**Assignment 9**

In this assignment we are required to run the Wikipedia dataset on a distributed mode.

**Issues faced:**

* After setting up my cluster with physical nodes (with my friends’ PCs’), I was facing error with configuring SSH.
* Later, I decided to go with AWS EC2 (Elastic Computing cloud) cluster setup.
* I was facing memory issues (Java Heap error) with my cluster and so decided to switch on AWS EMR (Elastic MapReduce) setup.

**Setup Process:**

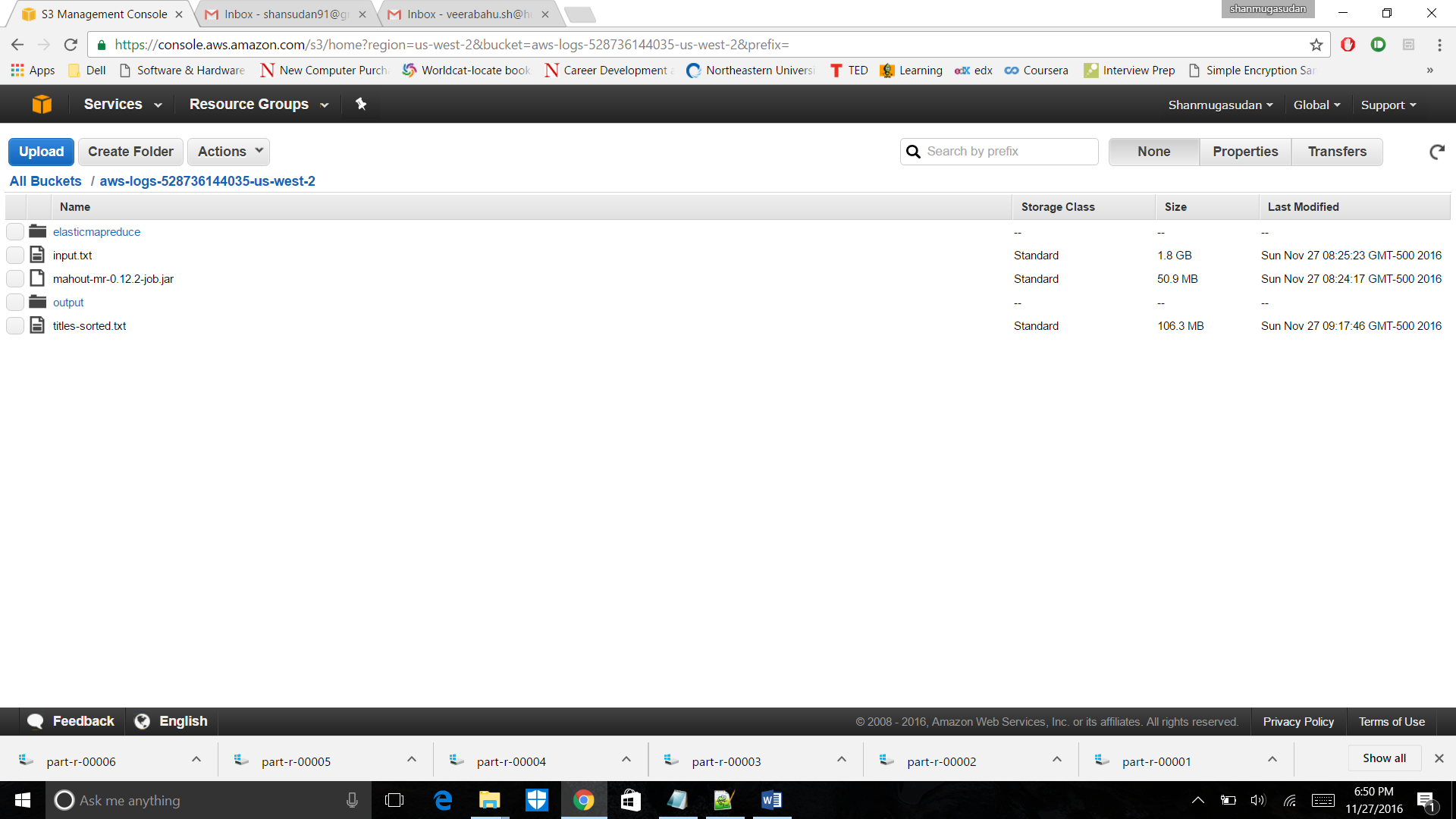
1. I signed up for AWS account and then created a AWS Educate account. Using my AWS Educate as a student, I received a promo code to use 100 free hours on AWS.
2. I activated my promo code, created a bucket on S3 (Simple Storage Service).

**Configuration parameter used:**

*org.apache.mahout.cf.taste.hadoop.item.RecommenderJob -s SIMILARITY\_COOCCURRENCE -i s3://aws-logs-528736144035-us-west-2/input.txt -o s3://aws-logs-528736144035-us-west-2/output*

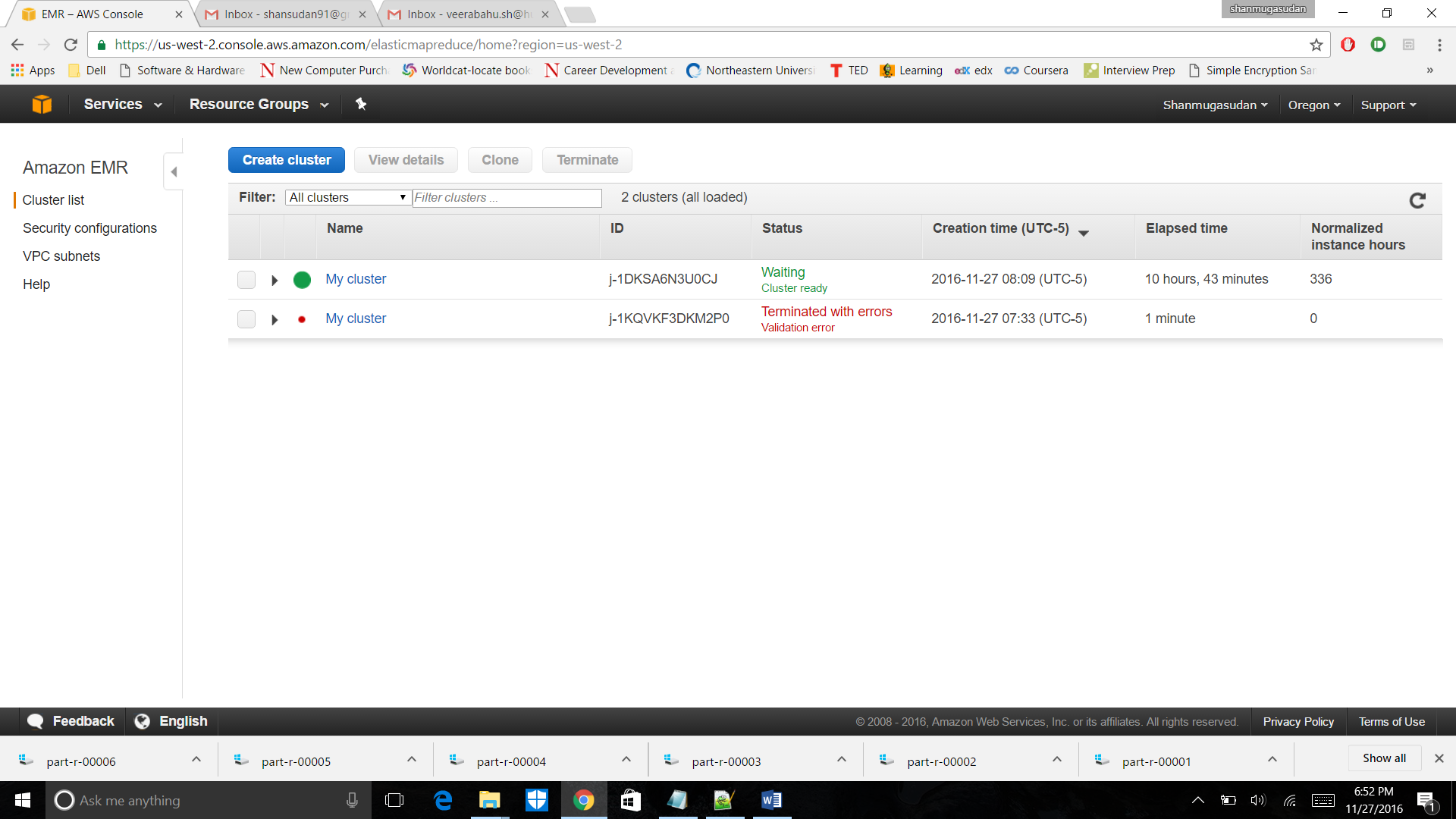
I have downloaded the mahout jar (mahout mr-0.12.2-job.jar) and pointed it under the custom JAR.

**Screenshot of my S3 bucket:**

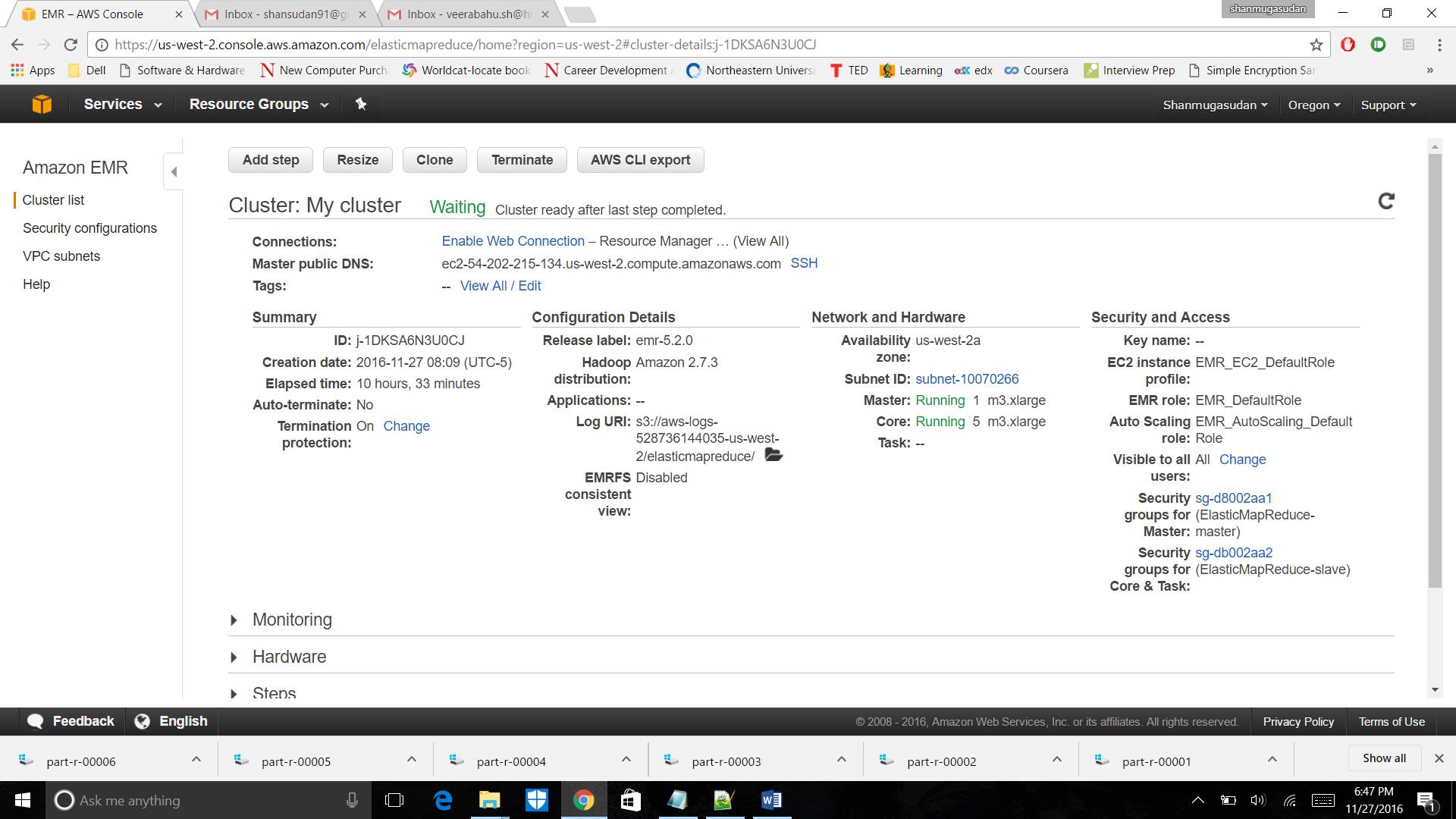


1. I created a EMR cluster as shown below. Initially I set it up with 2 core instances and 1 master instance, later I resized the core instances to 5 to speed up the process. As you could see, it took me around **8 hours to process with 2 core instances** and so I decided to auto-scale and the task was completed in **10 hours!!**

**Screenshot of my Cluster setup on AWS:**

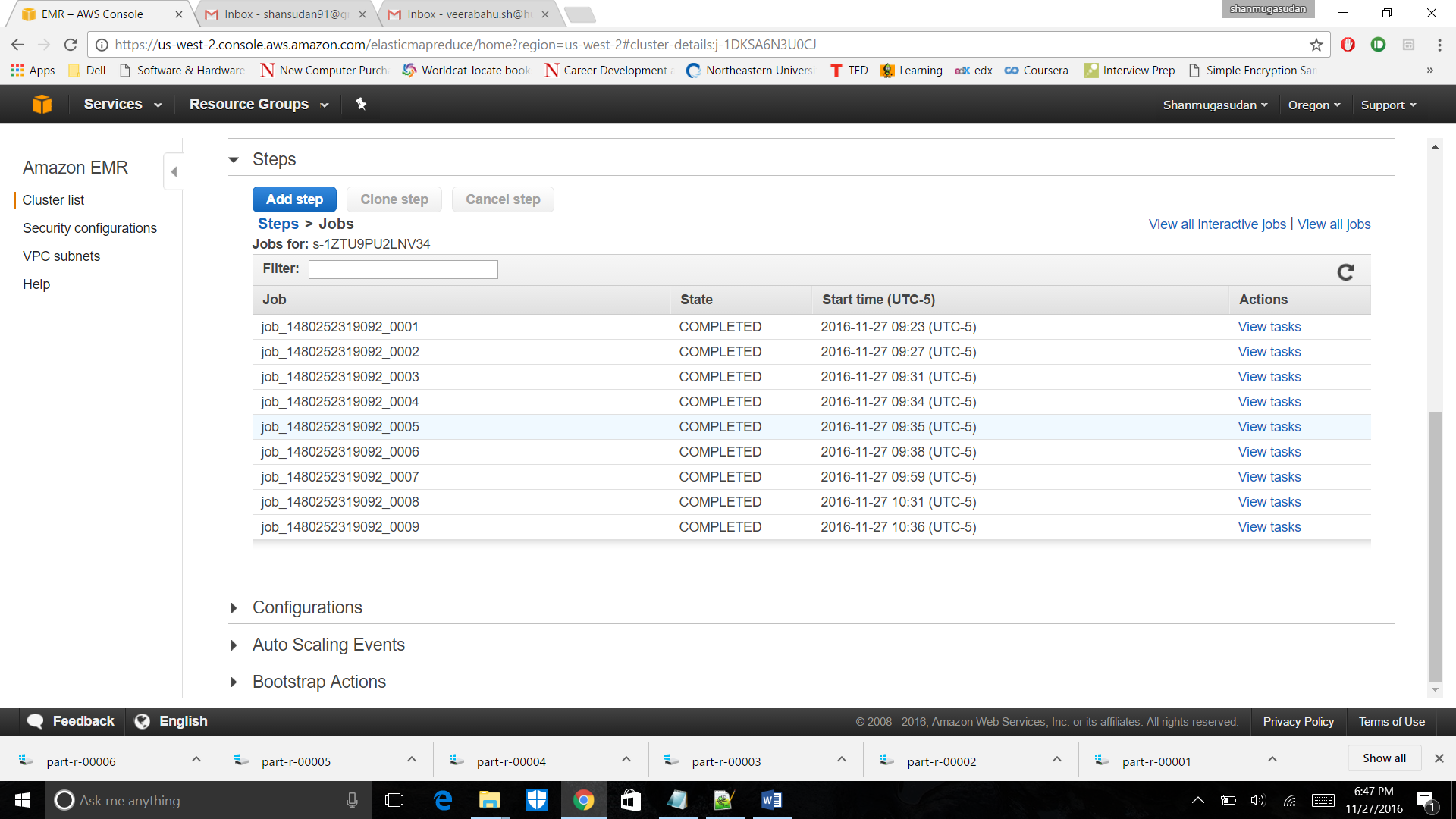


1. Below is the screenshot of my cluster configuration after resizing.

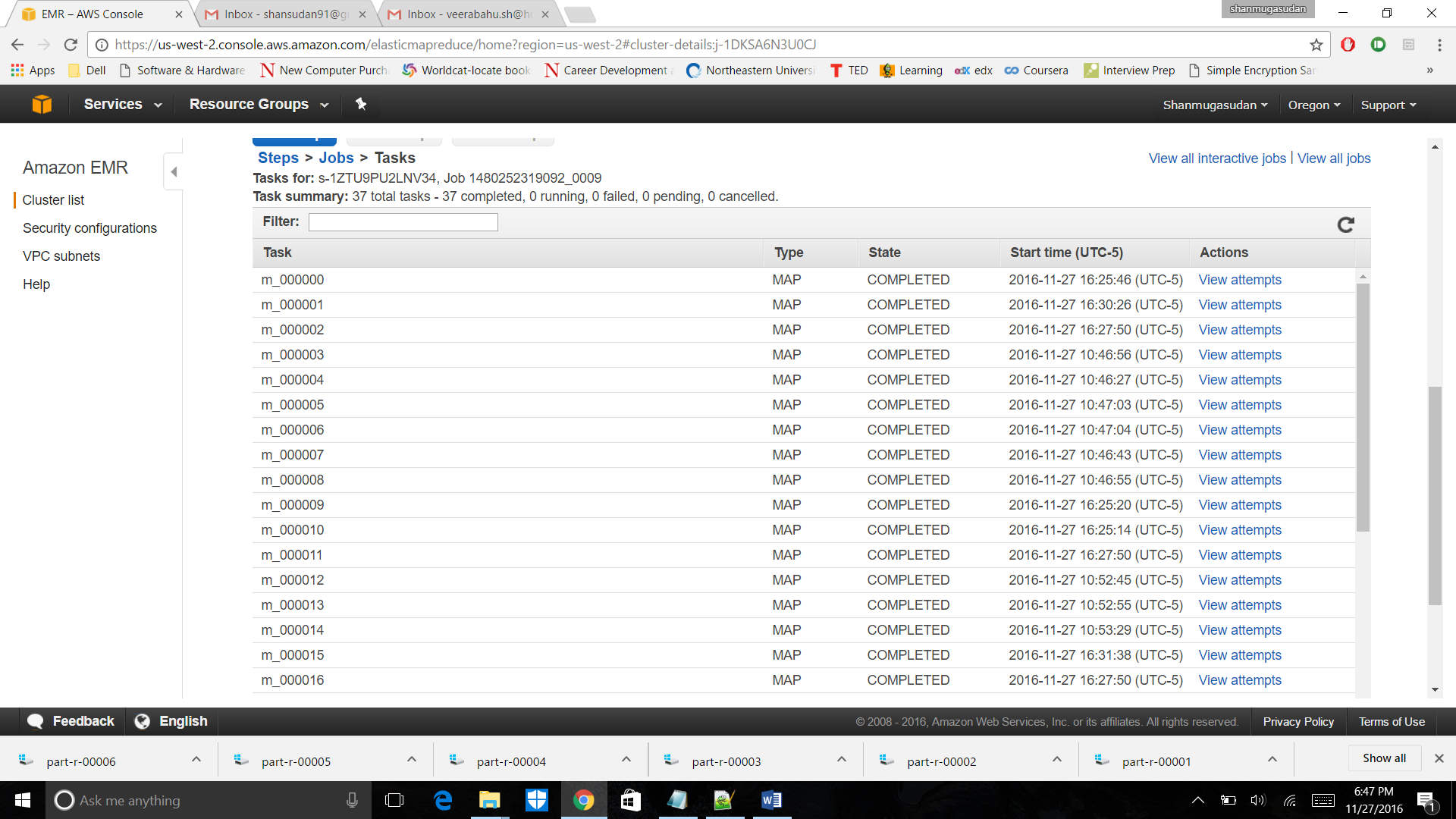
**Screenshot of Inside my Cluster on AWS**:

1. Under the steps tab, we can monitor the jobs created on the instances. By selecting “*view tasks”* we can view map and reduce tasks running on the jobs. Subsequent screenshots describe of the same.

**Sample Screenshot of the jobs:**

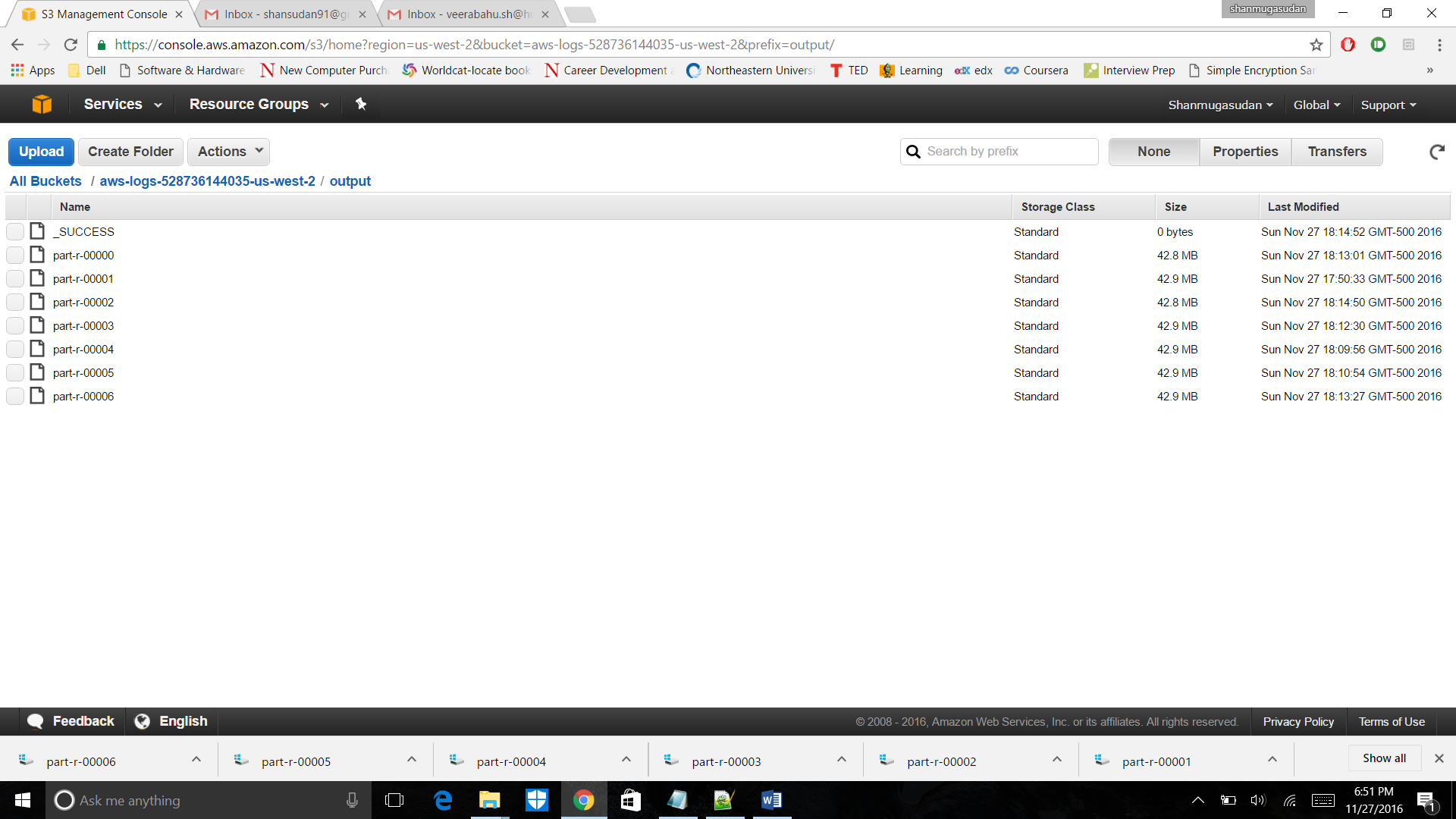


**Sample Screenshot of the tasks:**

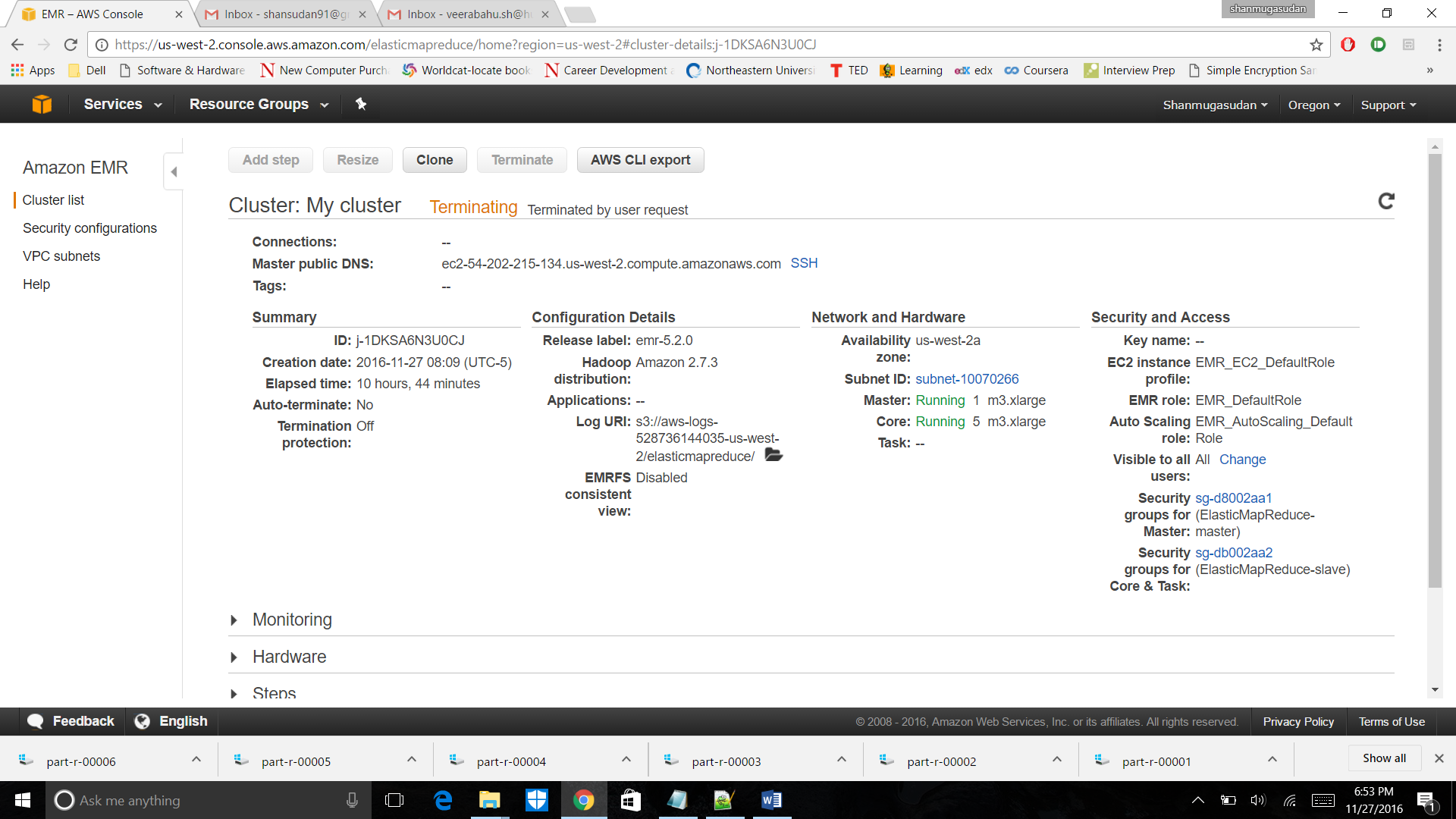


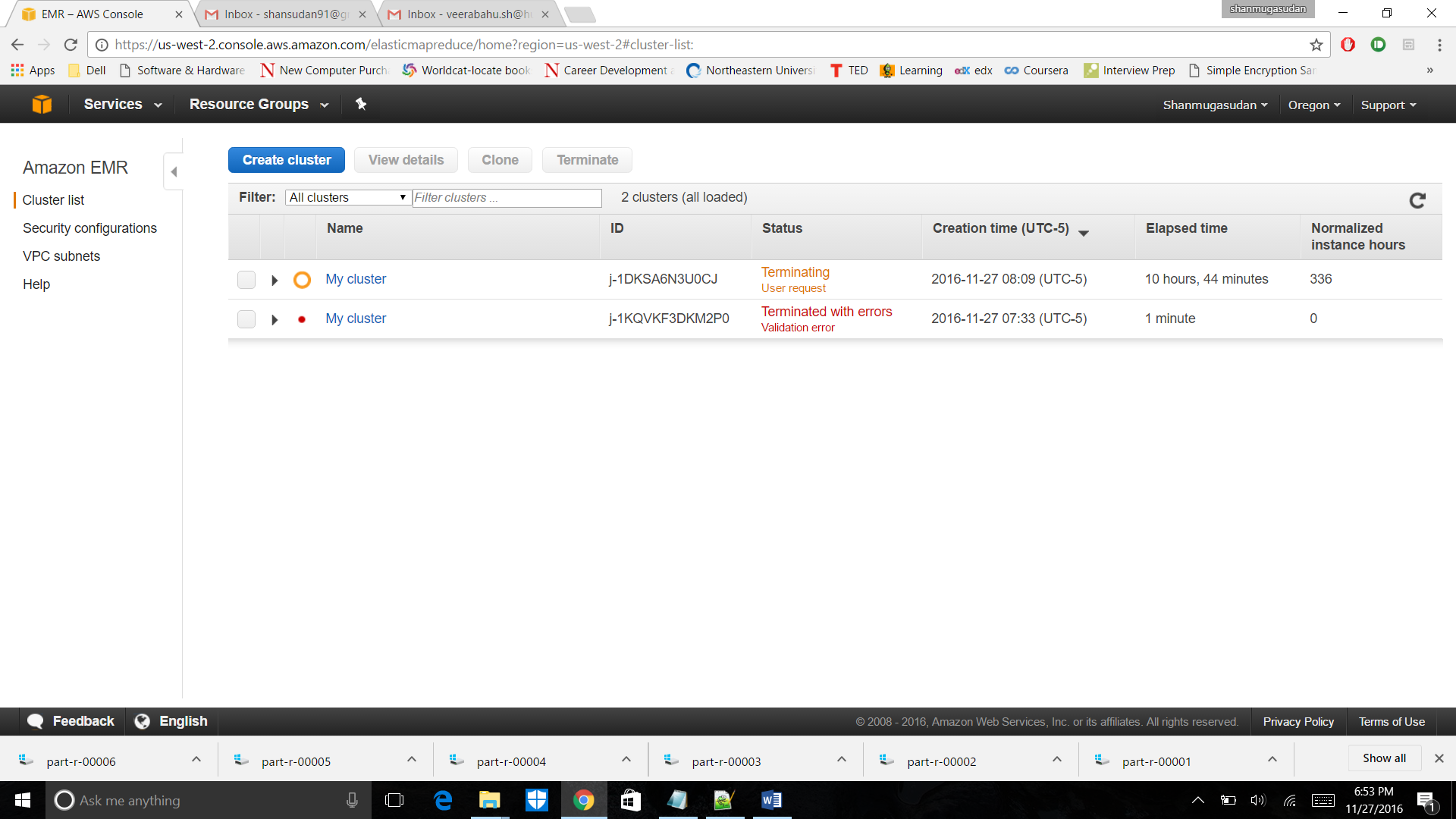
1. The output folder under my path, is shown below.

**Screenshot of Output folder:**



1. After processing the data, I have terminated the instances as show below.

**Screenshot of my Termination:**



1. Later, I have downloaded the files and below is a sample output file screenshot. As the output file is of 6 in count and each of size 43 MB, I wasn’t able to share the output file herewith.

**Sample Output file:**