**Project Documentation**

I need to implement mapreduce using python mrjob. I had provided with two data sets movies dataset and ratings dataset. I need to find the similarities between the movies based on the ratings given by the users. So to accomplish the task I had used 1 mapper and 4 reducers by using python mrjob package.

I applied the below mapper and reducer functions on small dataset.

**Mapper :**

In the mapper I had read the two datasets using the if else condition which helps to differ the two datasets. From the movies dataset I took movie id and movie title. From the ratings dataset I took movieid, userid and ratings.

**Reducer 1:**

In the reducer 1 I yield the required columns like user id as key and (movie\_title , rating) as a values. I passed the key, value pairs to the reducer 2.

**Reducer 2:**

I had used itertools package in python which helps me fining the combinations of movies which are rated by the same user i.e; I had emitted combinations on values of each key. Then from pair of combinations I yielded (title1,titl2),(rating1,rating2). So the key is titles and values is ratings. This results are passed to reducer 3 as key and values.

**Reducer 3:**

In this reducer I have combined the rating pairs of each key and yielded title and ratings.

output ex: (title1,title2),((2,3),(4,3),(3,4))

**Reducer 4:**

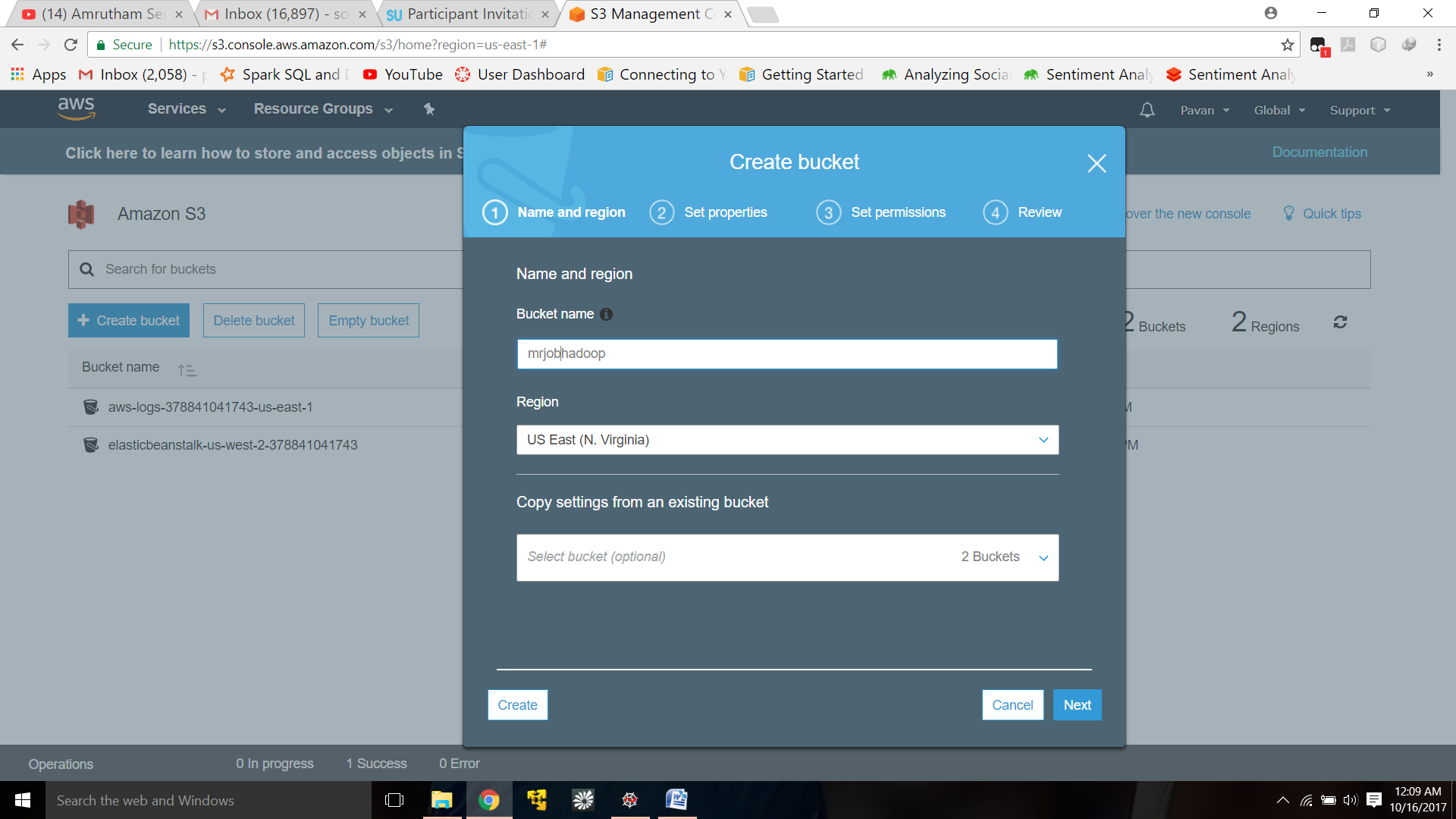
In reducer 4 I have formed to vectors using rating pairs. Using the numpy and scipy package i had calculated the correlation coefficient and cosine correlation.

Once doing the computation on all movies I had passed additional arguments in the code to customize the output. This arguments helps in getting the output of required movies, sorting based on similarities etc. Please find the below source code for small dataset. For lareg Dataset we used the same code but change the split function in code.

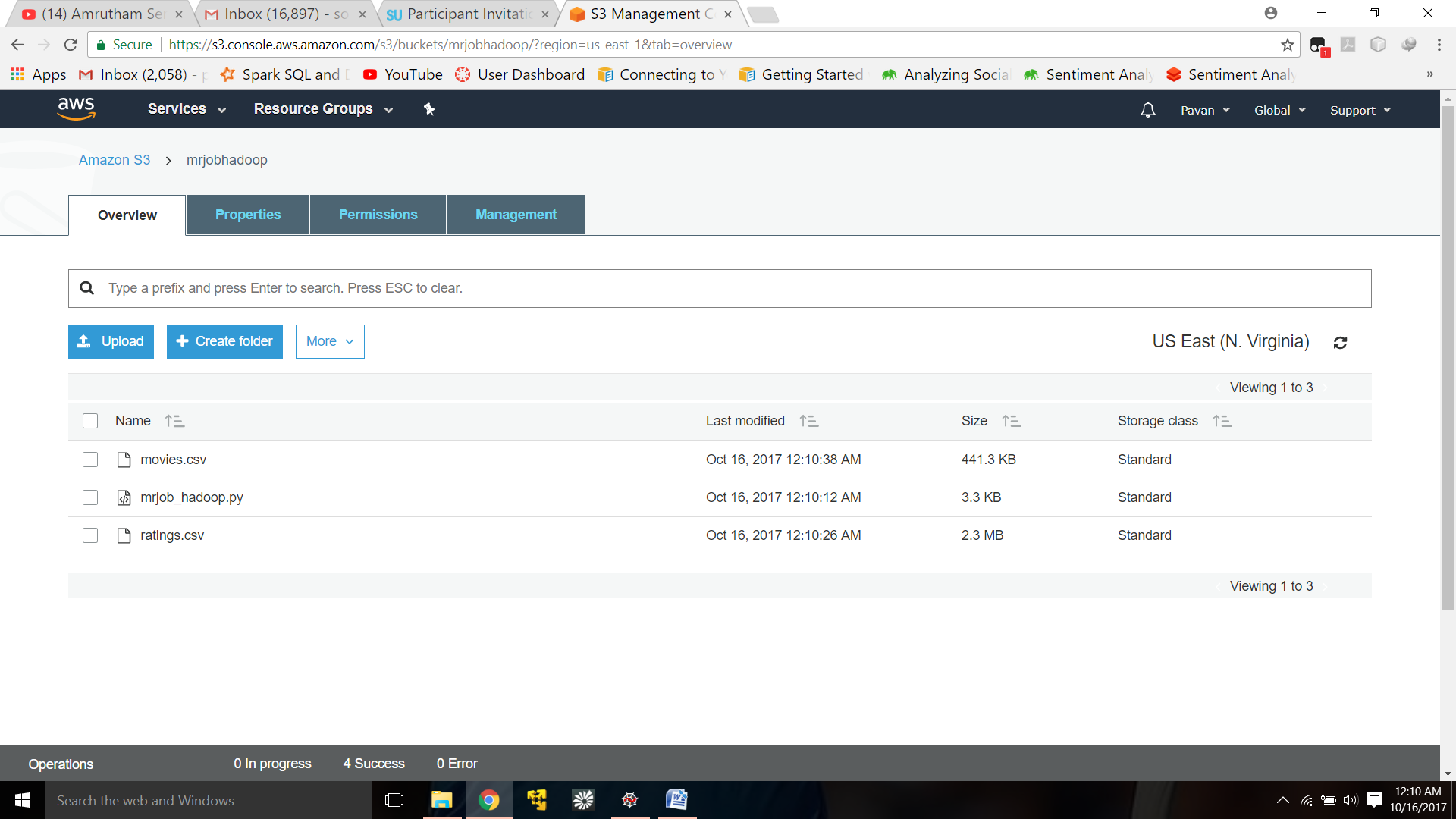
**Running on AWS:**

**note:** I was unable to find mrjob.conf file in my windows machine. So I tried in another way to run my python script on aws.

**step 1:**  Created s3 bucket.

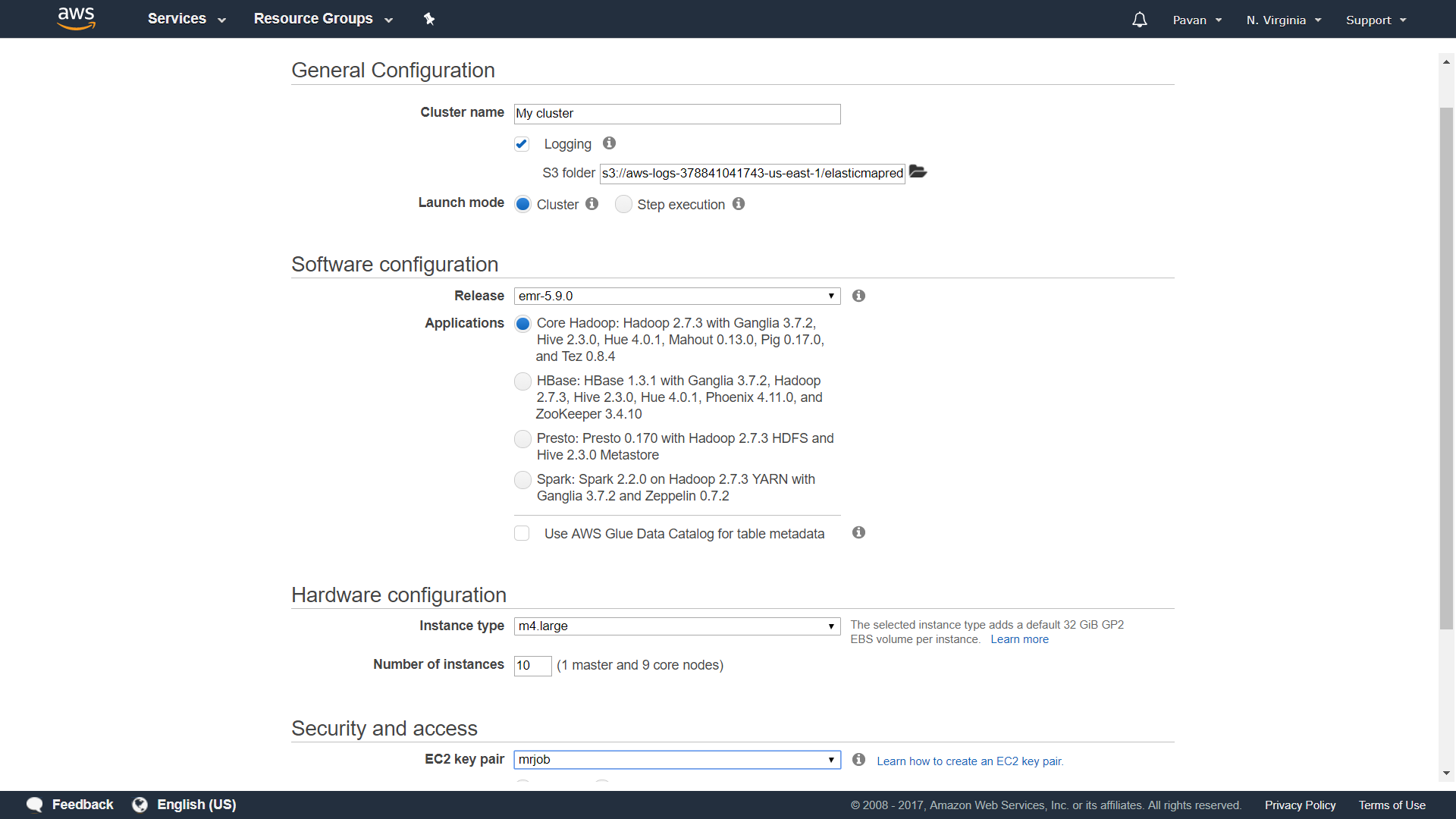


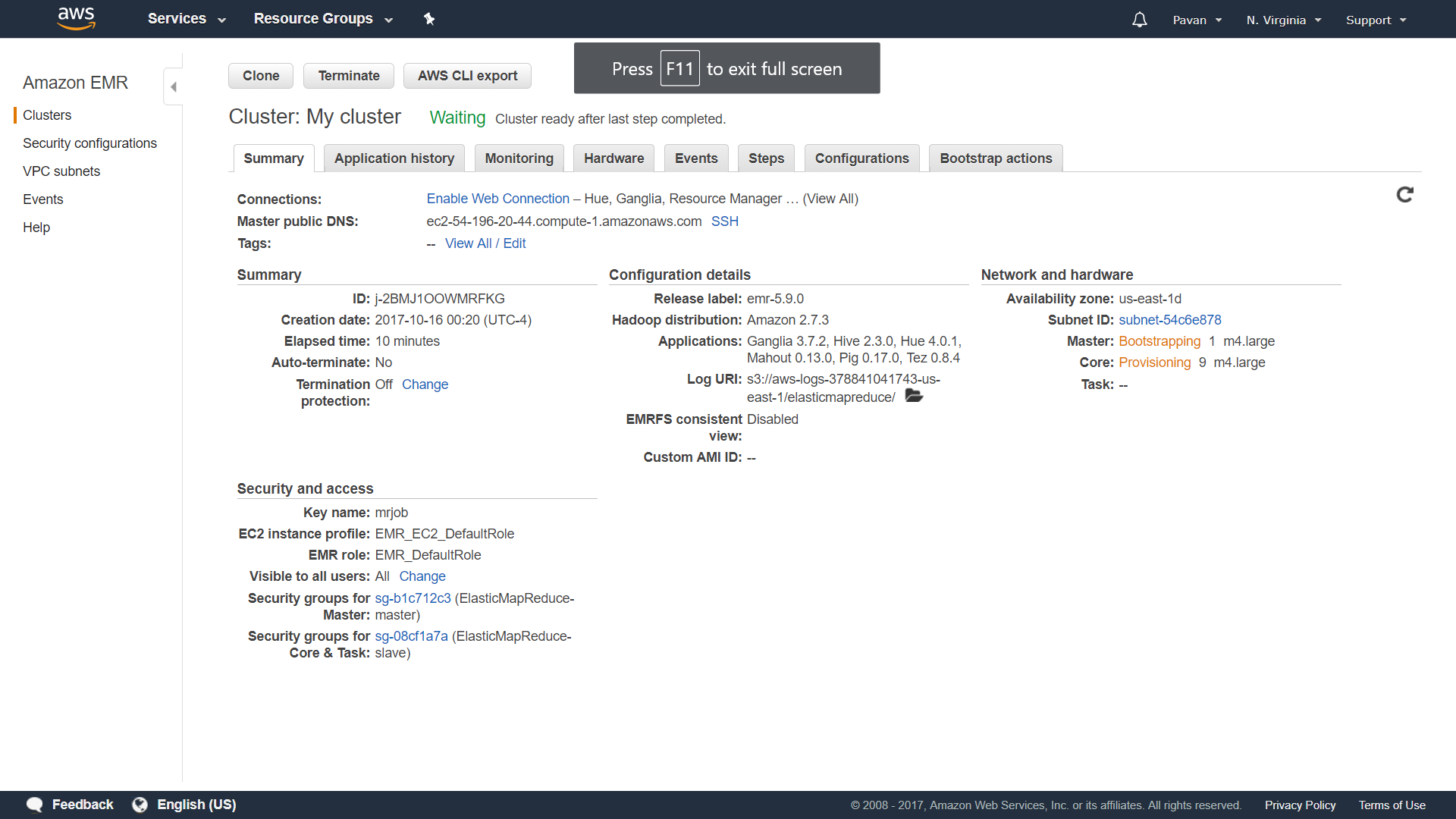
Uploaded required files into s3 bucket.

****

**Step 2:**

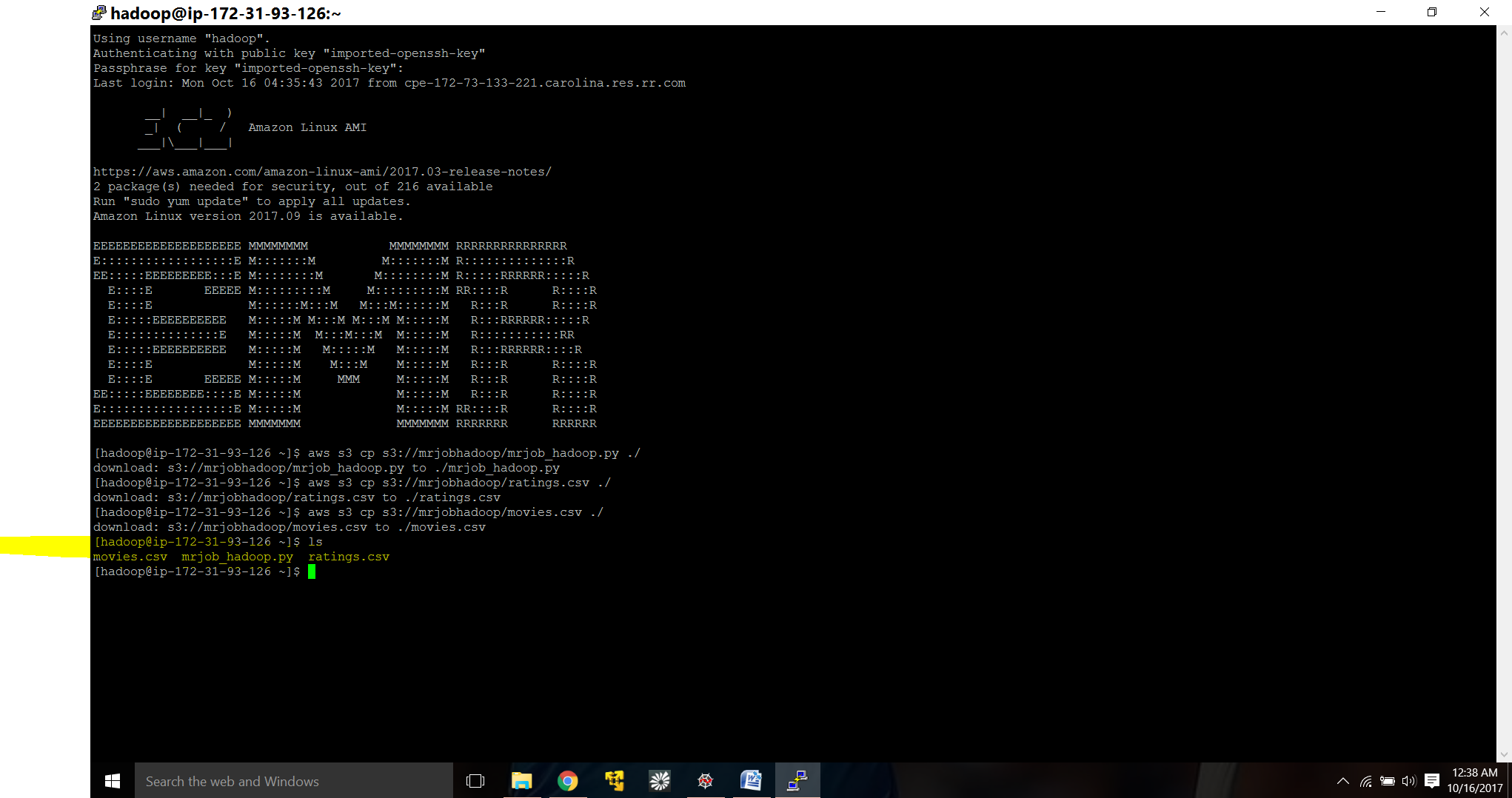
Created a cluster in AWS by selecting the required hadoop version, instance type, number of instances, security key as shown in the figure.





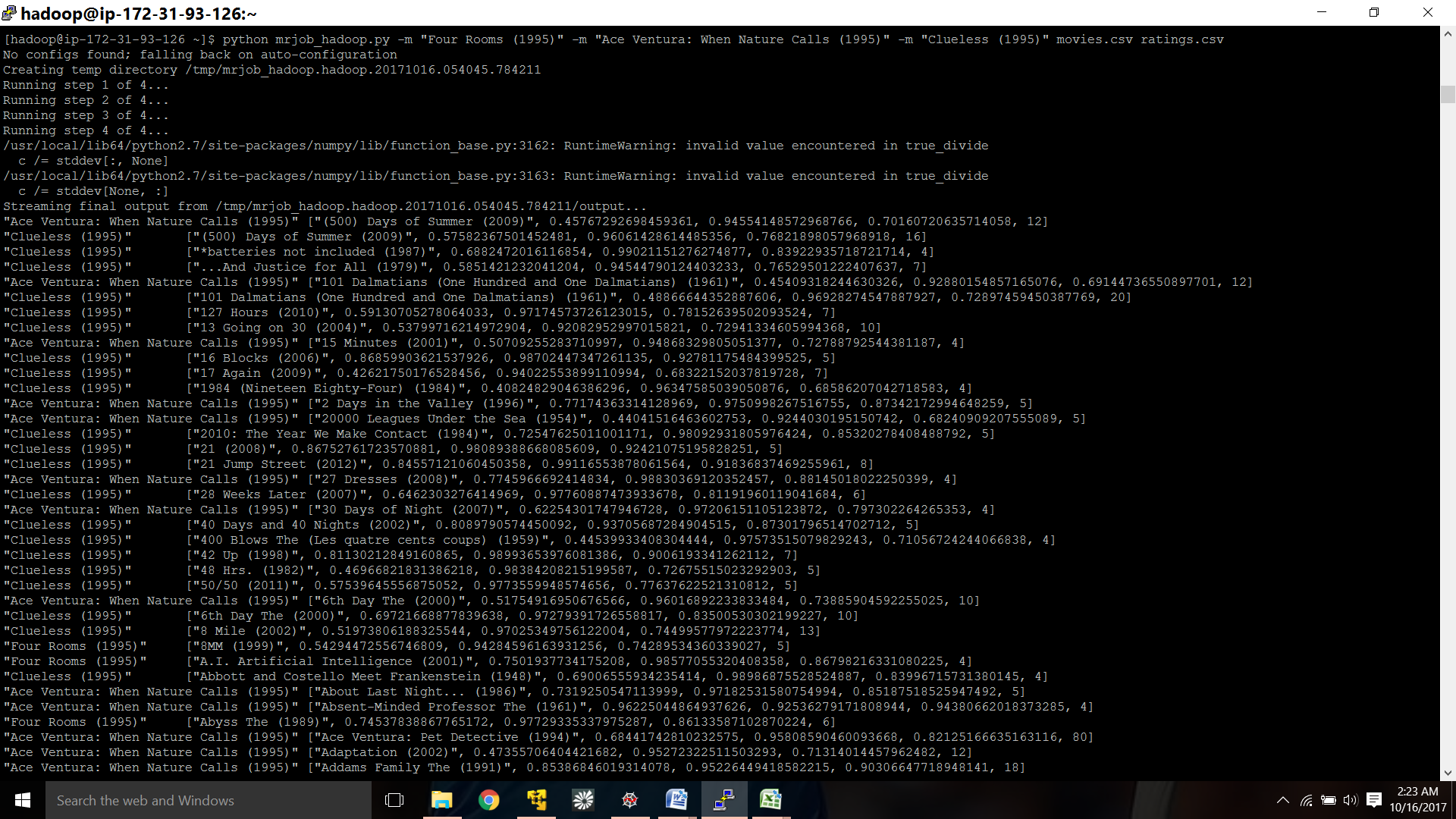
**step 3:**

Connected to the master node using putty. Copied the required script and data files from s3 to master node instance as shown below.



**Step 4:(output)**

Executed the similarity between the movies using below command. Please find the below screenshot for output.



output continuation:

