Tutorial- Counting Words in File(s) using Elastic MapReduce (AWS)

Prepared by- Srinivasan Rajappa

Under the guidance of Prof. Bina Ramamurthy

1 OVERVIEW

This document serves as a tutorial to setup and run a simple application in Elastic MapReduce which is a service provided by Amazon Web Services.

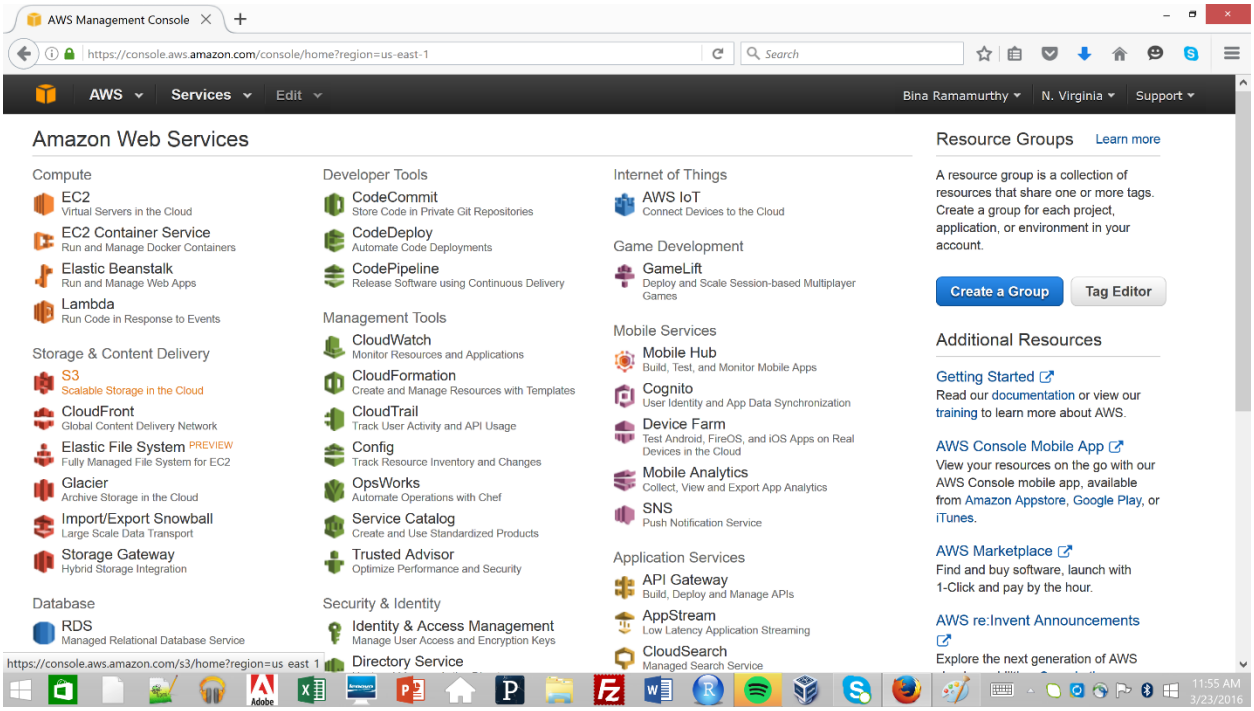
The process involves three phases viz. setting up an AWS Educate account, Creating Buckets within S3 and then to configure and run a cluster. In addition to covering these topics, additional notes and warnings are also be provided.

2 CREATING AWS EDUCATE ACCOUNT

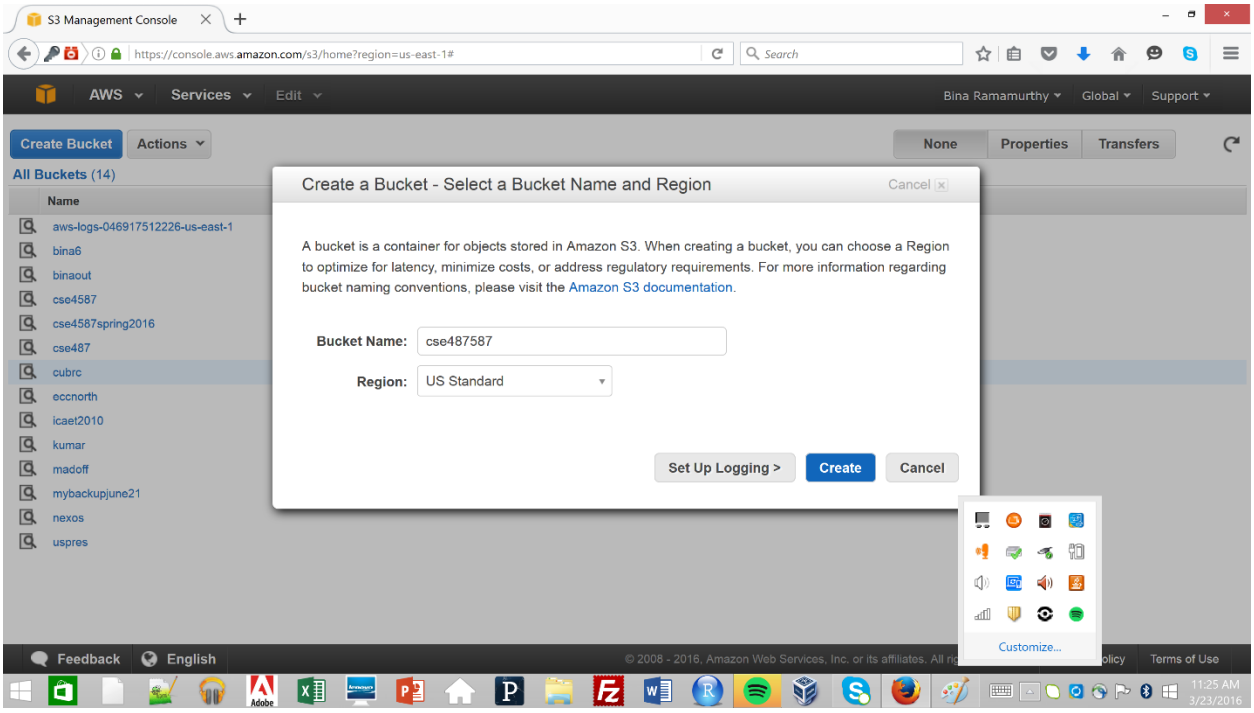
An account can be created by following this video link here. Please ensure that you register using your *buffalo.edu*[Why ?] email account as it will provide various benefits.

