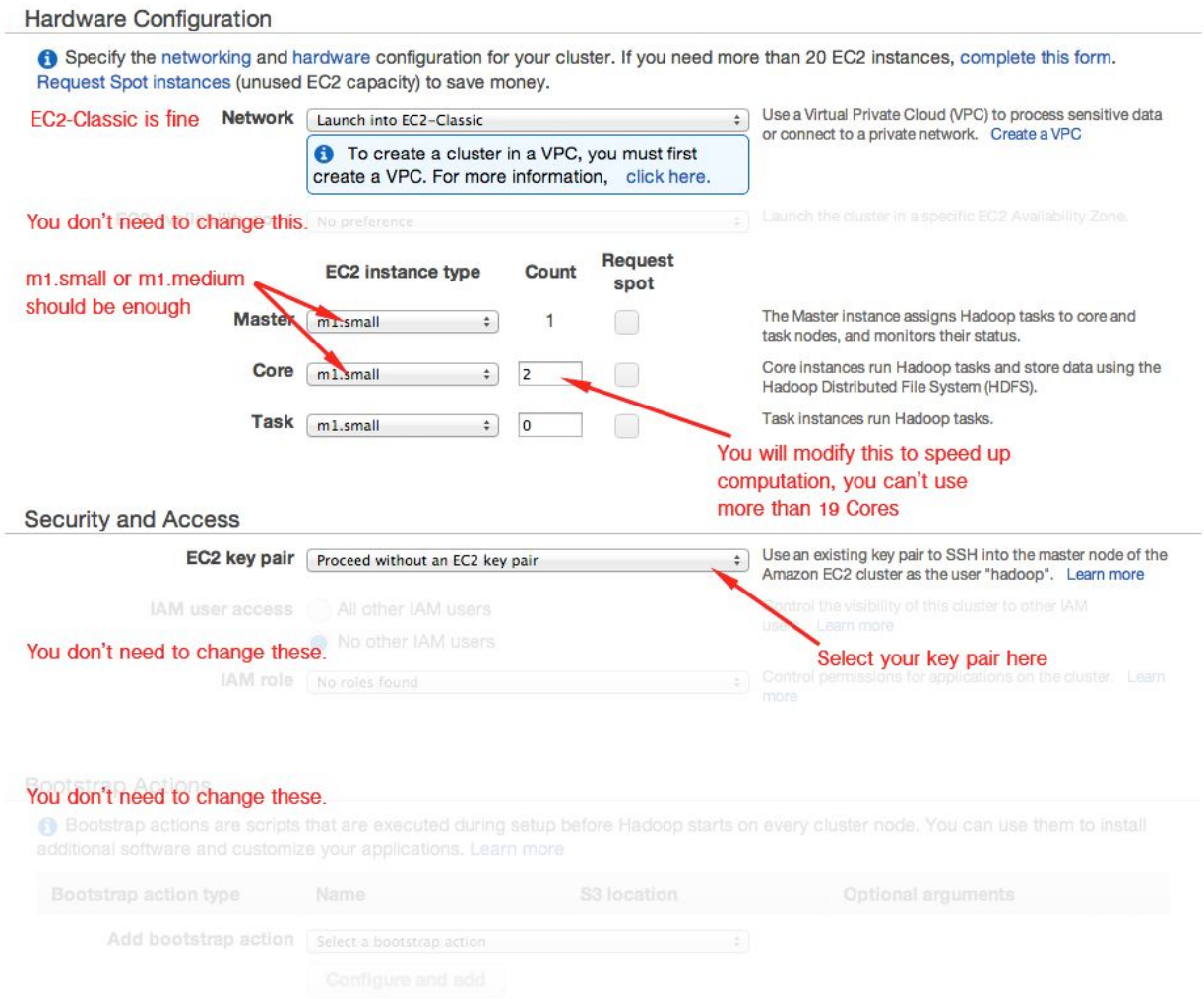


2. For **Step 2 : Hardware** configuration, you may see one of the following two views. Modify the EC2 instances as per your needs and select **Next. (**One Master instance and 1-15 Core instances should be sufficient. You may face Bootstrapping errors if you exceed a certain limit of core instances)

View 1 : Using VPC (Virtual Private Cloud)



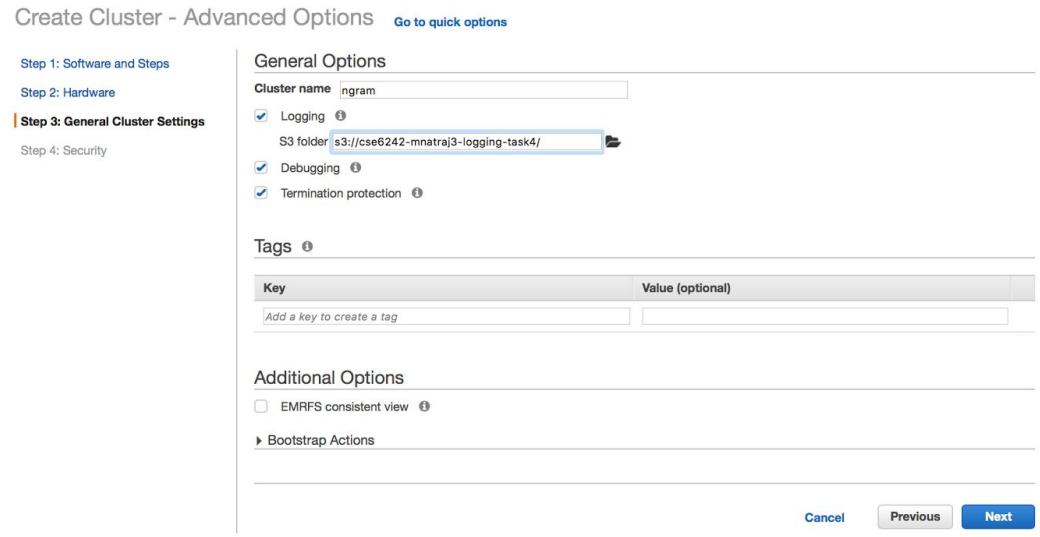
View 2 : Using EC2 - Classic



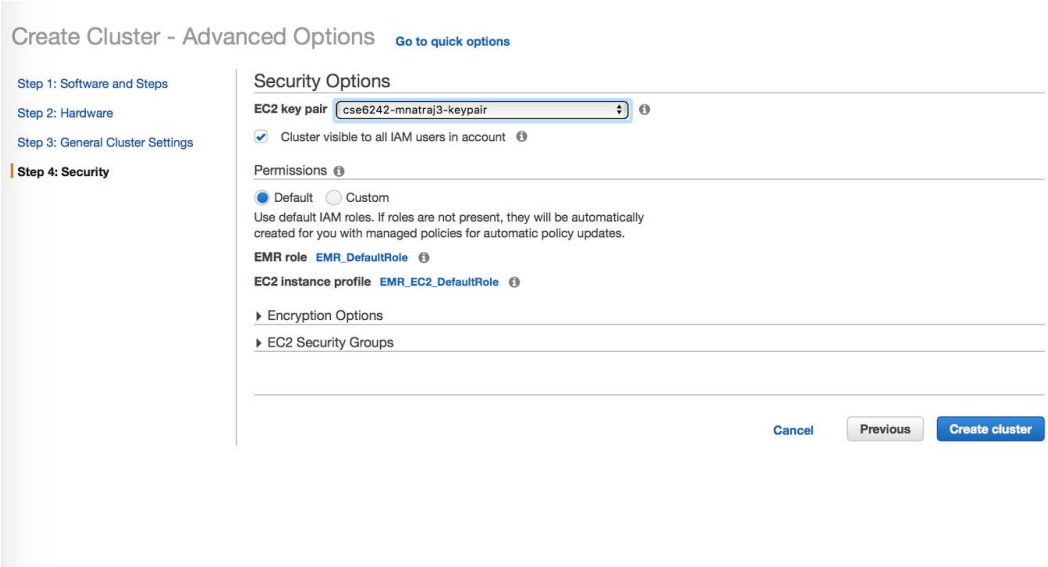
Note: If your account supports only EC2-VPC, you can select the default VPC from the Network list i.e. you will not see "EC2-Classic".

The costs listed in pricing are charged on an hourly rate, based on the number and type of nodes in your cluster.

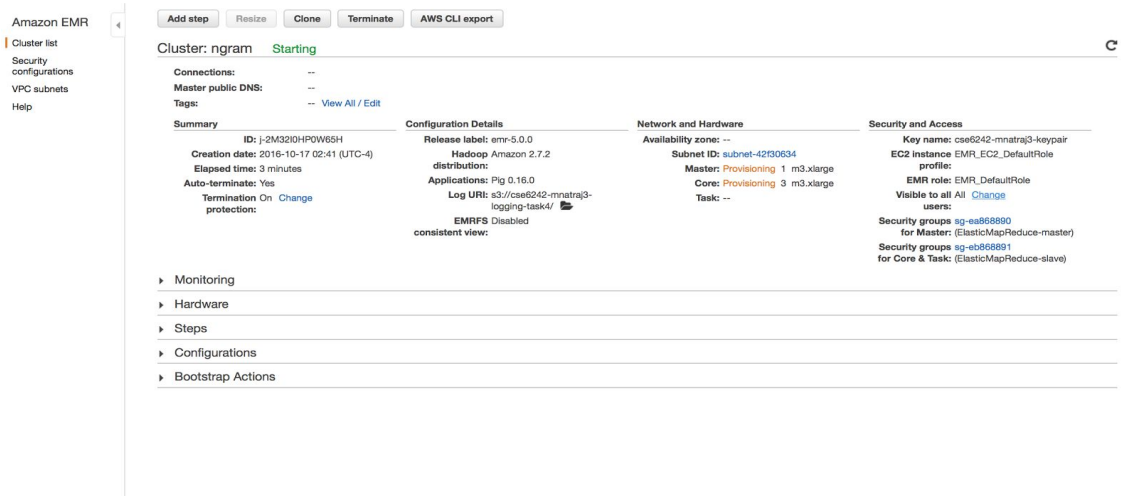
3. For **Step 3 : General Cluster Settings,** type a cluster name of your choice, and add the correct path to the logging folder (created in Step 2). Check Logging, Debugging and Termination protection as shown in the image below. Click "**Next**".



4. For **Step 4 : Security,** select your keypair and click **"Create Cluster"** to run the application.



5. The cluster must start running as follows,



You now can view the status of your application in this “Cluster Details” screen. It takes several minutes for the whole process to run.
Provisioning - Amazon locates resources for your application
Bootstrapping - Amazon sets up and configures the nodes to run your application
Running - Runs and writes to your output bucket.
Terminating - Amazon deconstructs the setups you used for the application

You can track its progress once it’s been created.

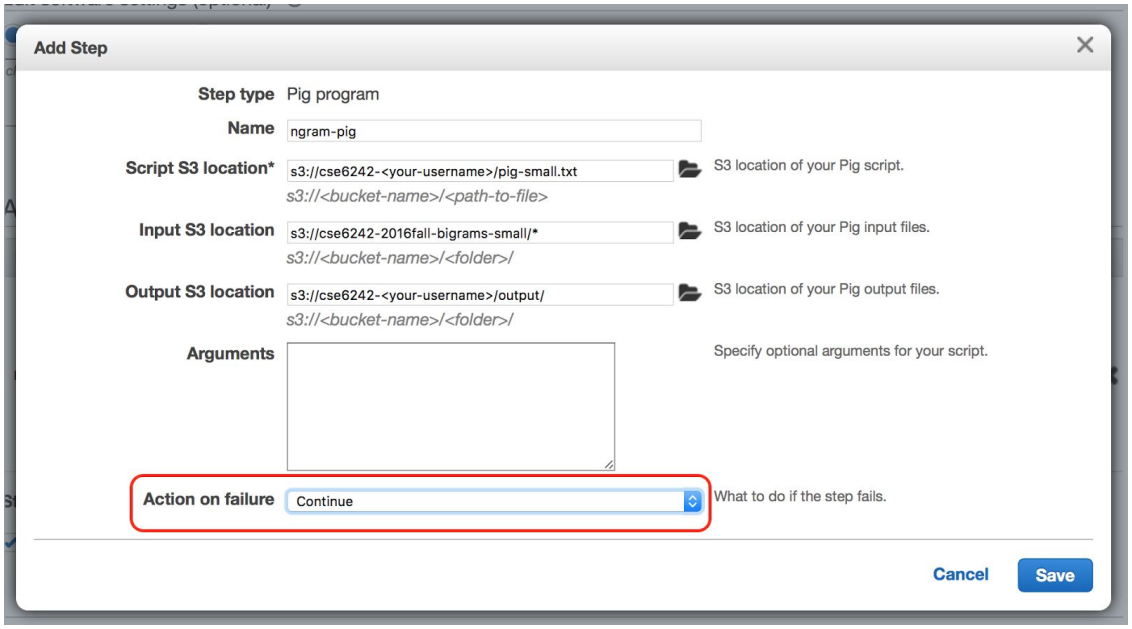
After the application terminates, you could go back to the S3 output bucket you chose. The results will be written to the output folder. You should have several partxxxx files in the output folder. These are texts of the output! You have just successfully completed a MapReduce job flow on AWS and are ready for large scale data analytics.

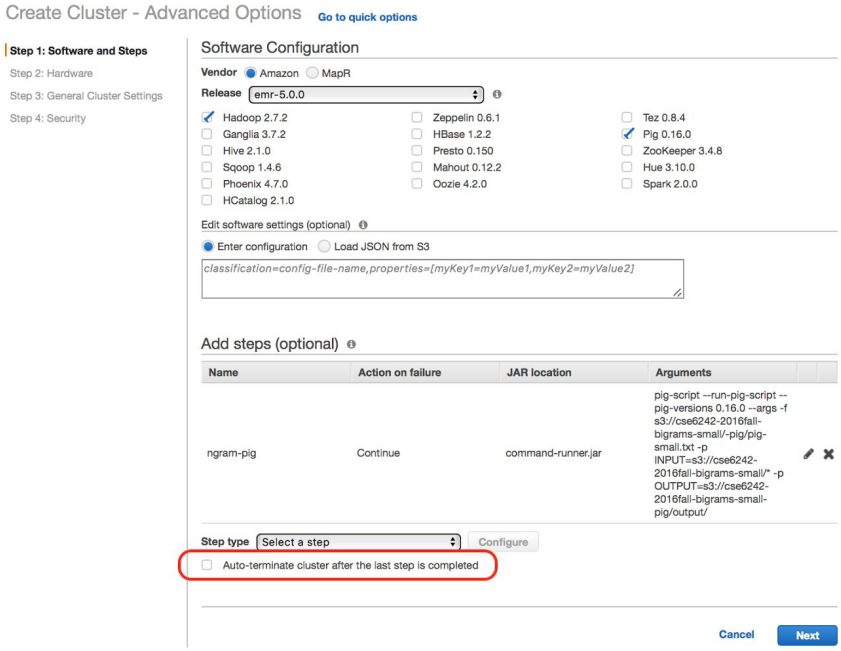
8. Debugging

A very important part of running Pig Scripts on AWS is the ability to also run your code directly on the master node. You can run your script step by step and identify the exact step where an error occurred. The steps to debug are given below.

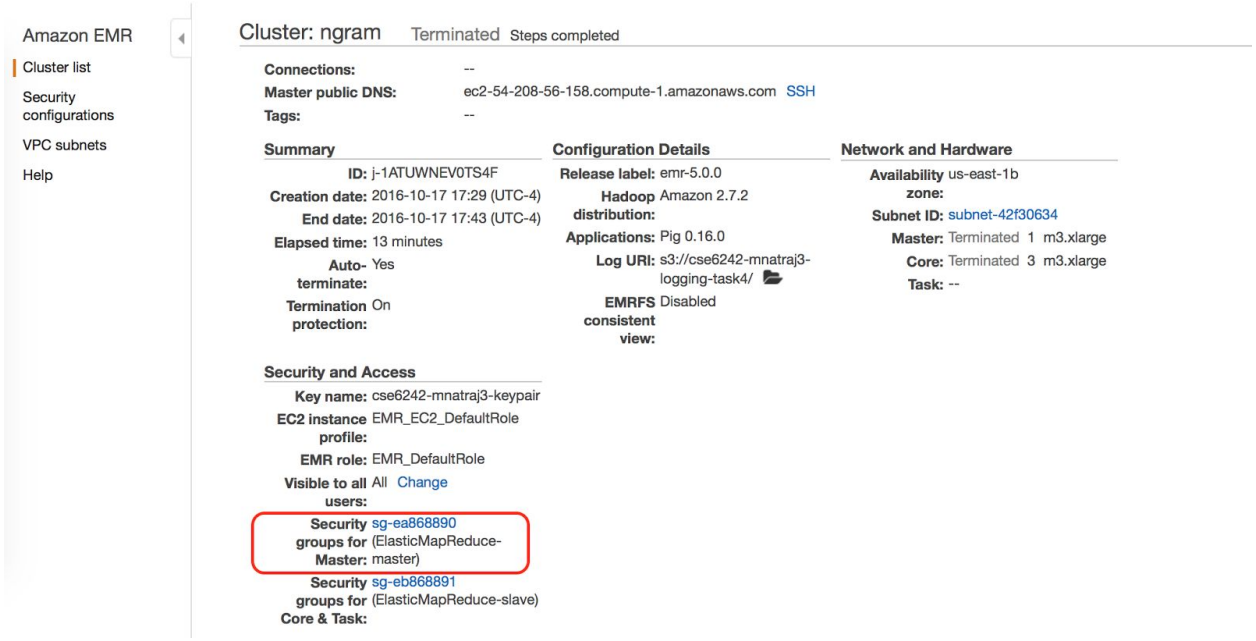
1. You must repeat all the steps in Section 7, except with three modifications:
    a. Ensure that you verify the script location, its input and output path. Do this each time you create/clone a cluster (many students make a mistake here)
    b. Modify the action on failure option to “Continue”
    c. Uncheck the “Auto-terminate cluster after….” option.

Warning : You must revert back these changes after debugging else you may leave the clusters running forever and you will be charged for this.

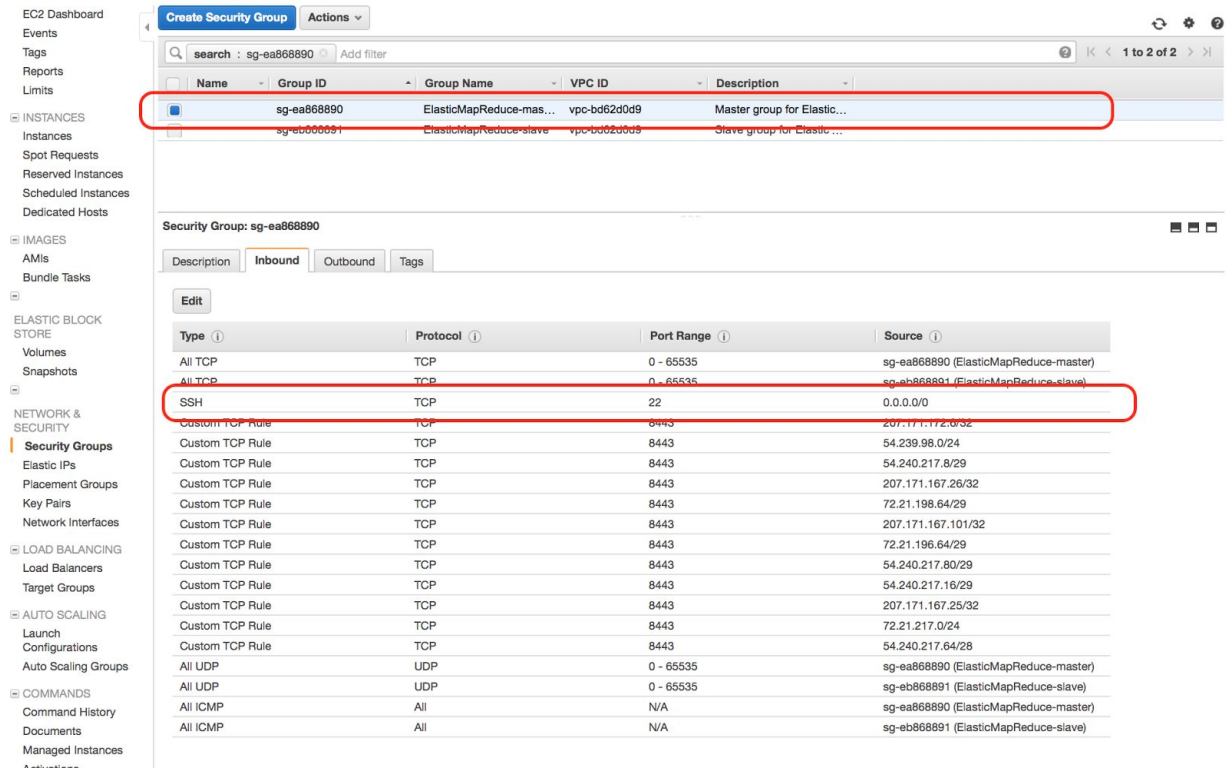

Add Step
Step type
Pig program
Name
ngram-pig
Script S3 location*
s3://cse6242-<your-username>/pig-small.txt
S3 location of your Pig script.
s3://<bucket-name>/<path-to-file>
Input S3 location
s3://cse6242-2016fall-bigrams-small/*
S3 location of your Pig input files.
s3://<bucket-name>/<folder>/
Output S3 location
s3://cse6242-<your-username>/output/
S3 location of your Pig output files.
s3://<bucket-name>/<folder>/
Arguments
Specify optional arguments for your script.
Action on failure
Continue
What to do if the step fails.
Cancel
Save


Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Software Configuration
Vendor
Amazon
MapR
Release
emr-5.0.0
Hadoop 2.7.2
Ganglia 3.7.2
Hive 2.1.0
Sqoop 1.4.6
Phoenix 4.7.0
HCatalog 2.1.0
Zeppelin 0.6.1
HBase 1.2.2
Presto 0.150
Mahout 0.12.2
Oozie 4.2.0
Tez 0.8.4
Pig 0.16.0
ZooKeeper 3.4.8
Hue 3.10.0
Spark 2.0.0
Edit software settings (optional)
Enter configuration
Load JSON from S3
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
Add steps (optional)
Name
Action on failure
JAR location
Arguments
ngram-pig
Continue
command-runner.jar
pig-script --run-pig-script --pig-versions 0.16.0 --args -f s3://cse6242-2016fall-bigrams-small/-pig/pig-small.txt -p INPUT=s3://cse6242-2016fall-bigrams-small/* -p OUTPUT=s3://cse6242-2016fall-bigrams-small-pig/output/
Step type
Select a step
Configure
Auto-terminate cluster after the last step is completed
Cancel
Next

2. Once the cluster is running, you can open the TCP Port of your Master node to allow SSH connections. Click on the security group of your master node.

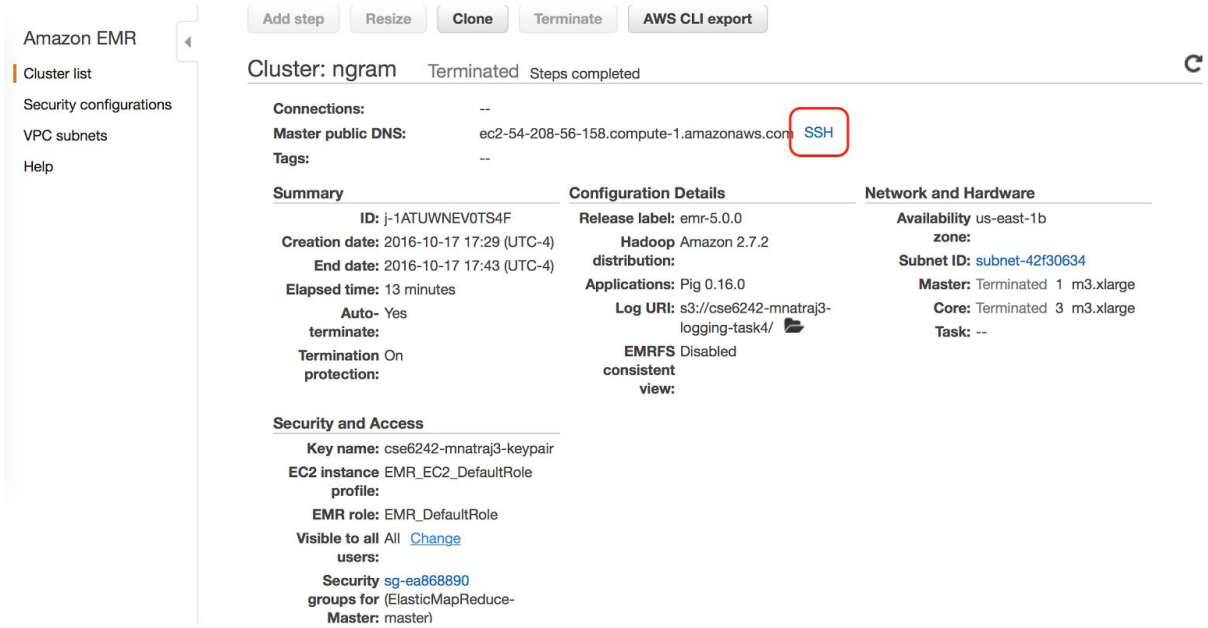


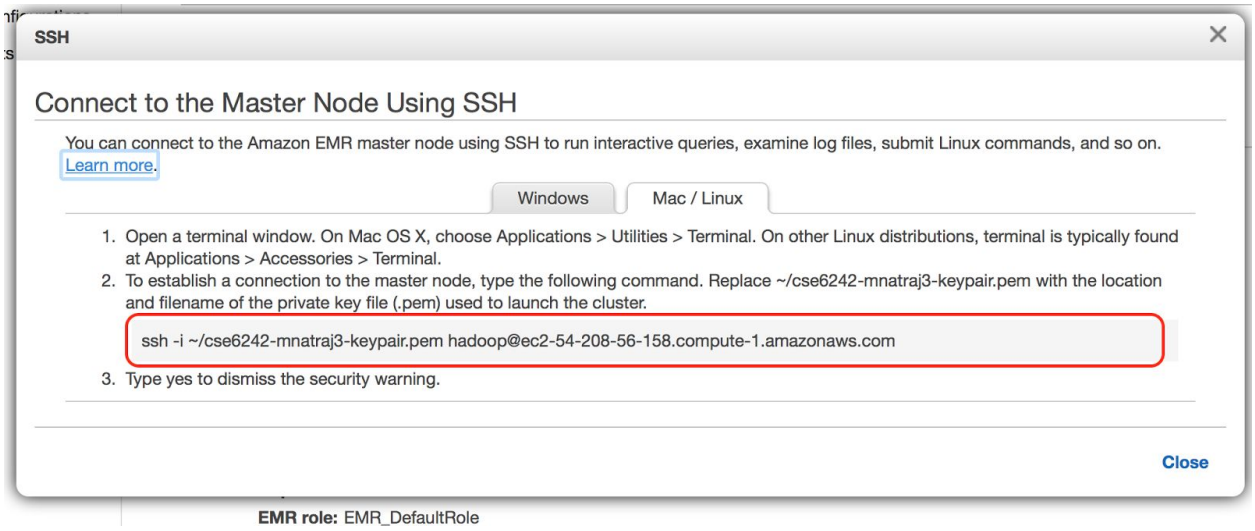
Add an entry for SSH in the inbound tab of your master node with the exact details as follows.



You can now SSH into your master node.

3. To SSH, first copy the command as follows.





Modify the path to your .pem file and run the command on your terminal. (ensure that the file permissions of your .pem file is set to 400).

4. You will now be logged into the master node. Type **pig** to be able to run commands on the pig shell.

5. Run your code line by line and spot the errors!