3 S3 BUCKETS

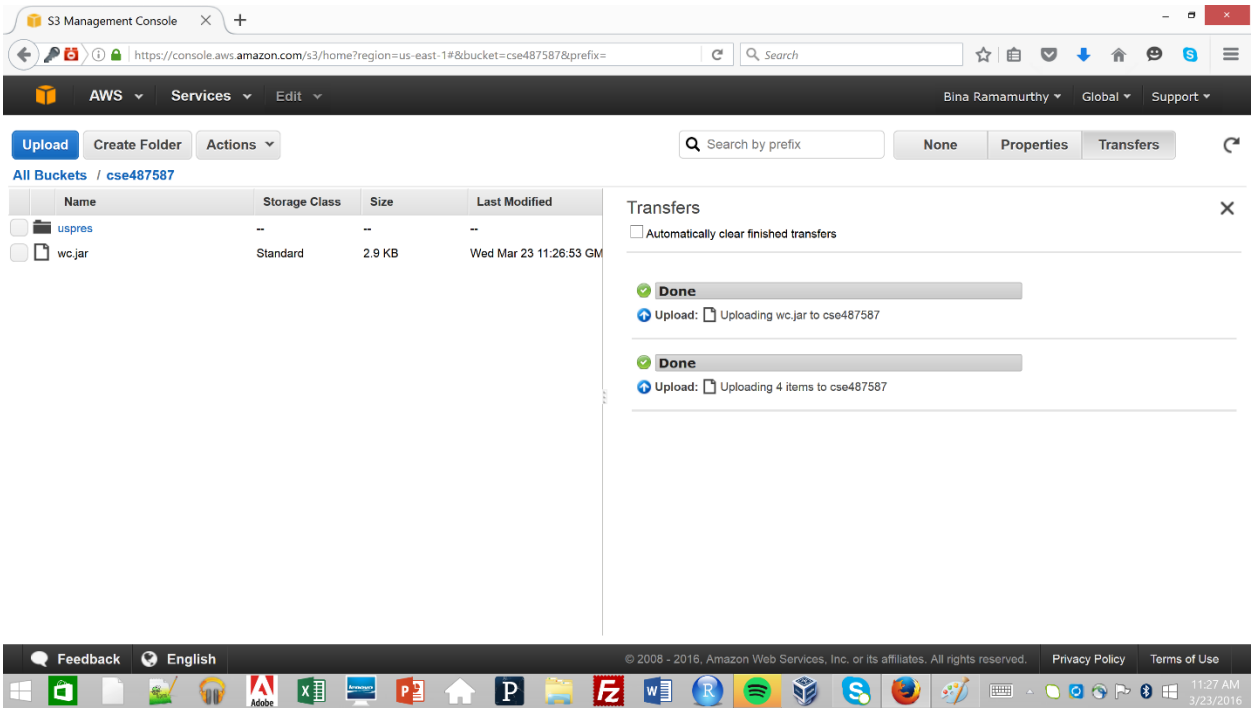
After successfully creating an account. Please login to this AWS console. Once you reach the page, click on Simple Storage Service (S3). Please refer to the following screenshot:



After opening the S3 webpage, please click on create a new bucket. Soon a popup requesting details will appear. Provide a universally unique name to the bucket, it should only comprise of lower case alphabets and numbers[Why ?]. Also, select the region as **US Standard**. Please refer to the following screenshot:

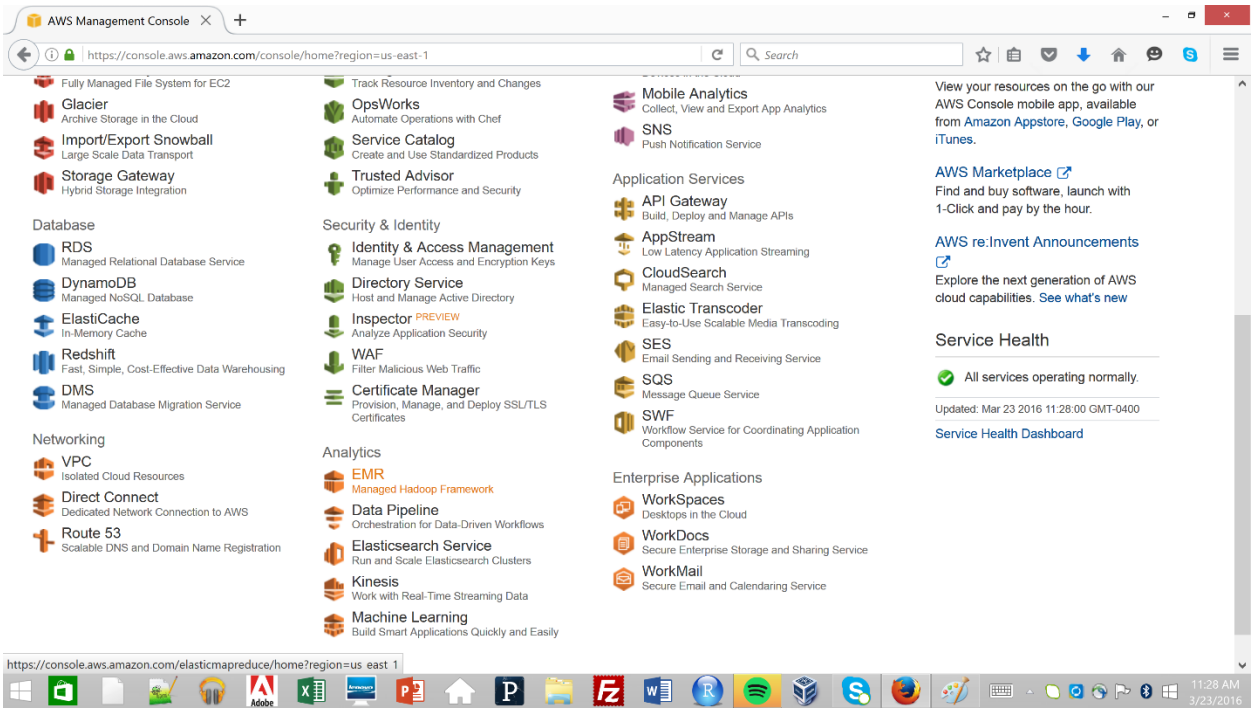


Once the bucket is created upload the `.jar` file and input files from local system or the VM. Click on Upload button and perform the upload of `wc.jar` and the input files. In this case, input files reside in ***uspres*** bucket file. Please refer to the following screenshot:

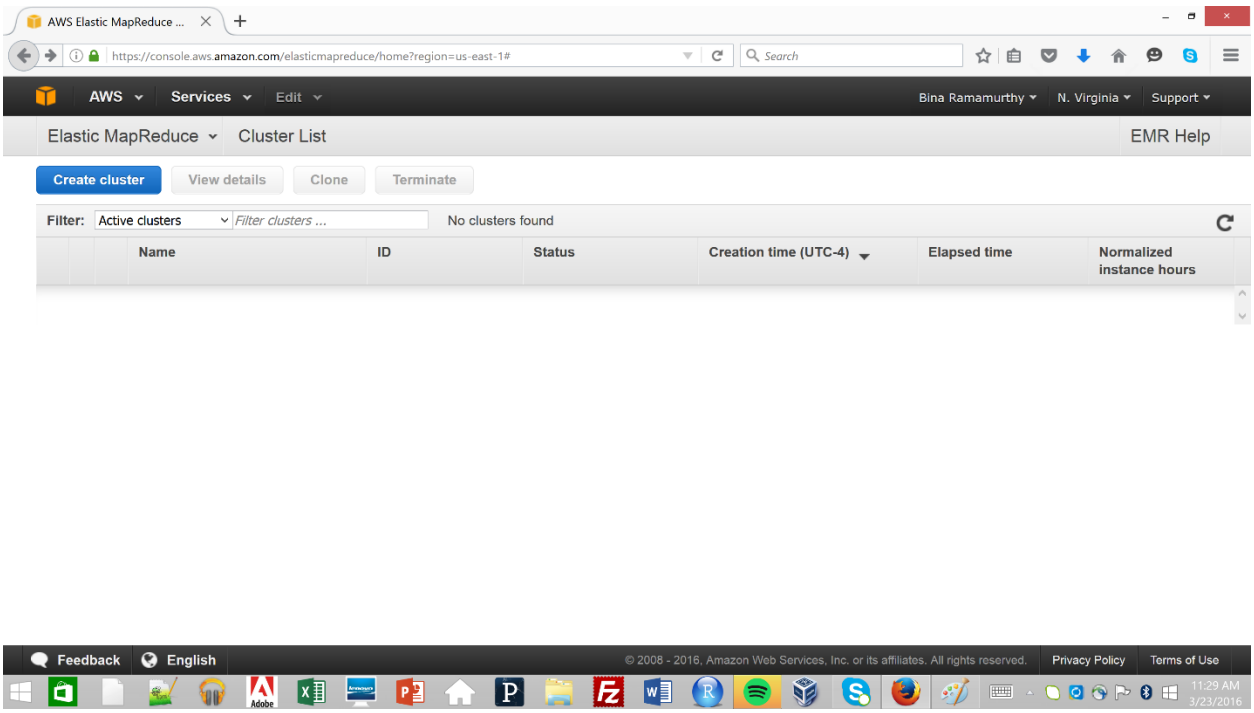


4 CONFIGURATIONS

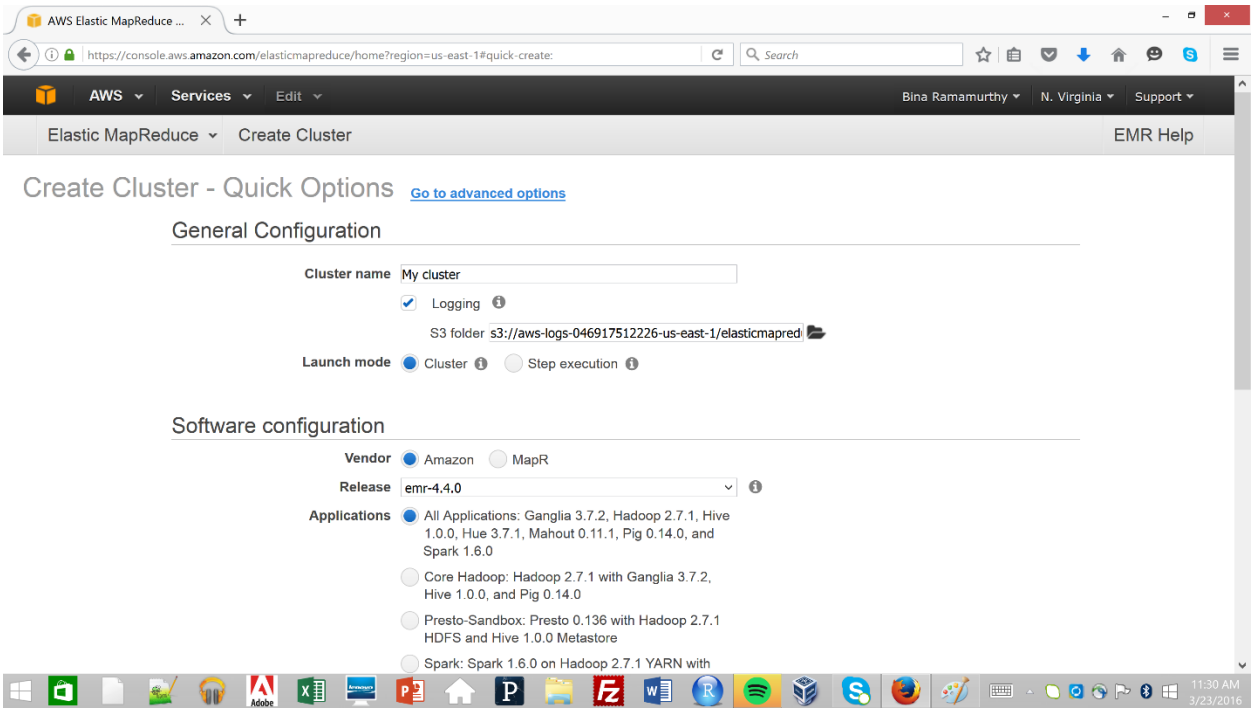
For configuration click on the AWS Service Cube icon present on the top left of the page. Once the page opens click on **EMR** residing in **Analytics** section. Please refer to the following screenshot:



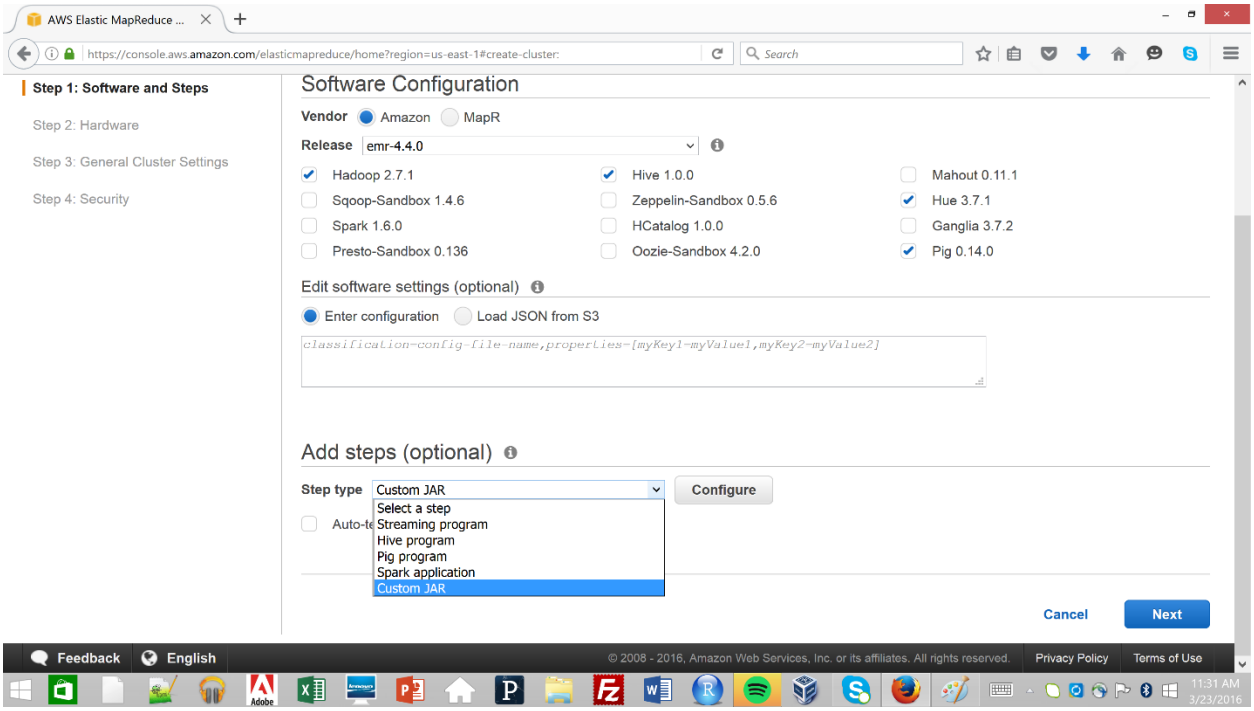
Here click on Create Cluster. Please refer to the following screenshot:



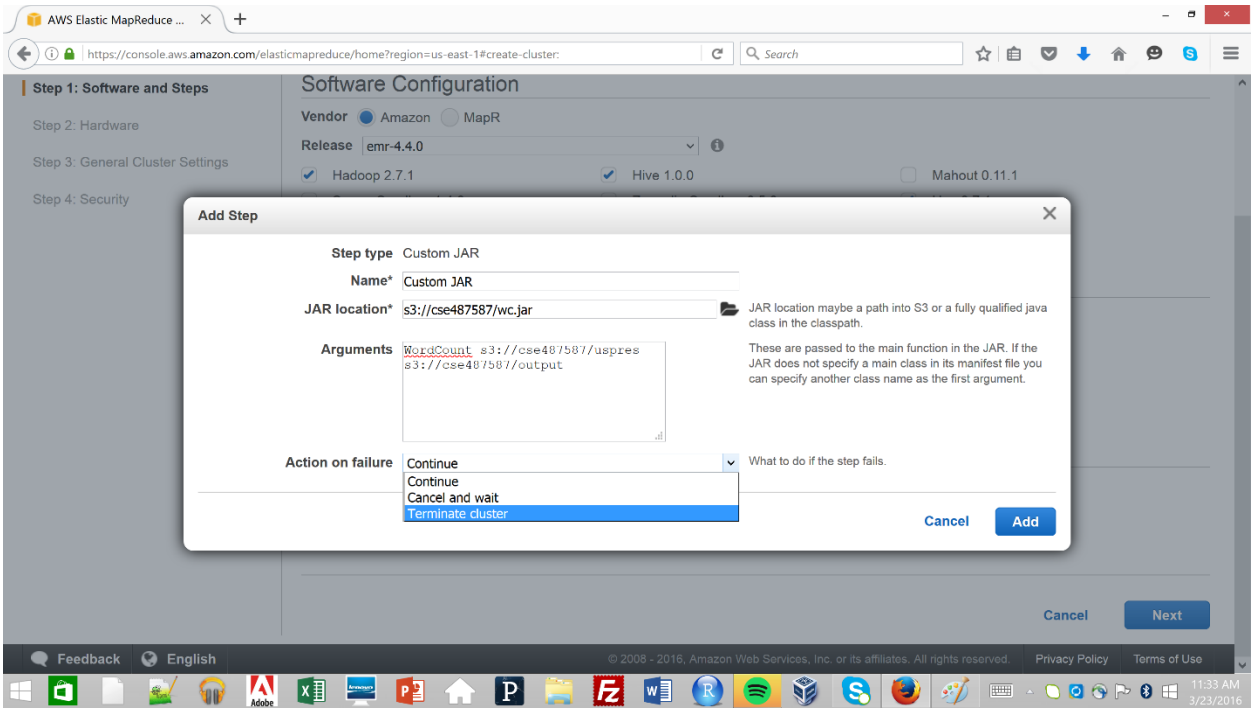
On opening a new web page, please click on the Advanced Options. Please refer to the following screenshot:

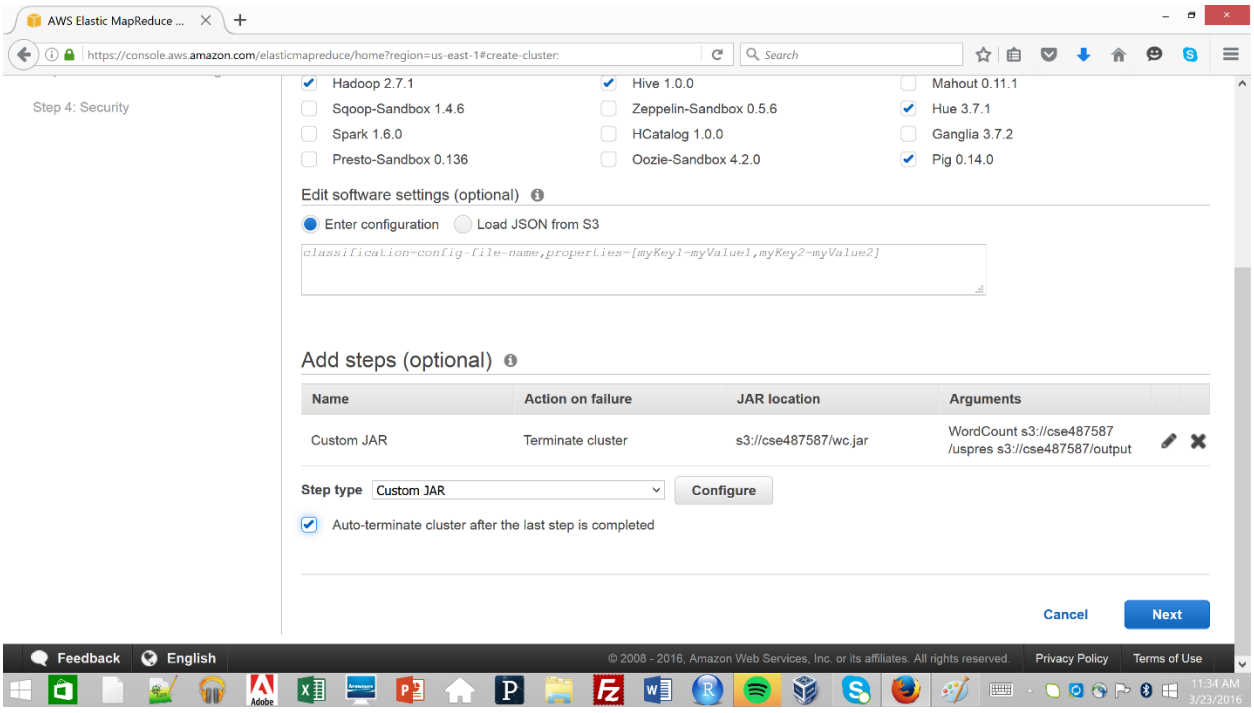


Once in the advanced options, go to section **Add Steps** and select **Step Type** to Custom JAR. Please refer to the following screenshot:

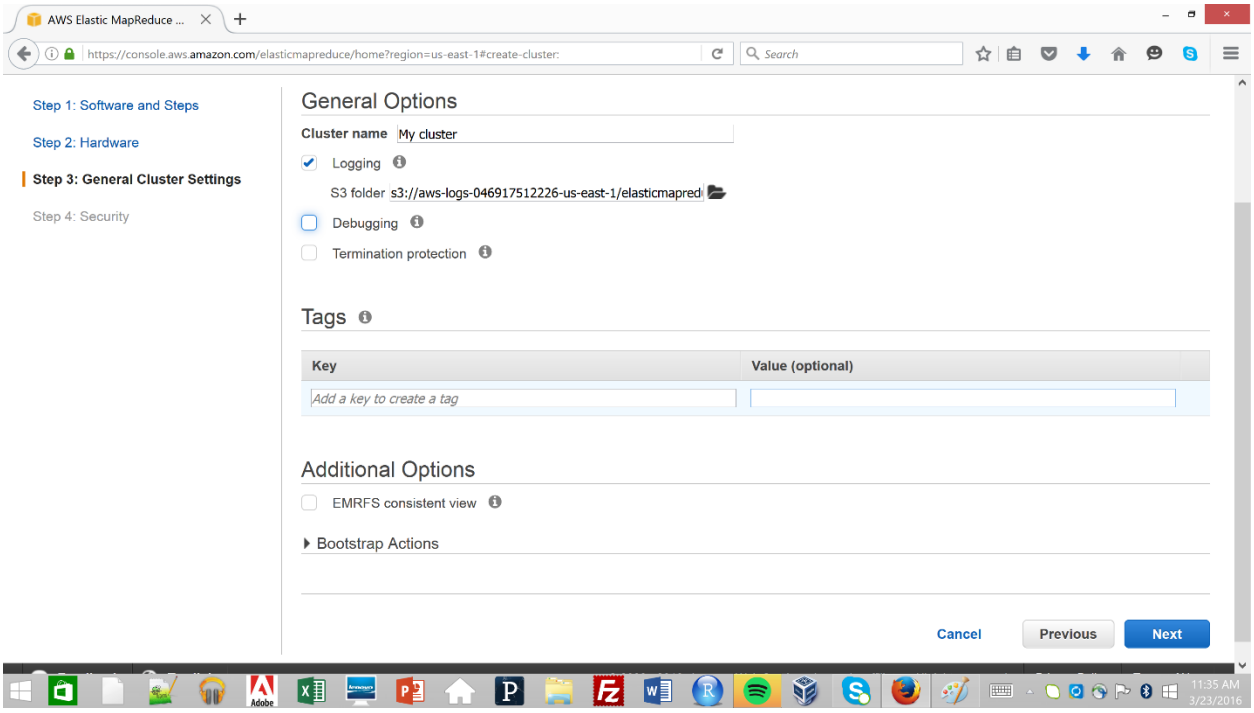


Later click on **Configure** button, it will open another popup. Fill in the details by providing the JAR location in S3 and corresponding arguments. Please ensure that **Action on Failure** has been selected to "*Terminate cluster*" [Why ?]. Moreover, please check "Auto-terminate after the last step is complete". Please refer to the following screenshots:

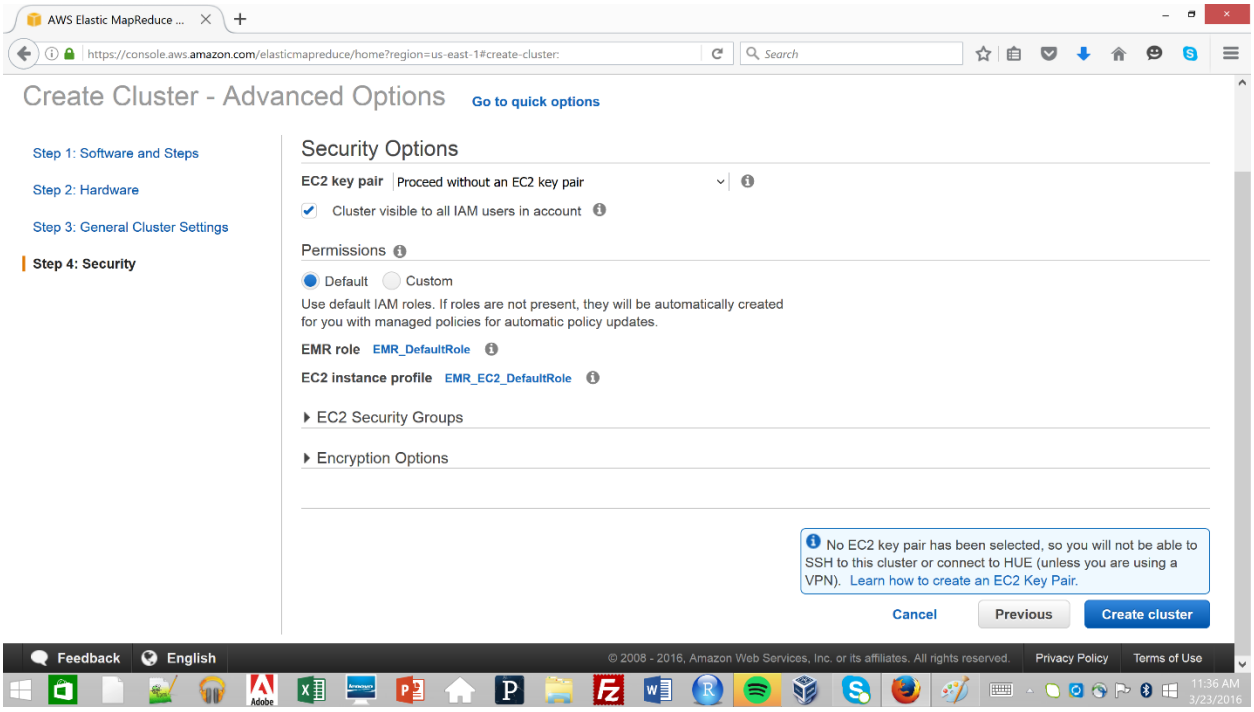




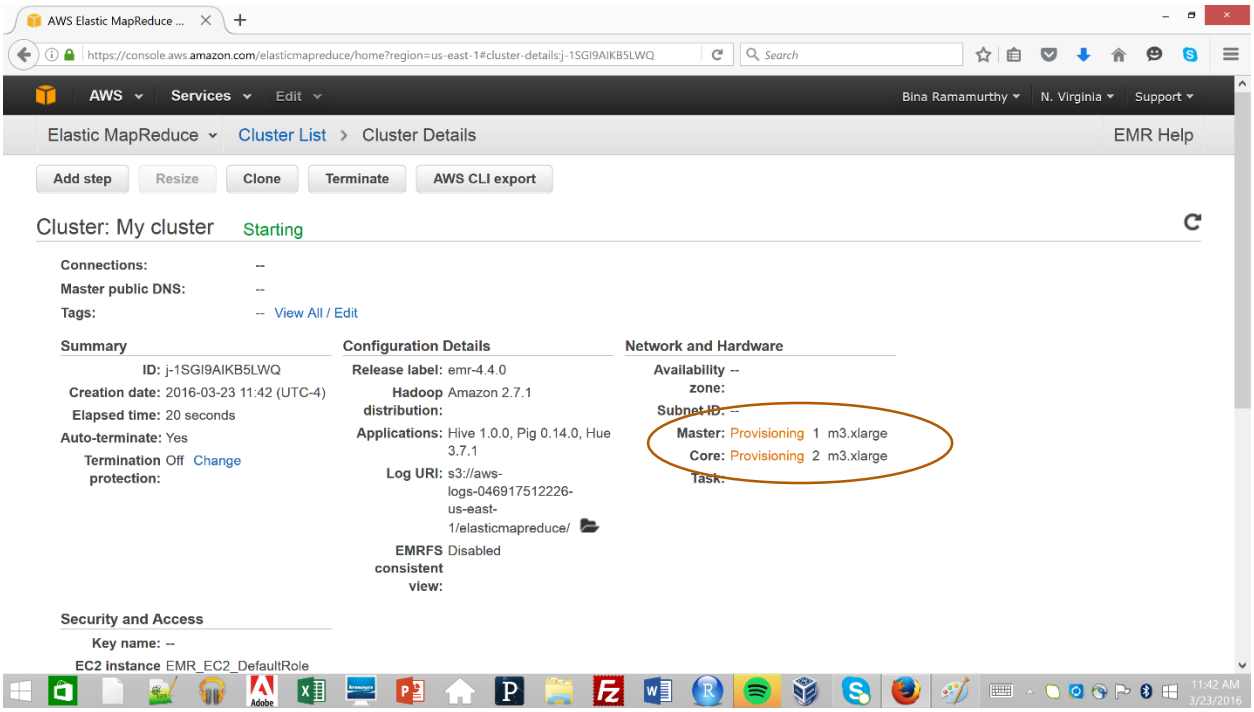
Click **Next** button and click **Next** button again skipping the Hardware configurations. Now after you reach **General Cluster Settings**, please uncheck **Debugging** and **Termination Protection**. Please refer to the following screenshot:



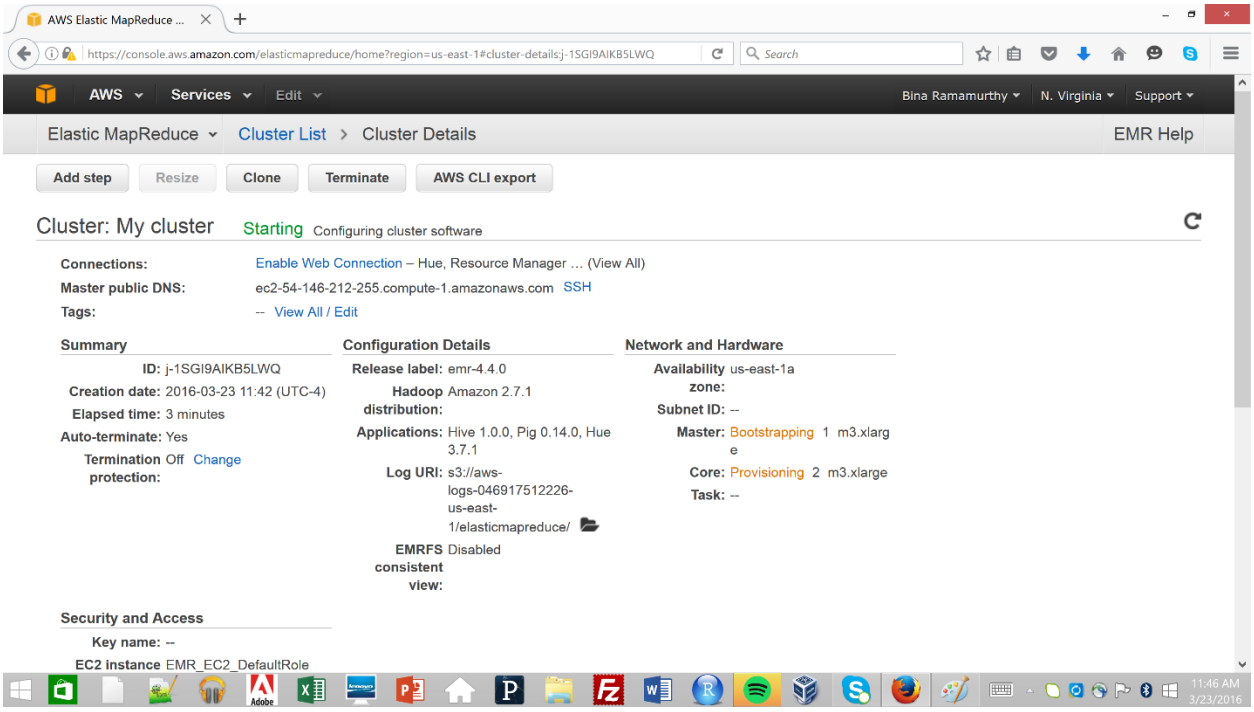
After these steps click on **Next** button. In this **Security** section, please ensure the following configuration and then click on **Create Cluster**.



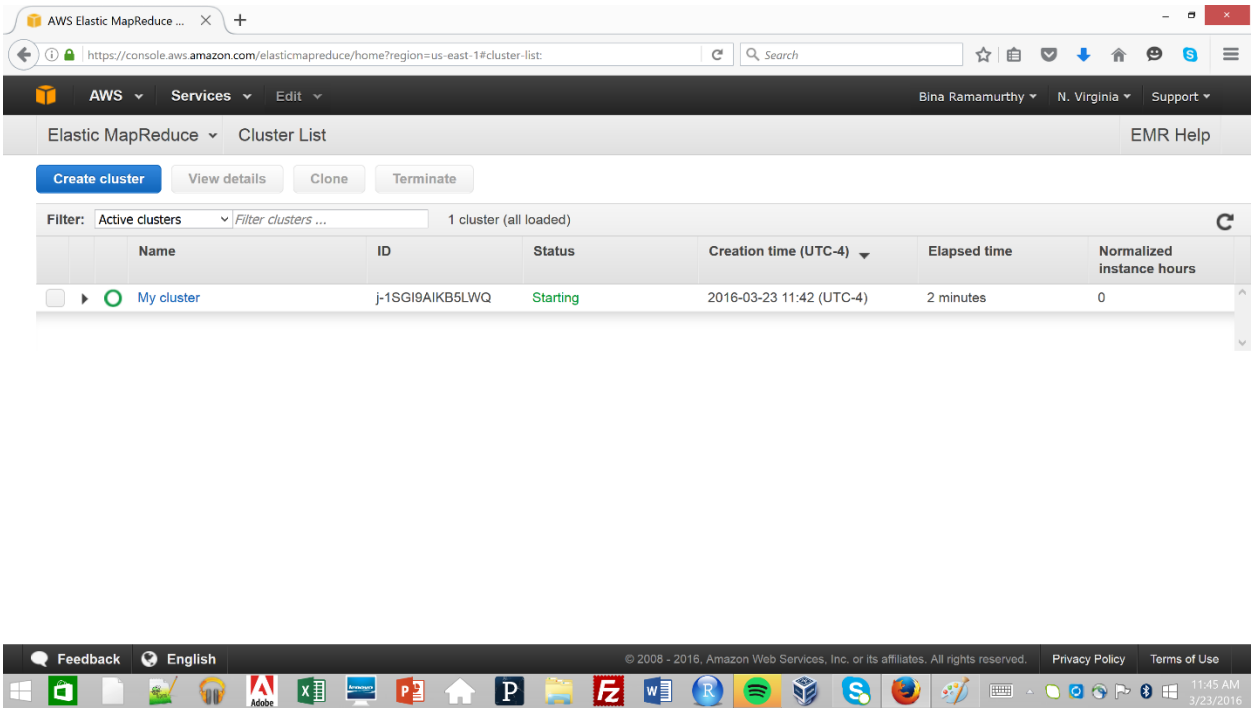
This will take you to a new page where details of the running cluster is shown. Please refer to the following image:



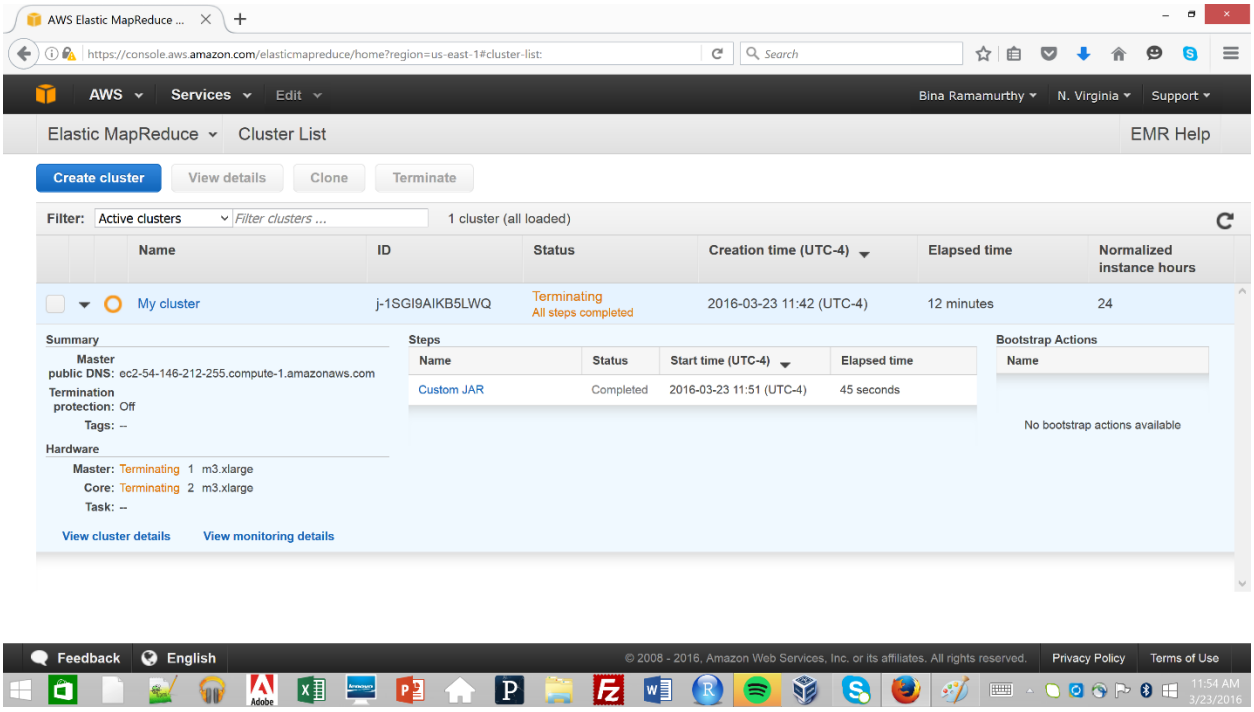
The subsequent processes typically take around 20 minutes. The next step involves the changing of the state of **Master** and **Core**, they change from *Provisioning* to *Bootstrap* to *Running* to *Terminated*. In the following screenshot there is a change recorded:



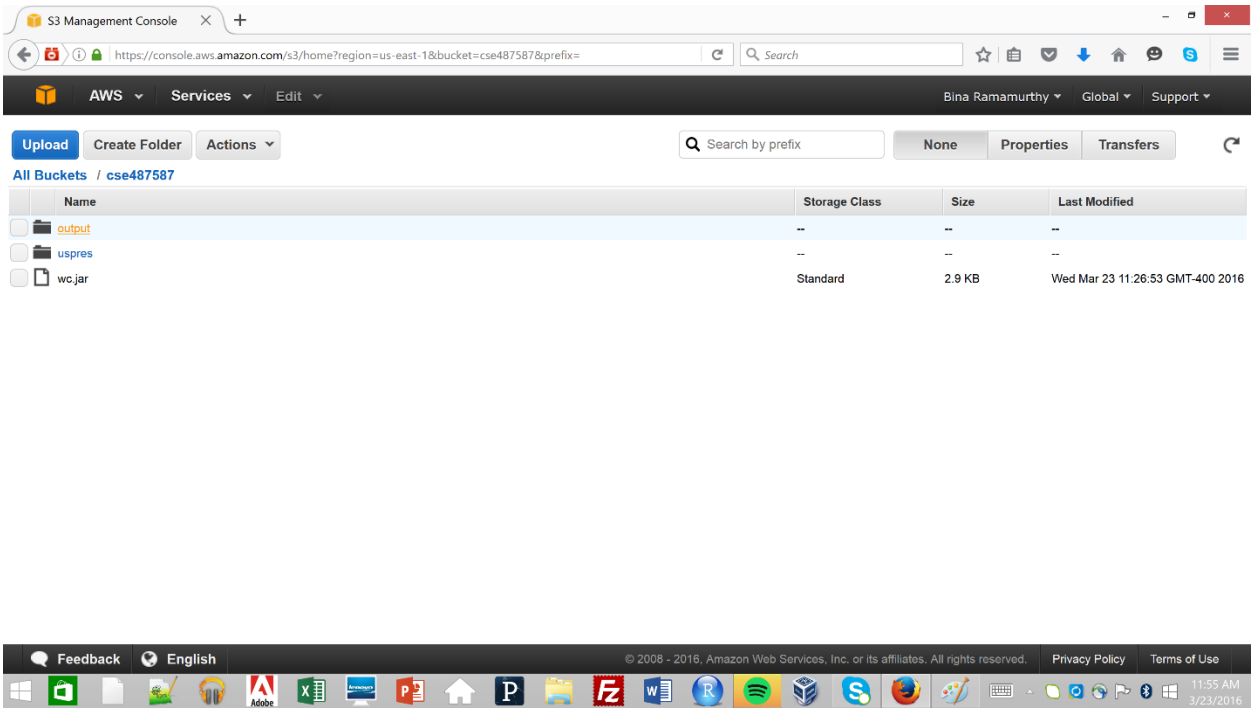
You can check the status of the active cluster by clicking on the **Cluster List**. The **Status** shows that the cluster has started running the application. Please refer to the following screenshot:



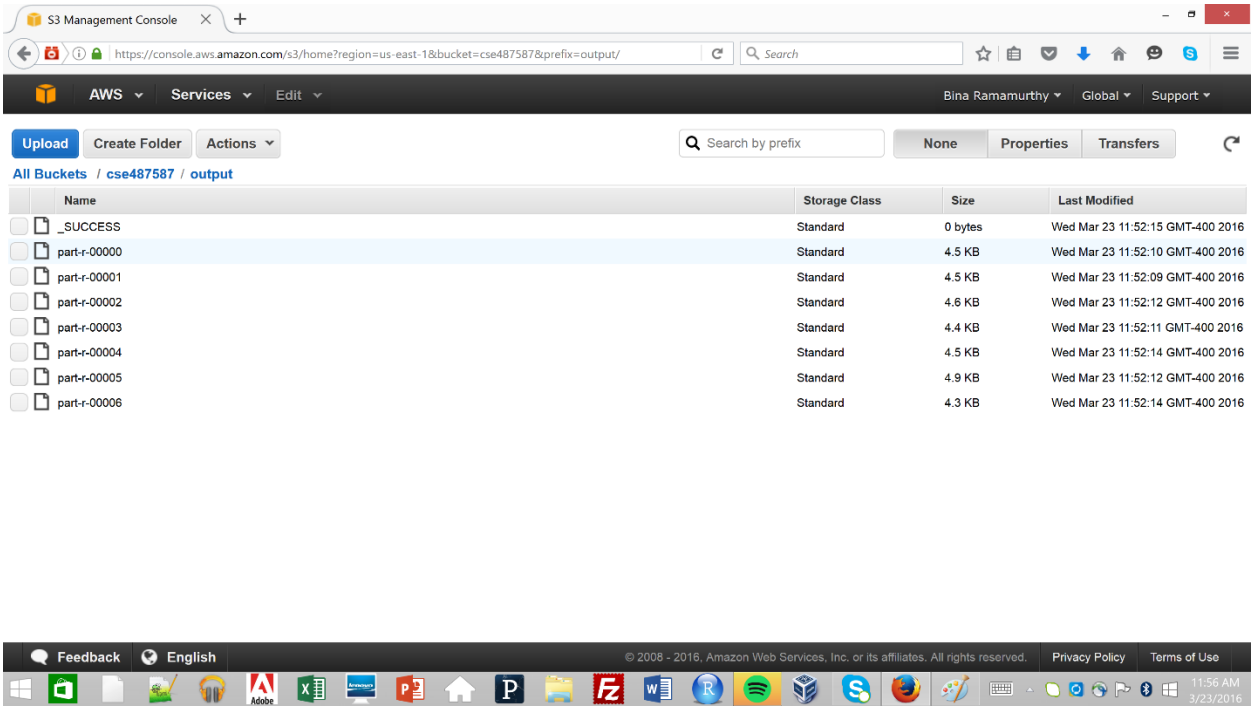
When the state changes to **Running** the color changes of the green icon changes to full green ( ). This is not shown in the screenshot because the cluster terminated soon enough to notice these changes. Later, when the cluster completed the run, you will see something similar to the following screenshot:



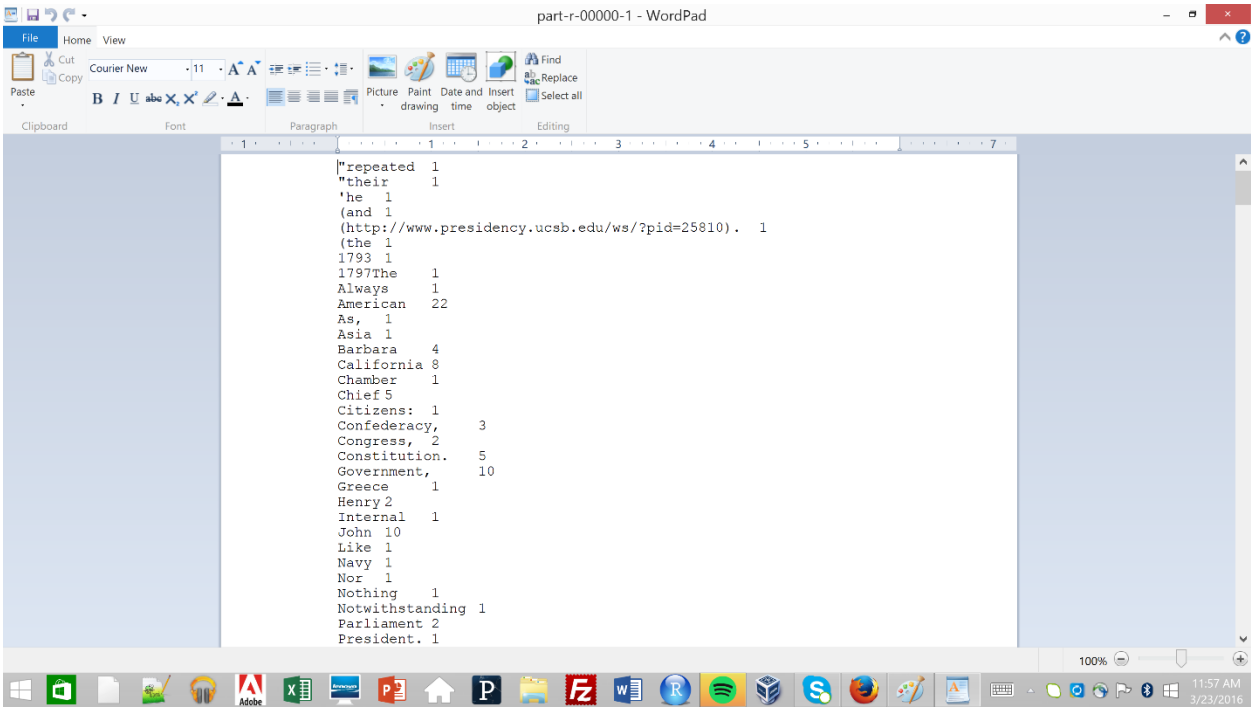
You can later check the output directories in the S3 Bucket which you created earlier. Please refer to the following screenshot:



Inside the **output** directory we can find the following list of files. Please refer to the following screenshot:



Once in this bucket you can check the details in the respective files. Here is a screenshot showing the the content in on these files:



5 NOTES AND WARNINGS

1. On creating an AWS Educate account using buffalo mail id, you would get benefits which include a free credit of $100 which can be used to configure and run clusters. Each run will cost around $1.
2. A bucket name must always be unique, it's a rule.
3. If auto terminate is unchecked and Terminate Cluster is ignored then your cluster might run repeatedly. This might not be in your best interests as each run operation costs around $1. It is better to continuously check the status of the active clusters, just like here below: