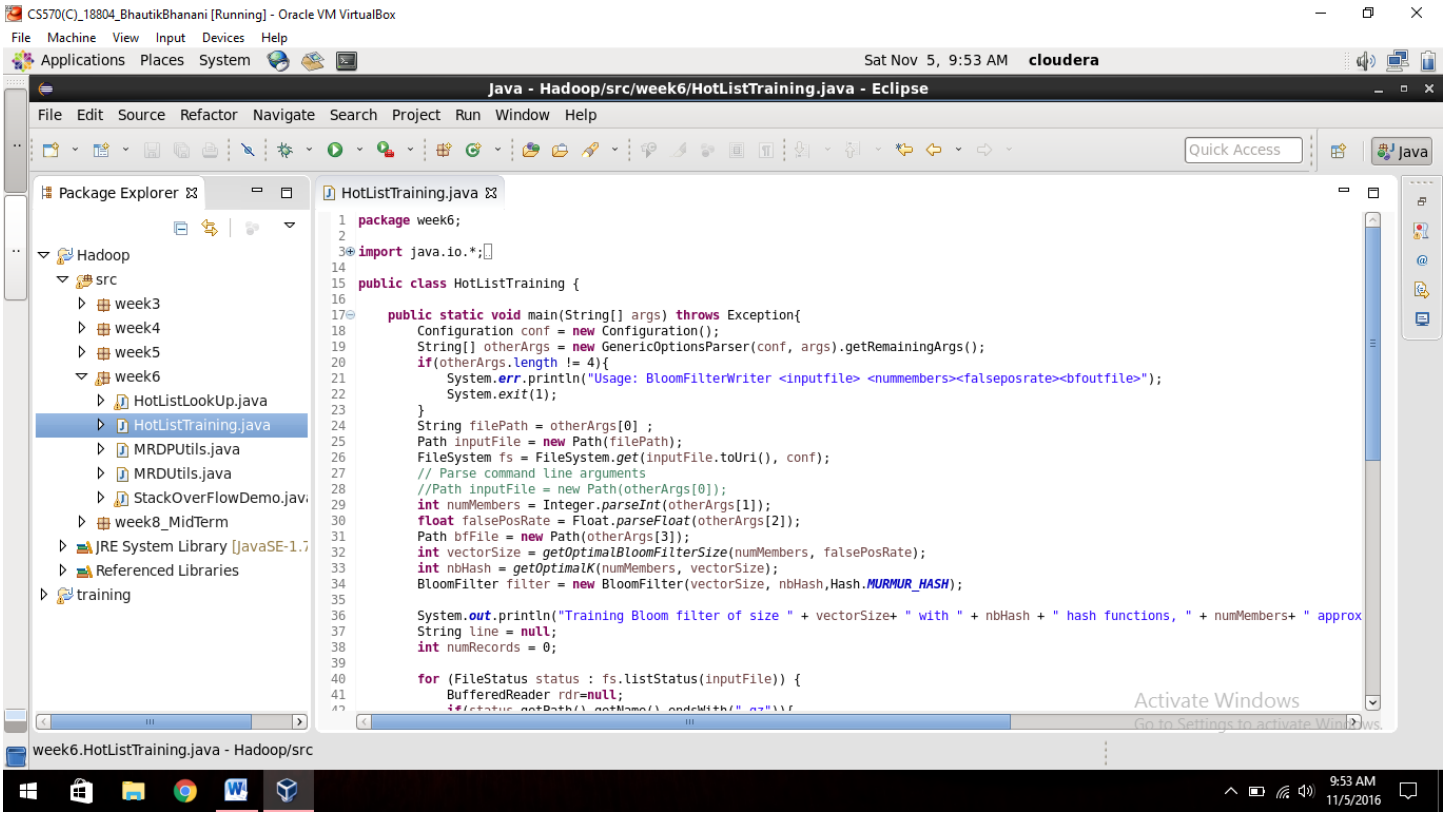
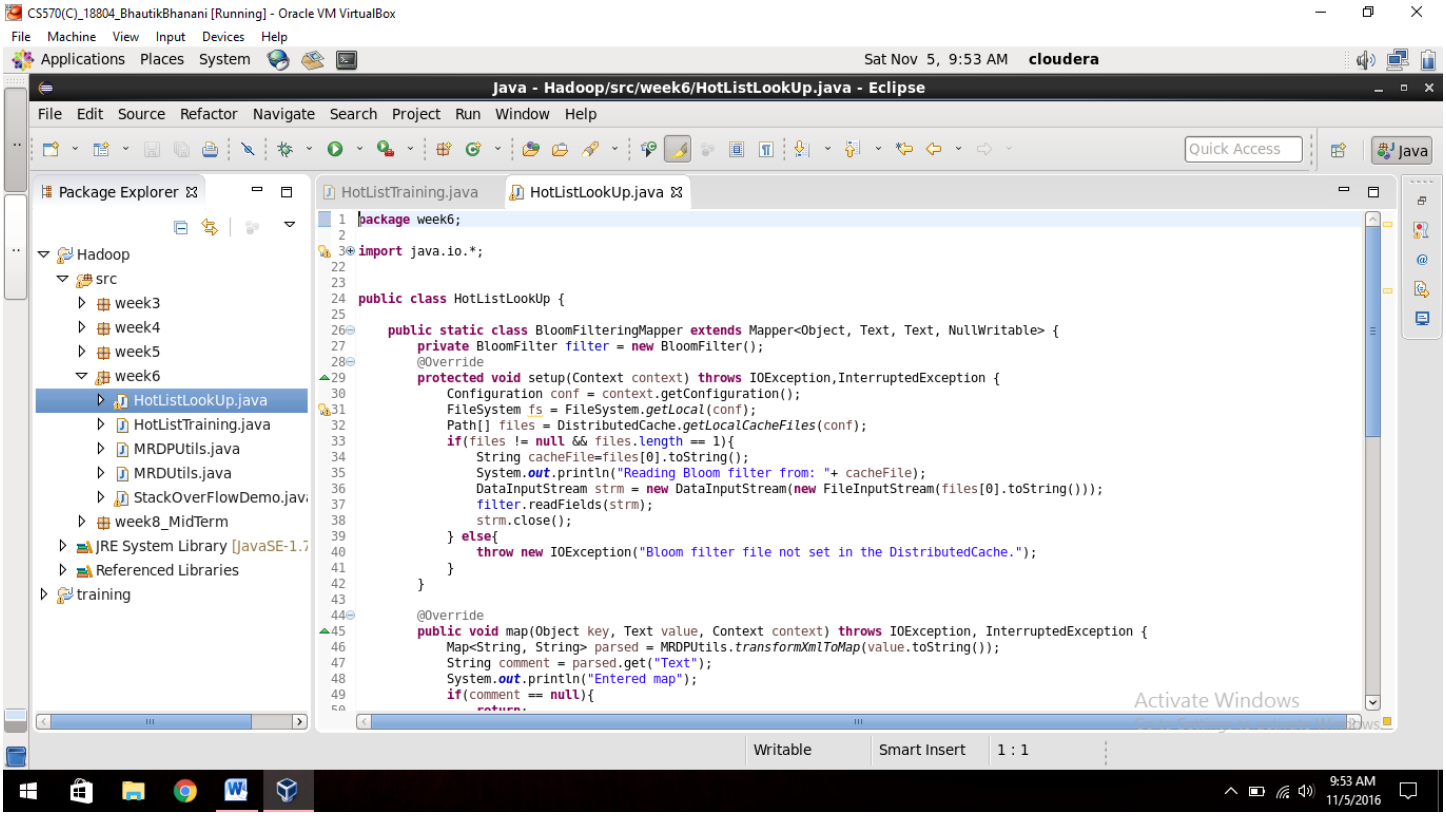
MapReduce to implement Bloom Filter for Hot List

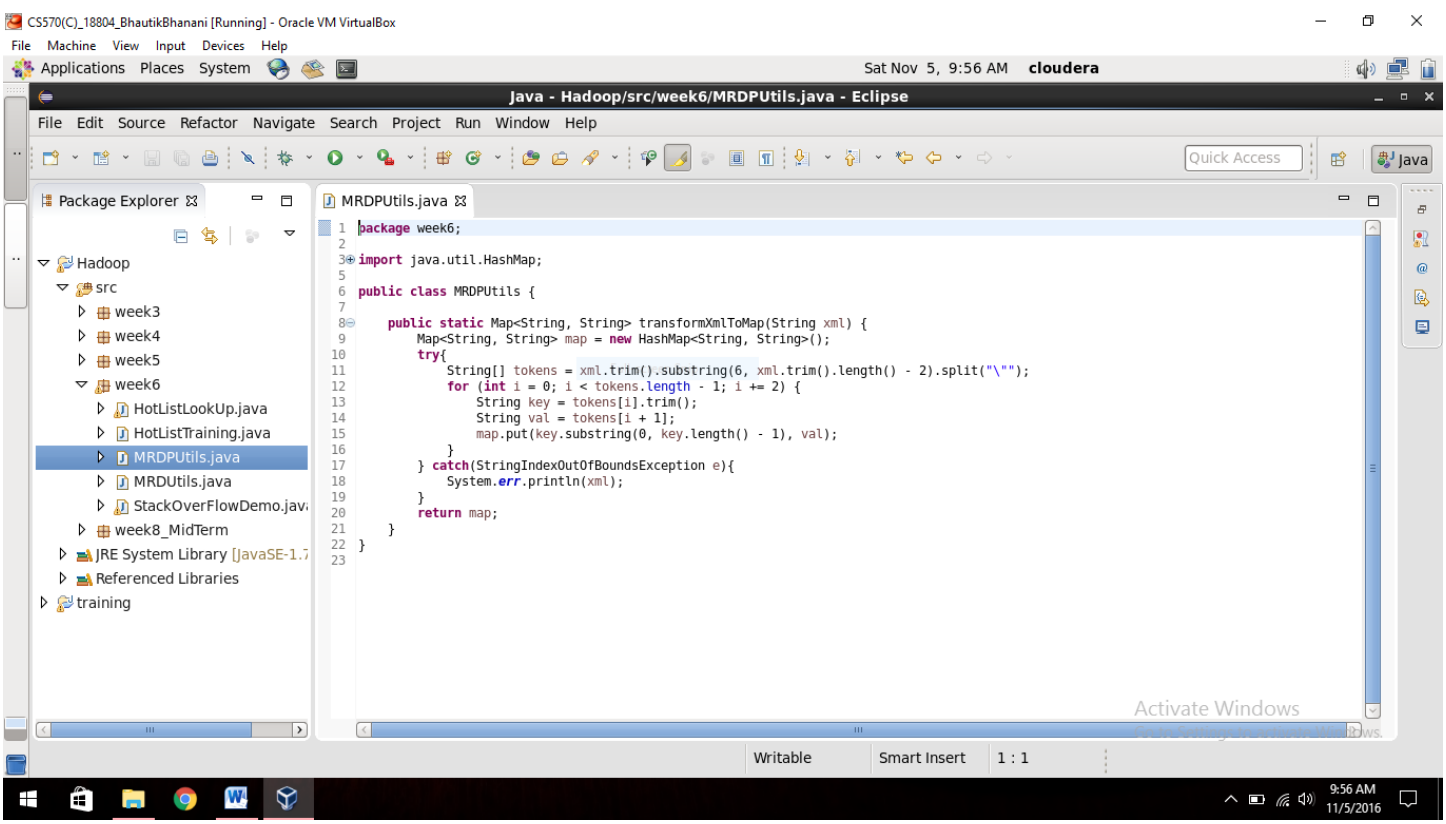
Solution:

Step 1: Open Cloudera and open Eclipse in it. Create one class under Hadoop project and name it ‘HotListTraining’ & ‘HotListLookUp’.





Step 2: Create MRDUtils.java file.



Step 3: These are the codes for two java files.

HotListTraining.java

```
// copy from repository
```

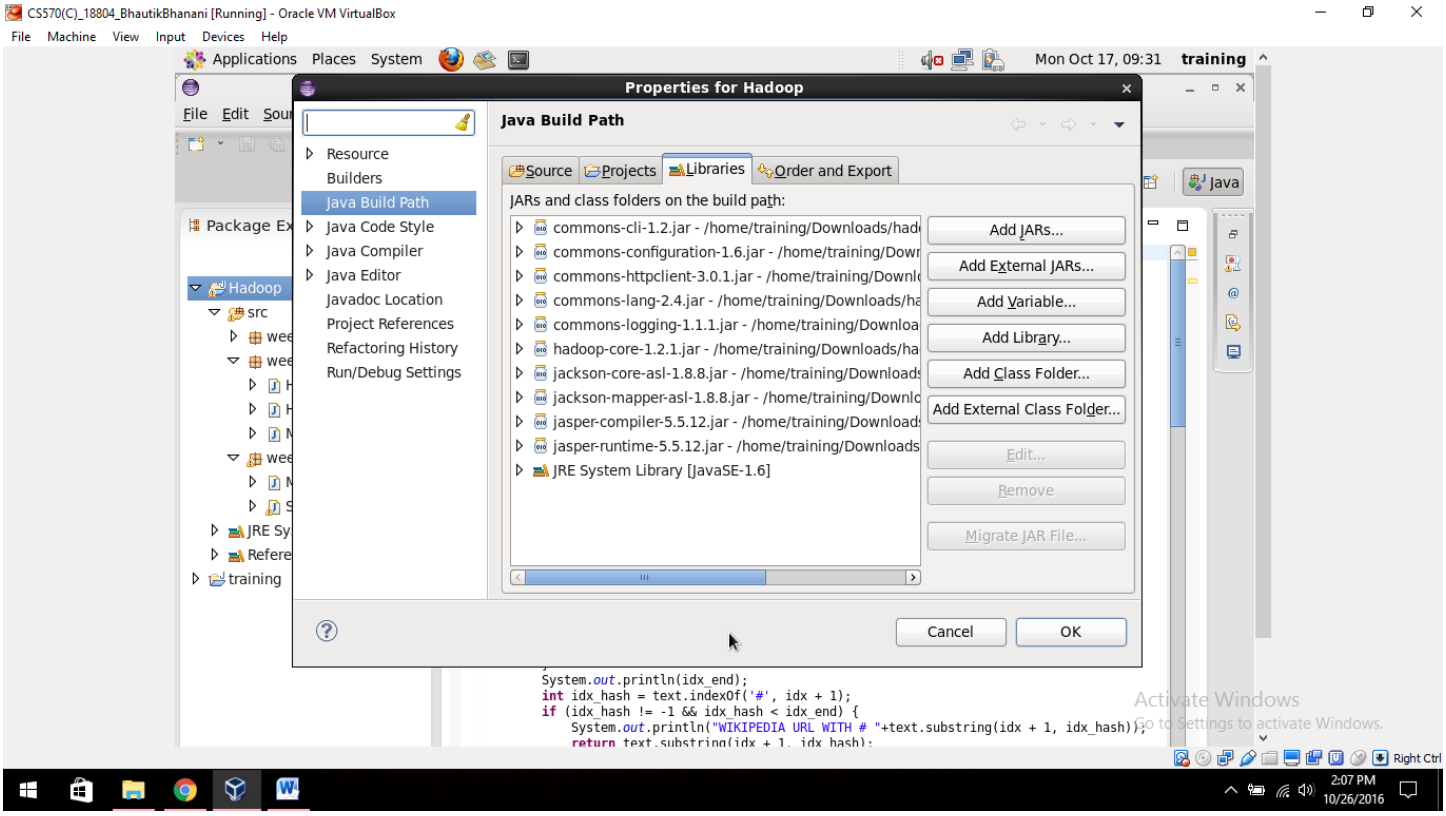
HotListLookUp.java

```
// copy from repository
```

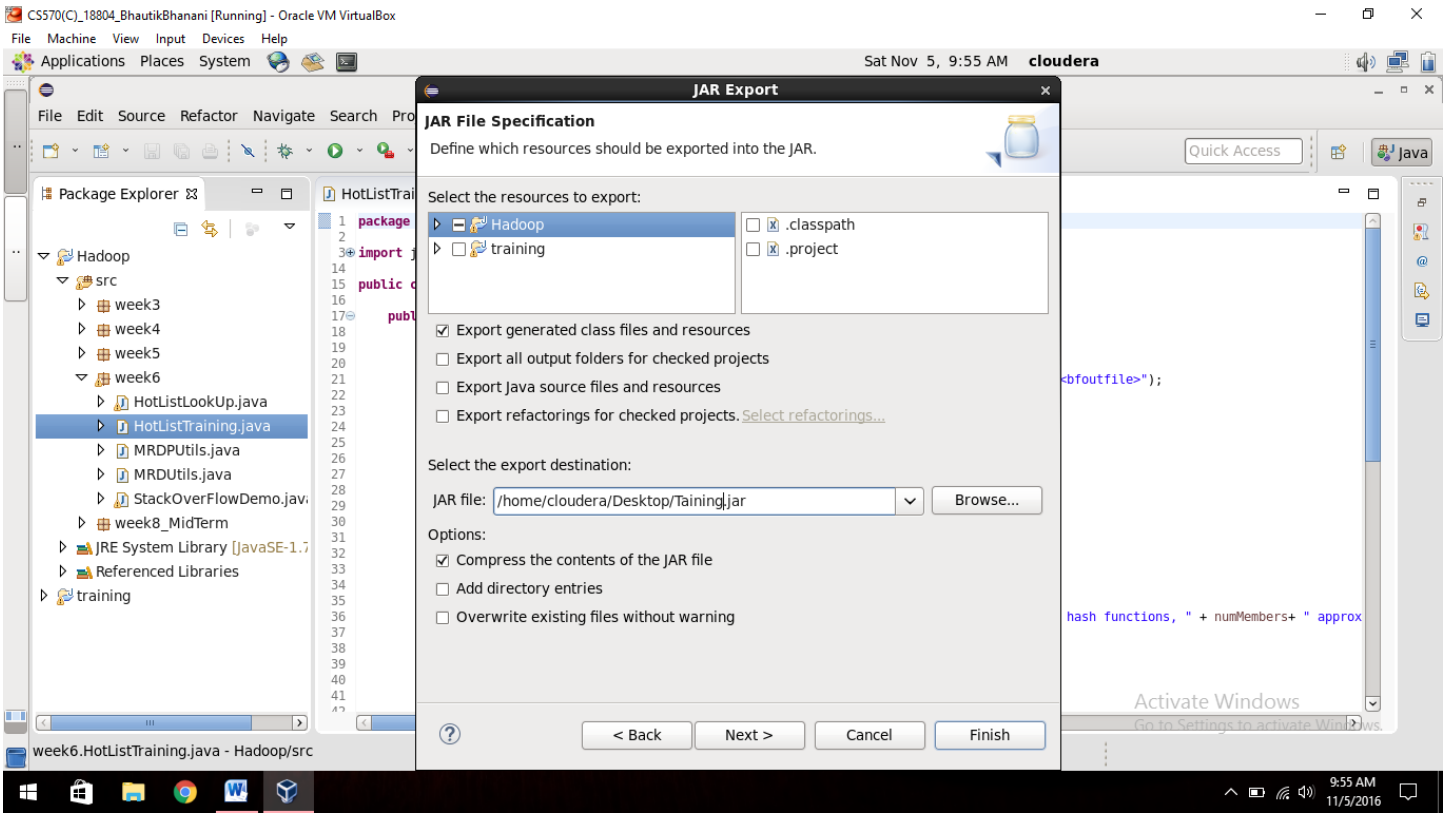
MRDUtils.java

```
// copy from repository
```

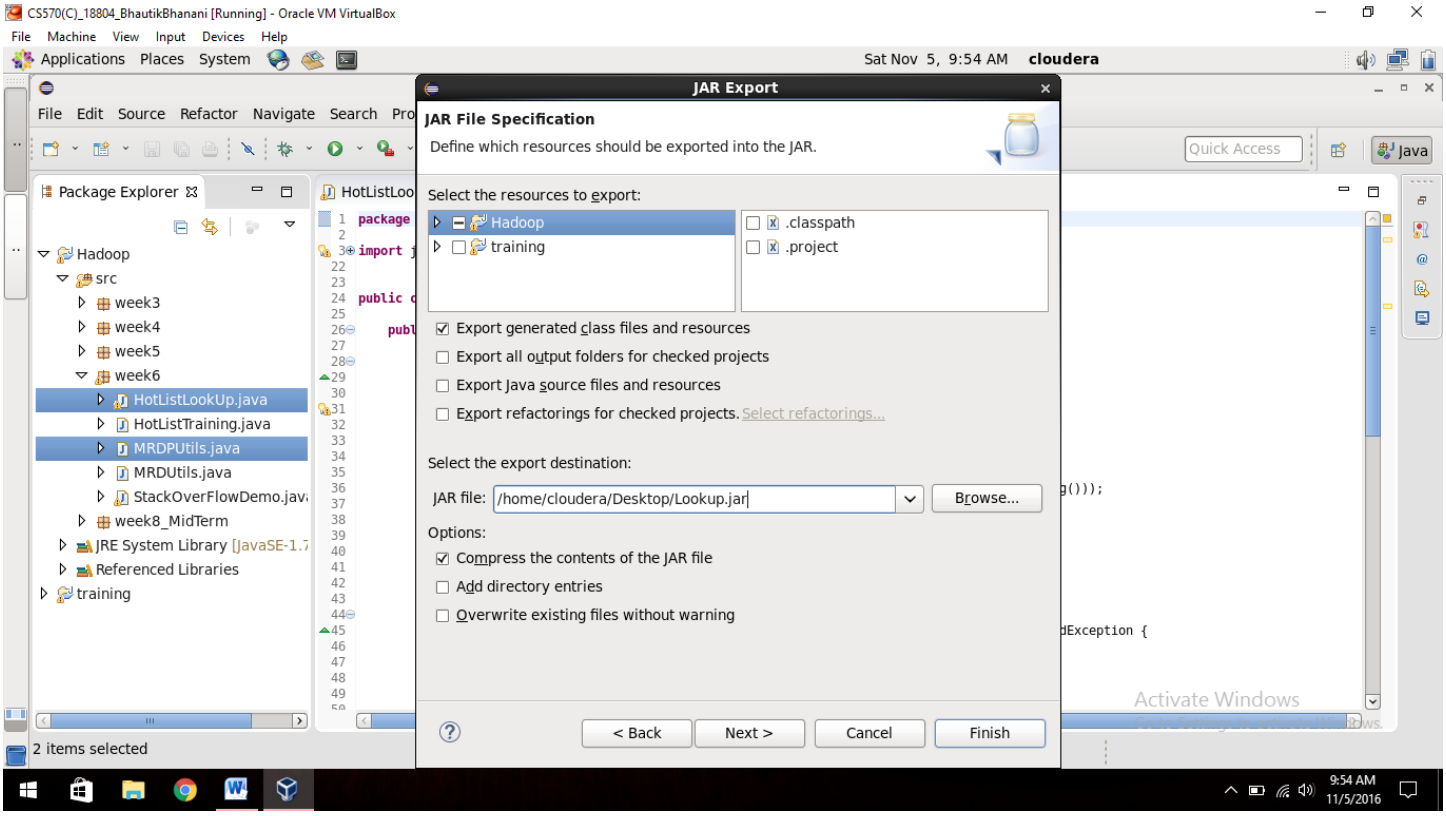
Step 4: Now import necessary hadoop jar files into “Java Build Path” of your project.



Step 5: Now export HotListTraining.class into jar file.

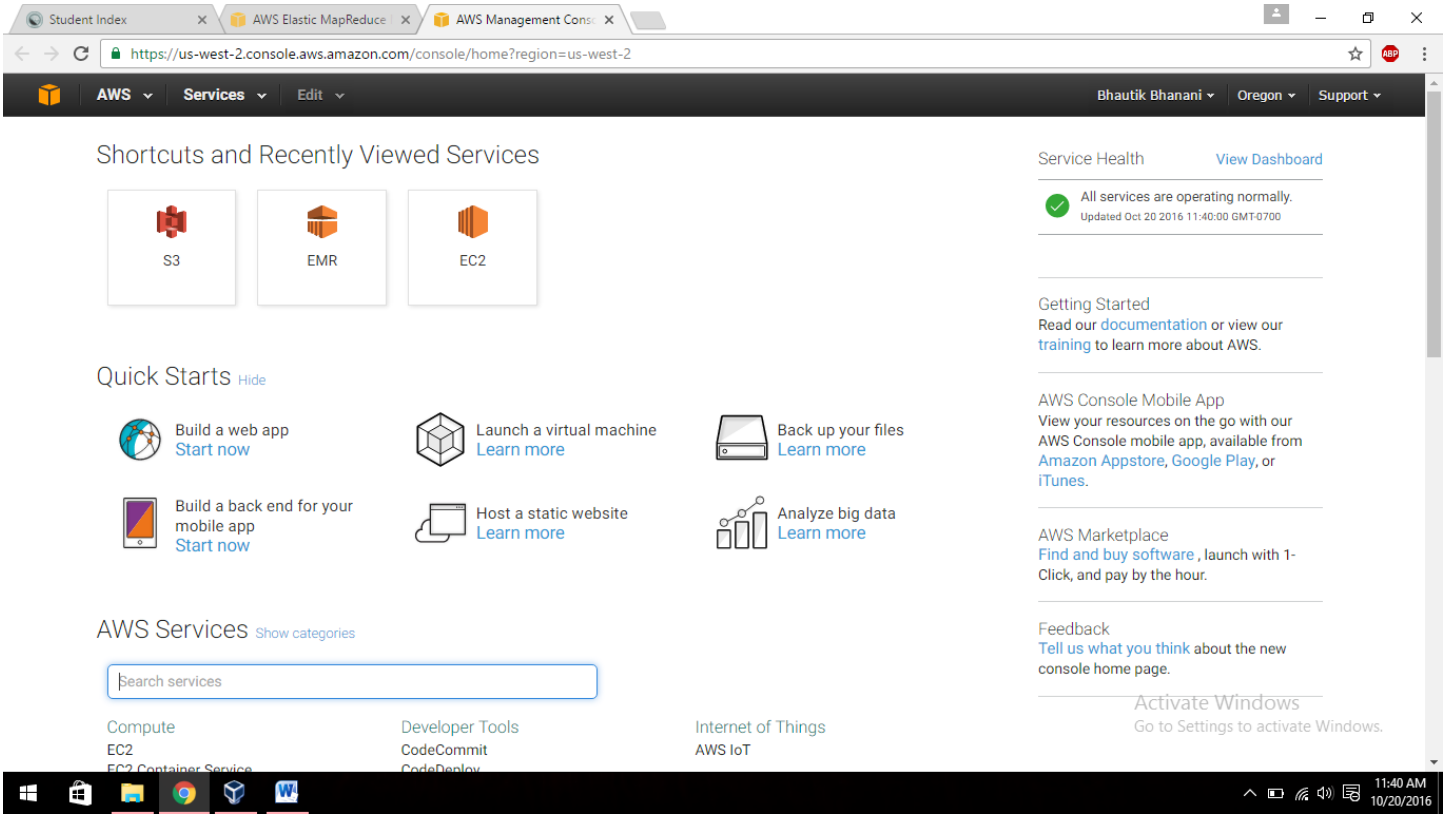


Step 6: Now export HotListLookUp.class and MRDPUtils.class, both into jar file.

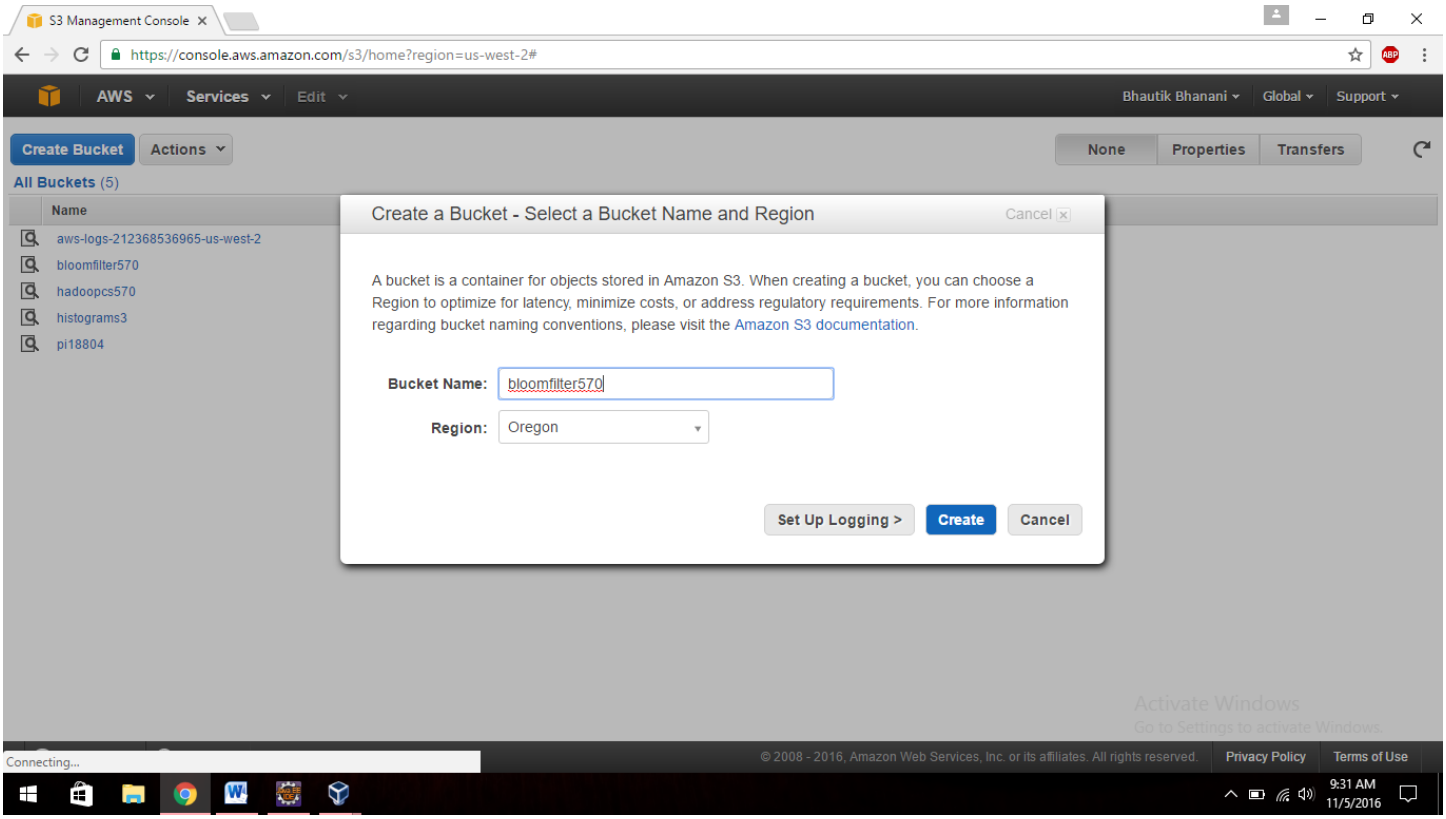


Now run StackOverFlow Indexing project on AWS

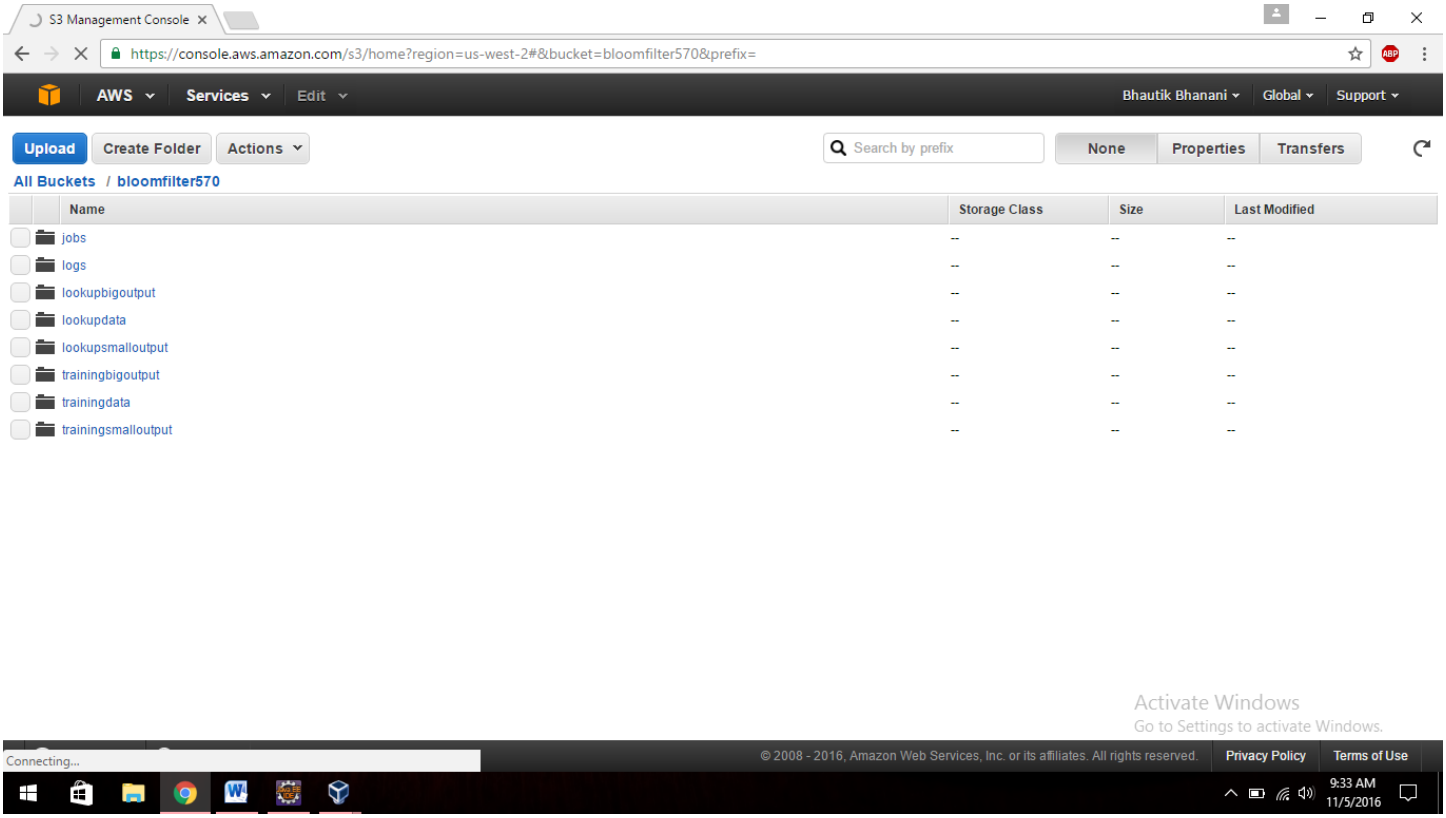
Step 1: Login to your AWS account and AWS dashboard.



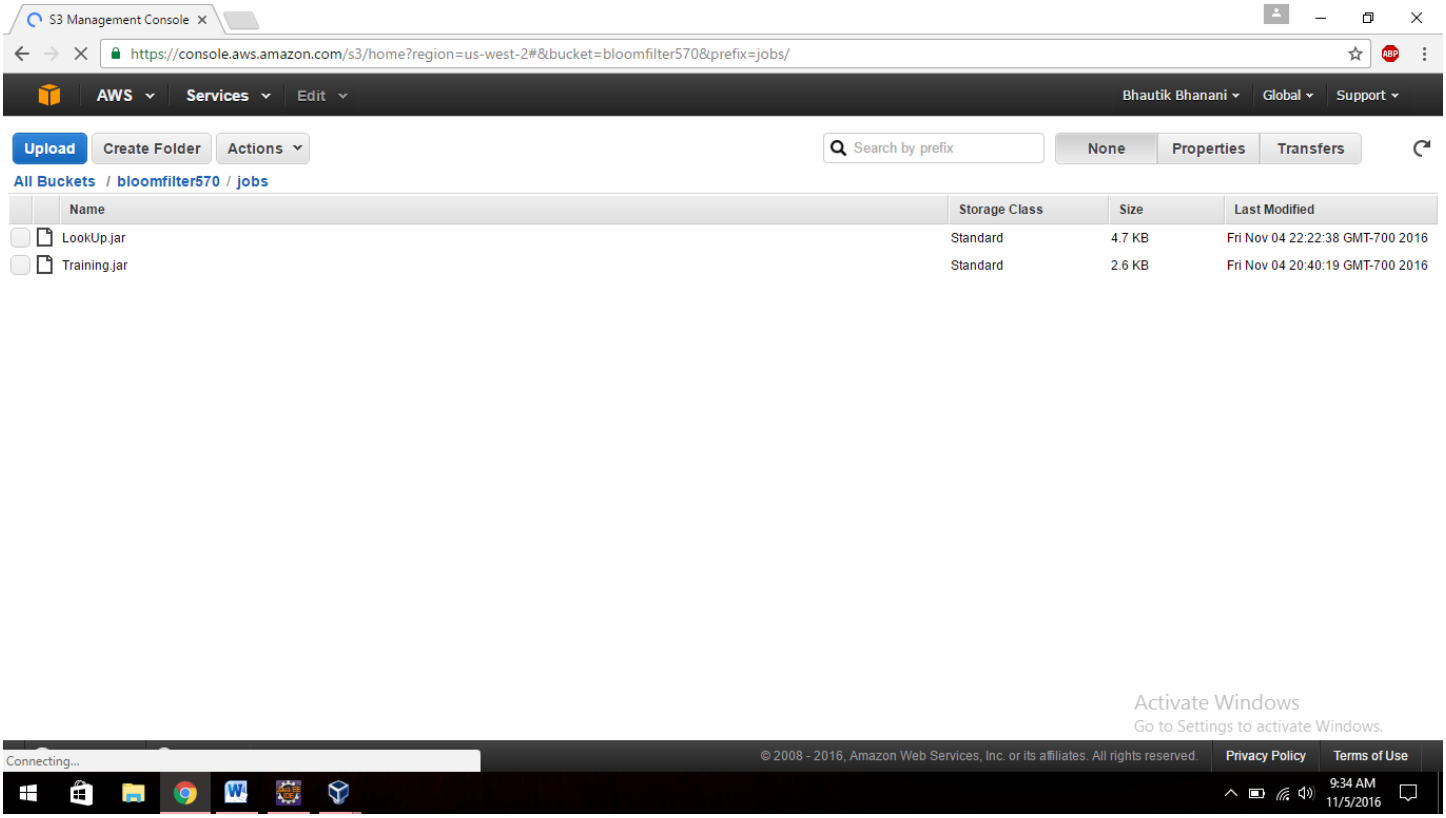
Step 2: Go to **S3**, and create one bucket for Hadoop project.



Step 3: Under your newly created bucket, create following folders: jobs, log, data folder for training and lookup face and output folder for training face. Do not create output folder for lookup face, because we will give output folder name in arguments and AWS will create it.

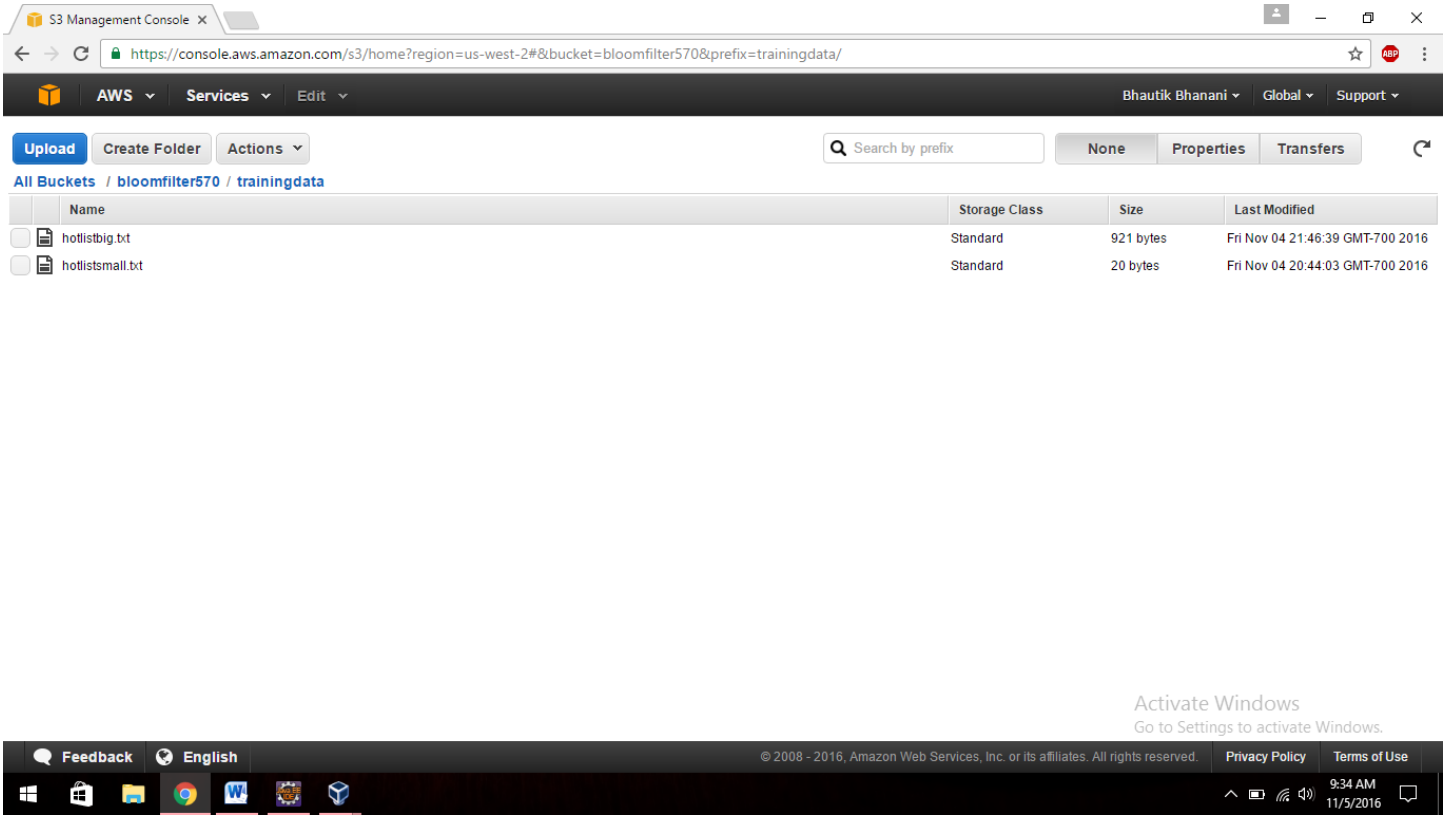


Step 4: Upload your exported jar files under **job** folder.

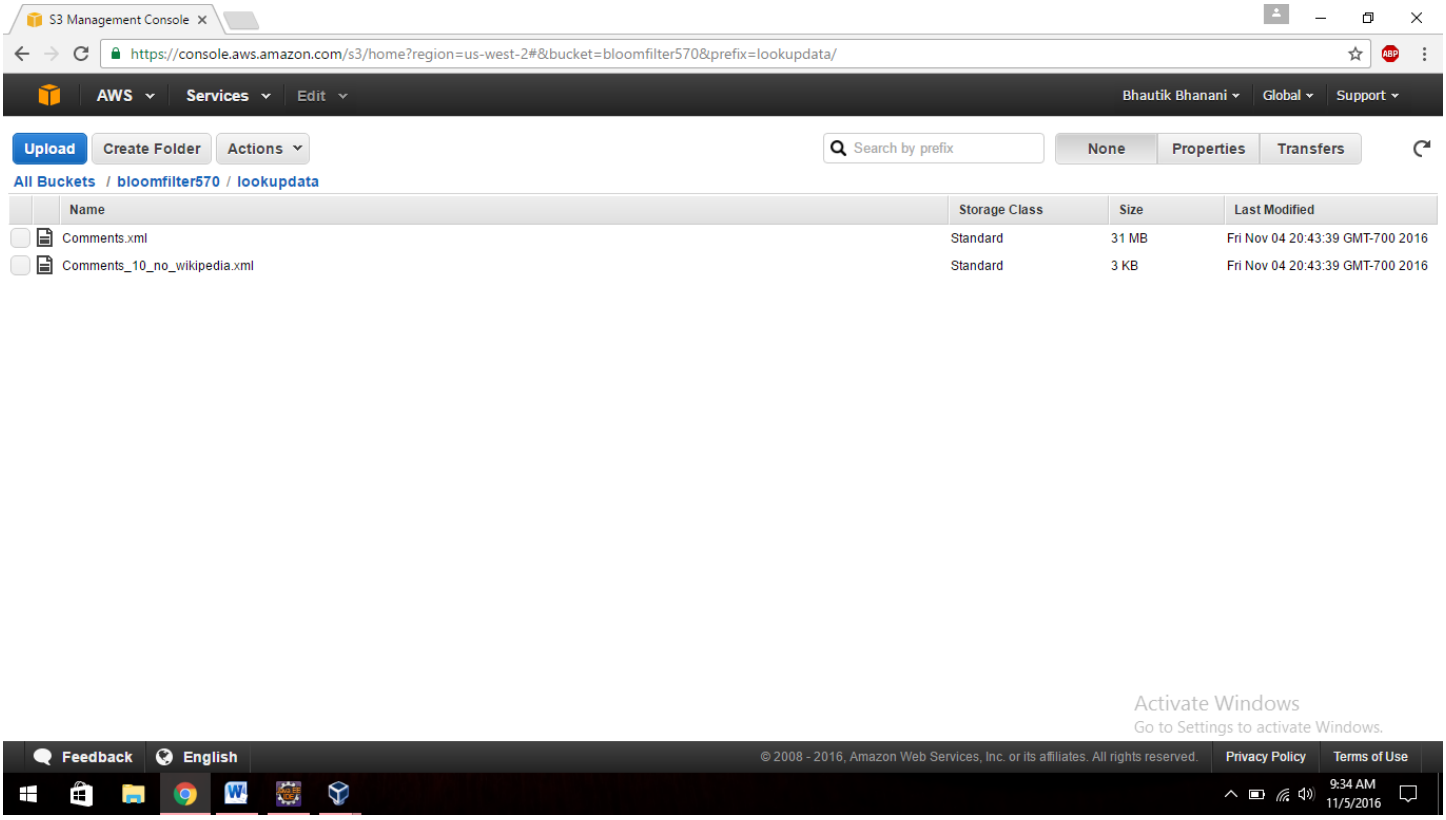


Step 5: Upload input file in input folder.

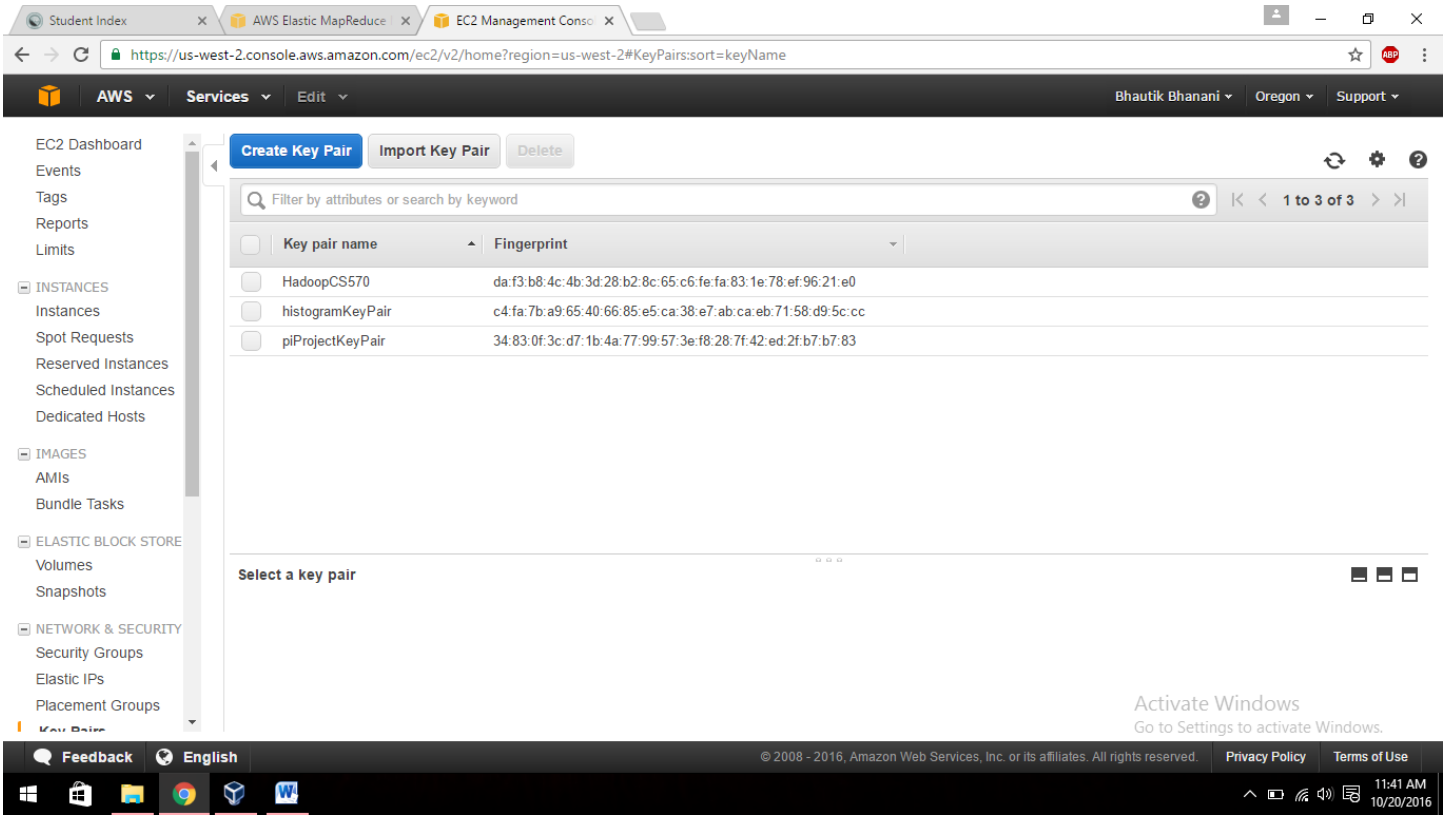
Training Data:



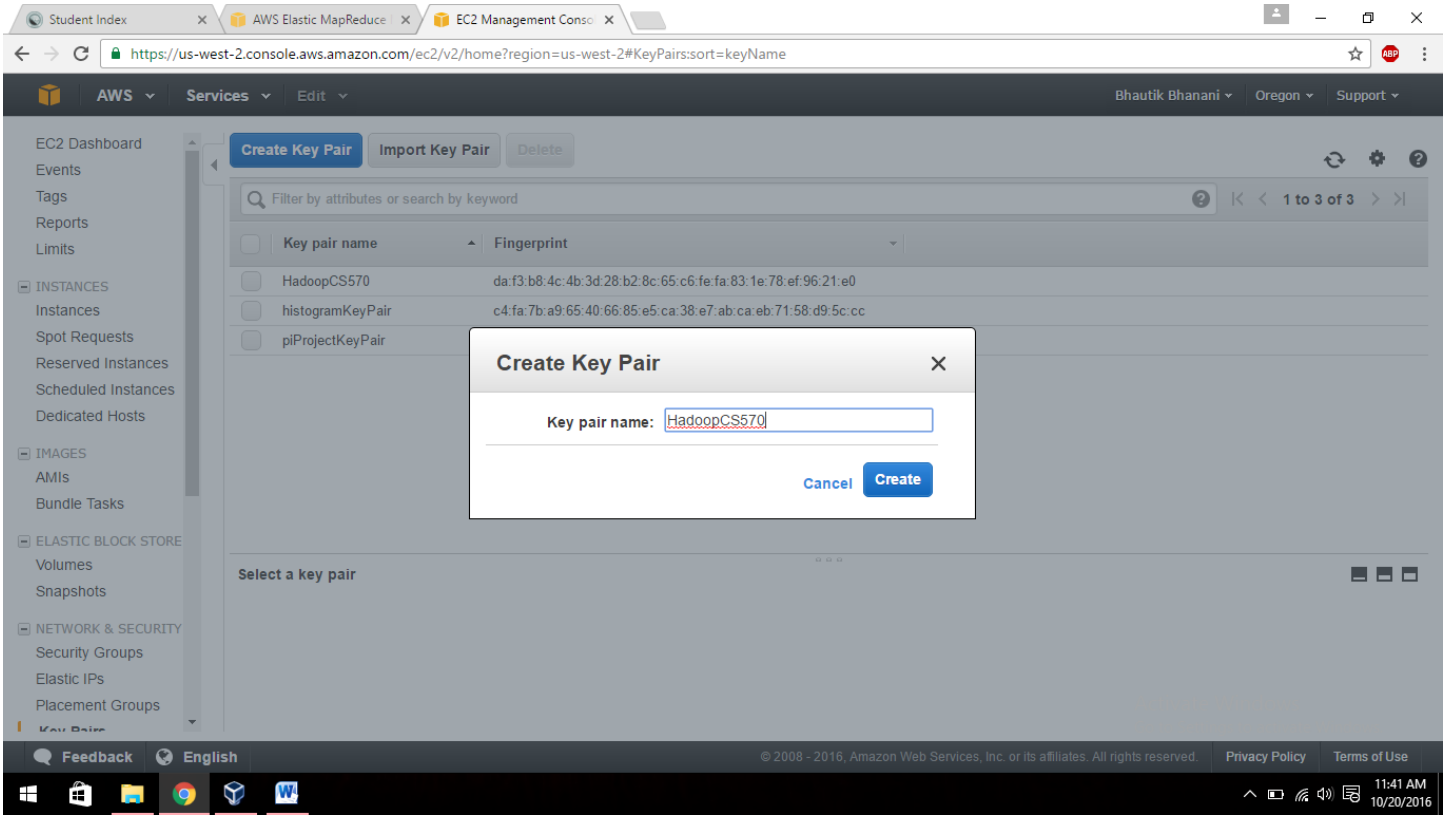
Lookup data:



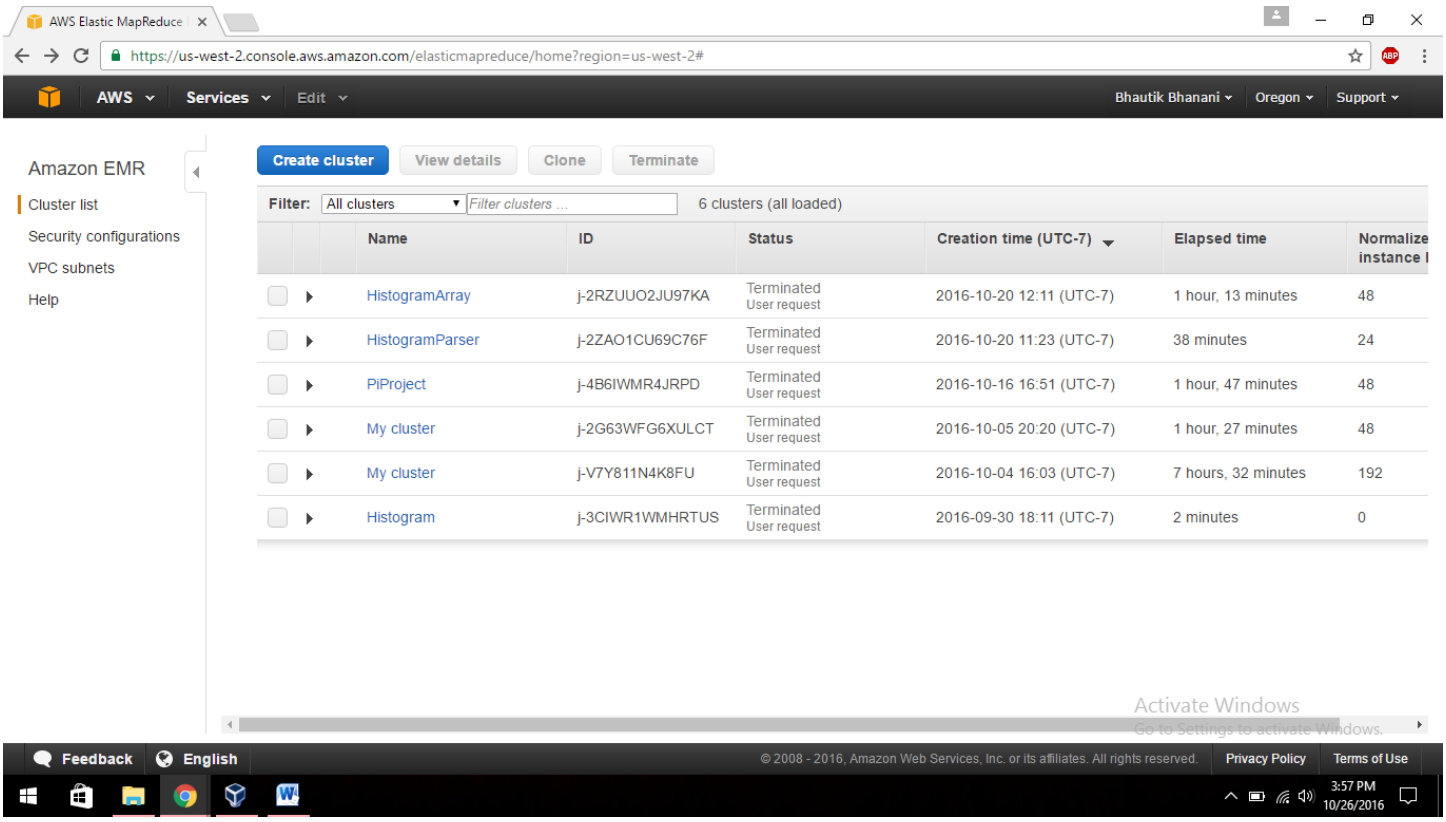
Step 6: Now go to **EC2** from AWS dashboard.



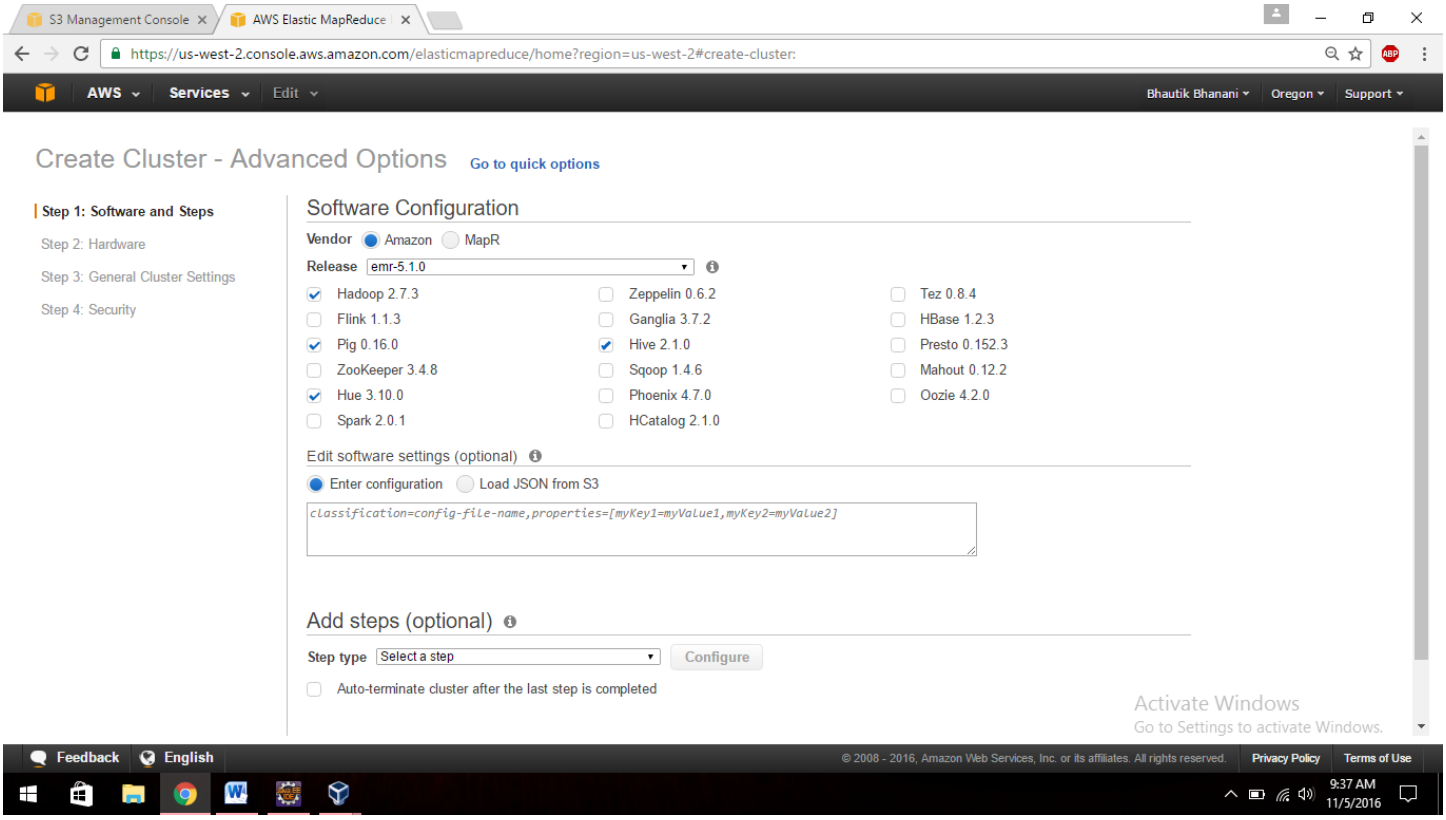
Step 7: Create key pair for your project.



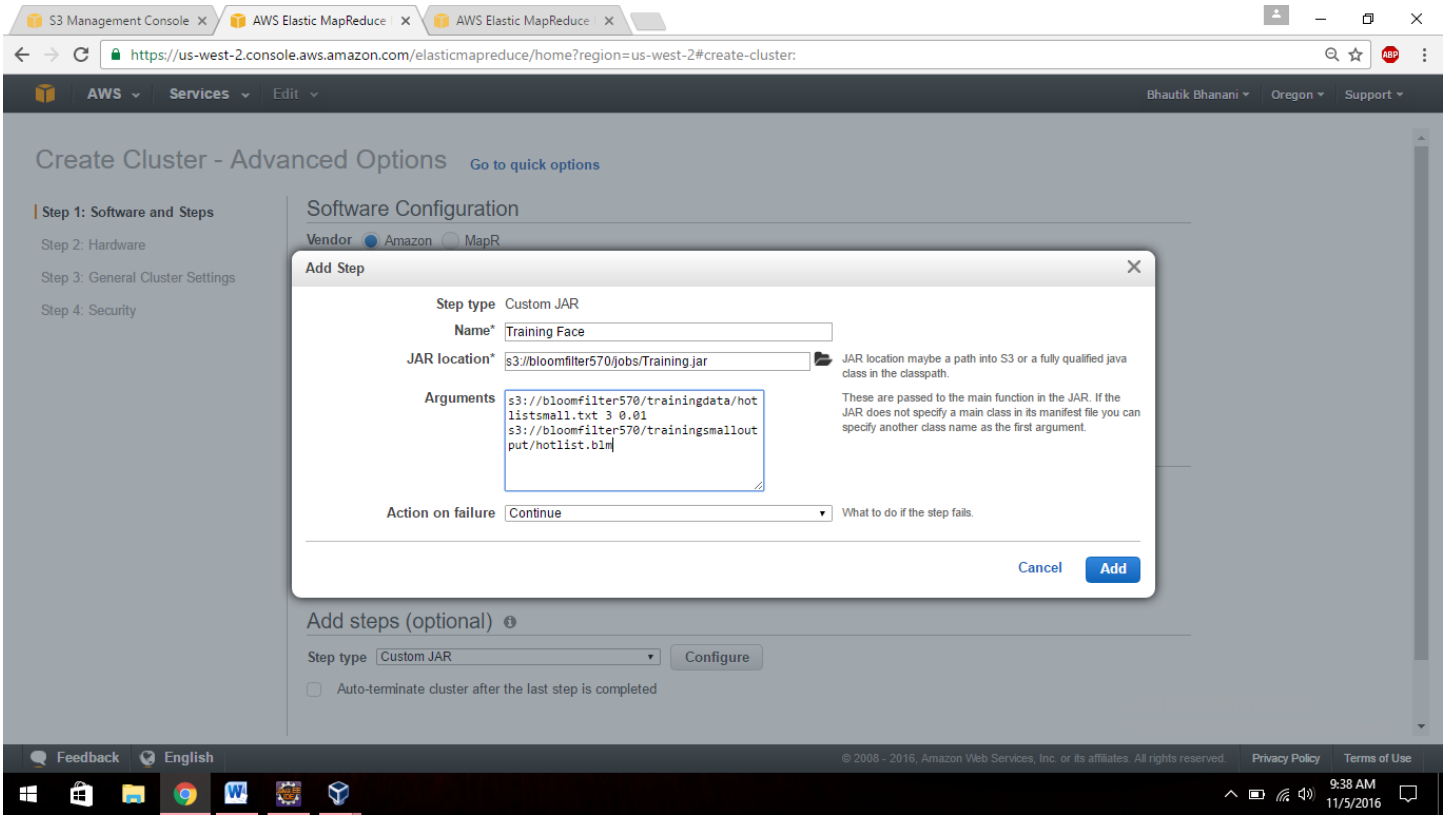
Step 8: Go to **EMR** from AWS dashboard.

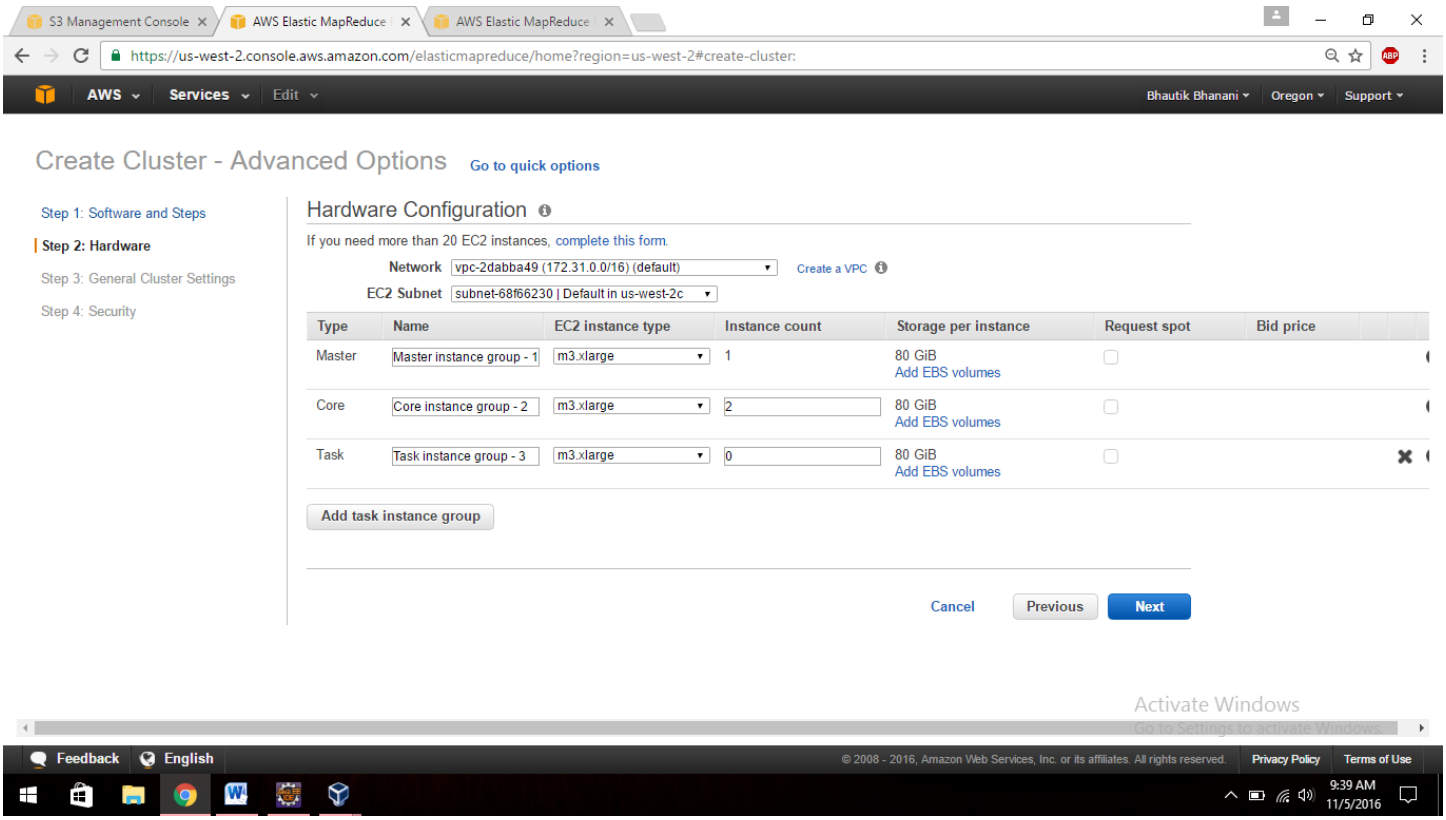


Step 9: Create new cluster of Bloom Filter project.



Add one training small data face as custom jar.





S3 Management Console | AWS Elastic MapReduce | AWS Elastic MapReduce

https://us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#create-cluster:

AWS | Services | Edit | Bhautik Bhanani | Oregon | Support

Create Cluster - Advanced Options

Go to quick options

- Step 1: Software and Steps
- Step 2: Hardware
- **Step 3: General Cluster Settings**
- Step 4: Security

General Options

Cluster name: BloomFilter

[x] Logging

S3 folder: s3://bloomfilter570/logs/

[x] Debugging

[x] Termination protection

Tags

Key	Value (optional)
Add a key to create a tag	

Additional Options

[ ] EMRFS consistent view

▸ Bootstrap Actions

Cancel | Previous | Next

Feedback | English

Privacy Policy | Terms of Use

9:39 AM 11/5/2016

S3 Management Console | AWS Elastic MapReduce | AWS Elastic MapReduce

https://us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#create-cluster:

AWS | Services | Edit | Bhautik Bhanani | Oregon | Support

Create Cluster - Advanced Options

Go to quick options

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- **Step 4: Security**

Security Options

EC2 key pair: HadoopCS570

[x] Cluster visible to all IAM users in account

Permissions

(●) Default ( ) Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole

EC2 instance profile: EMR_EC2_DefaultRole

▸ Encryption Options

▸ EC2 Security Groups

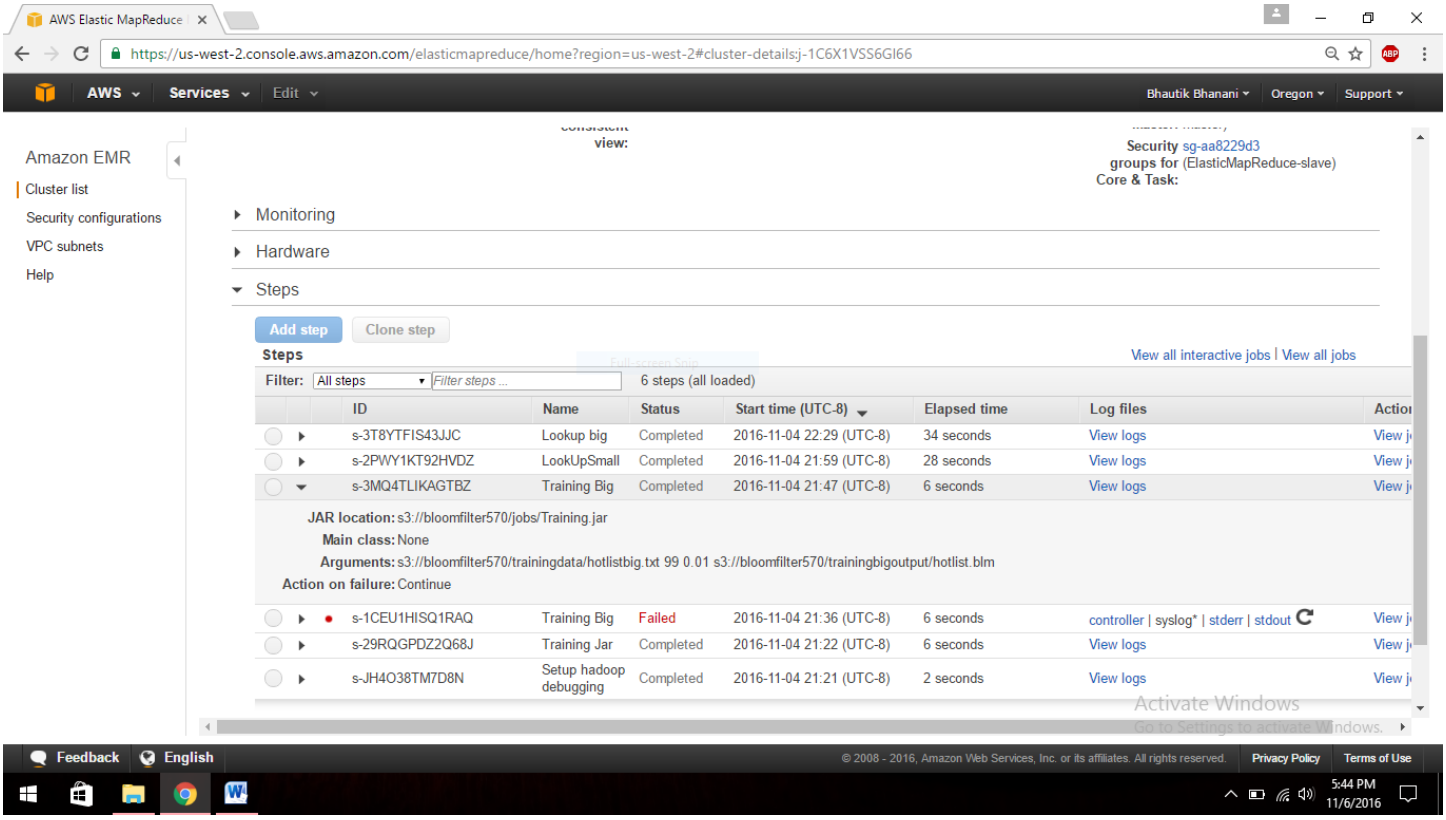
Cancel | Previous | Create cluster

Feedback | English

Privacy Policy | Terms of Use

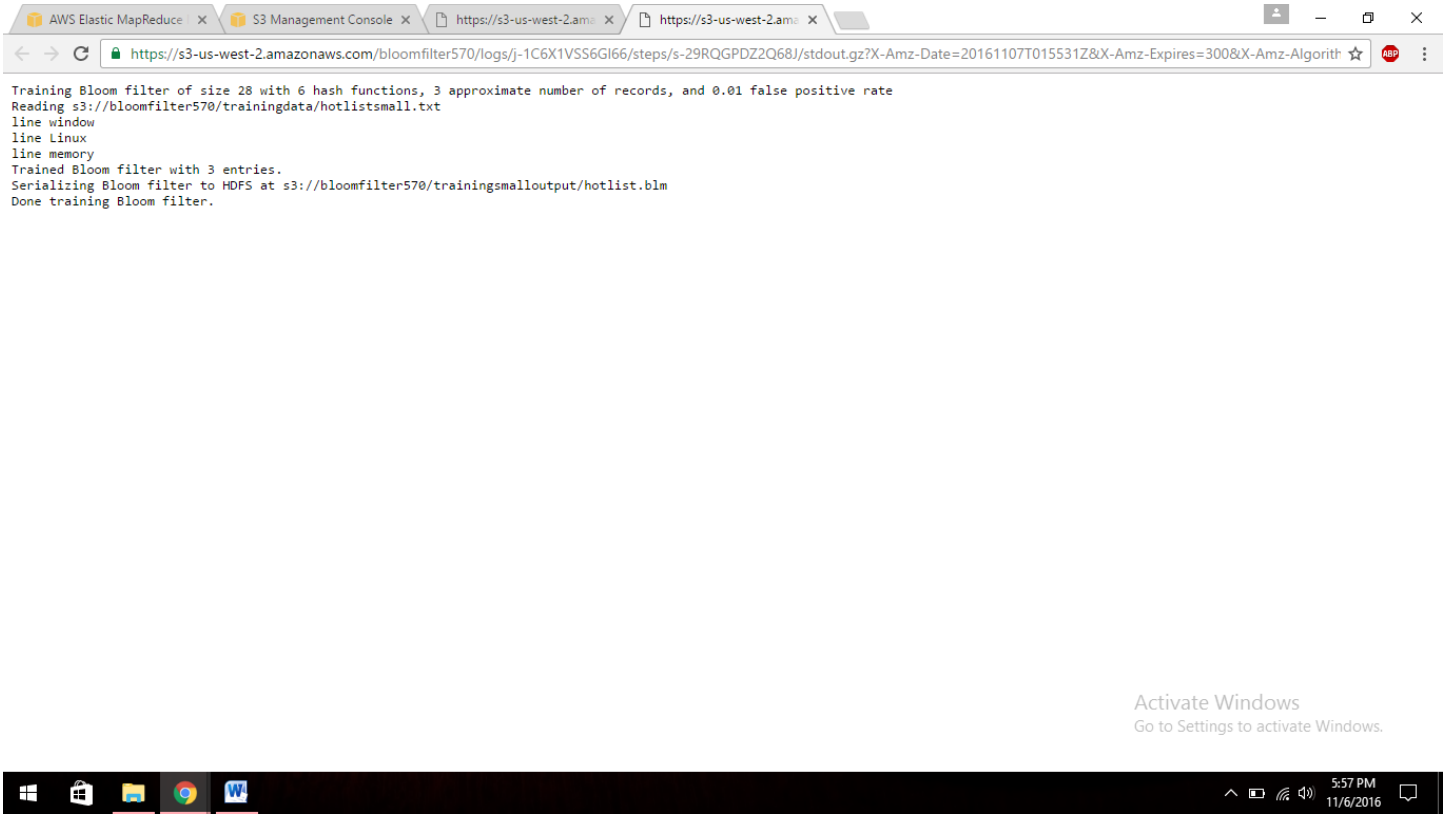
9:39 AM 11/5/2016

Step 10: When first training face completes, create second step for training with big data. Change argument as per big data.

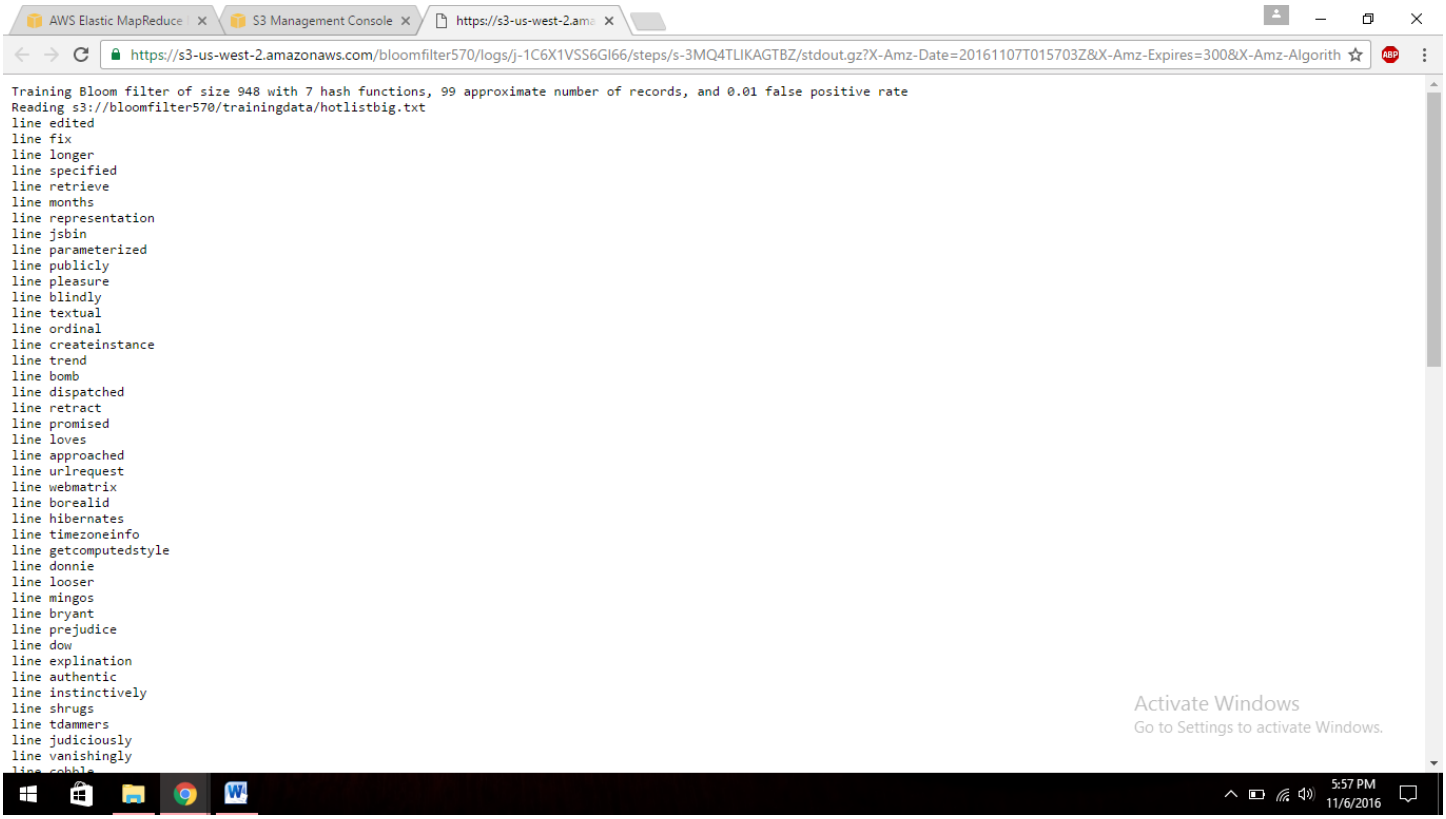


Step 11: When both training face completes, check outputs.

Small data:

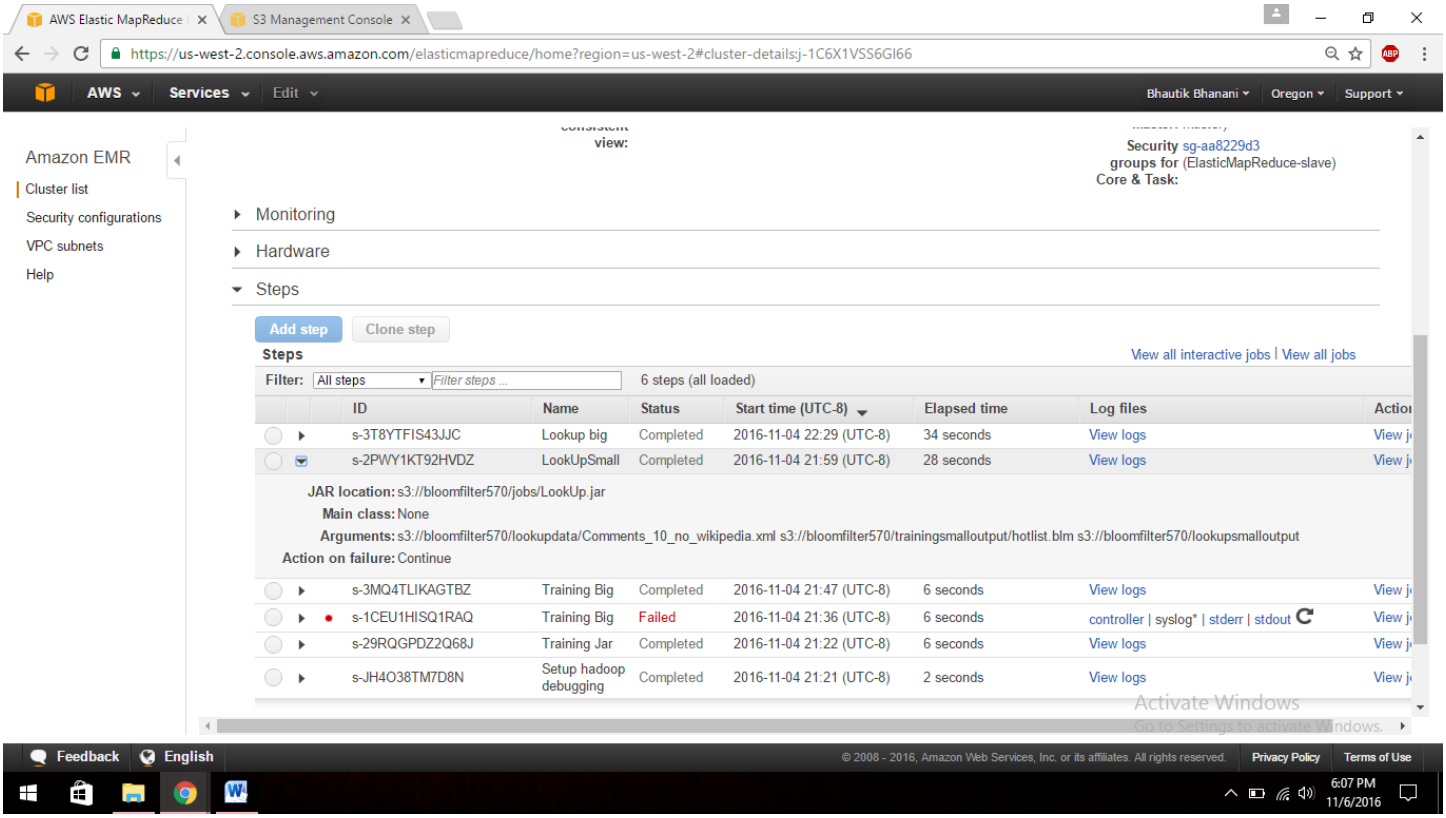


Big data:

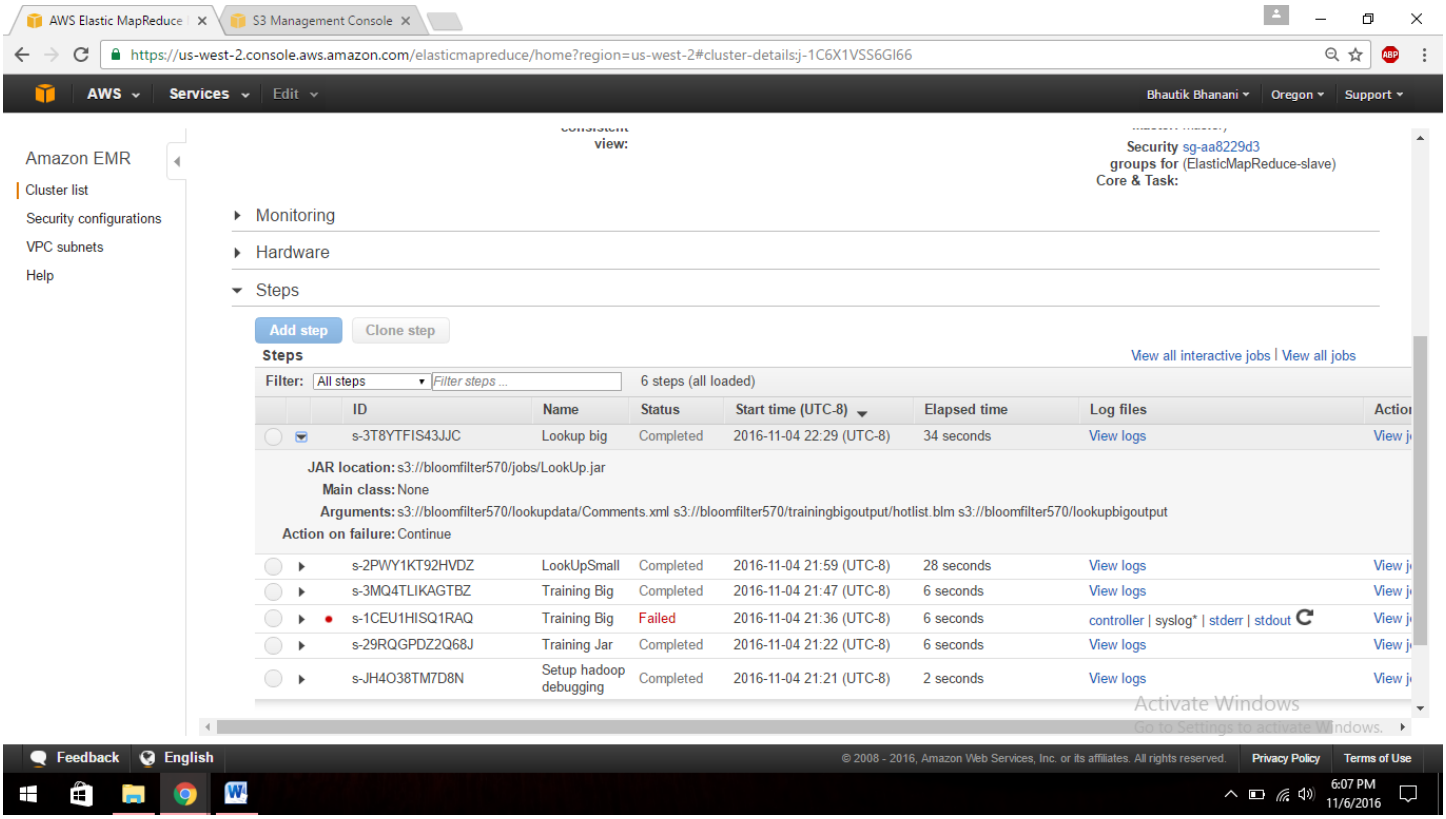


Step 12: After this, now create steps for lookup face for small and big data.

Argument: “/%PATH%/comment_10_no_wikipedia.xml /%PATH%/hotlist.blm /%OUTPUT_PATH%/”

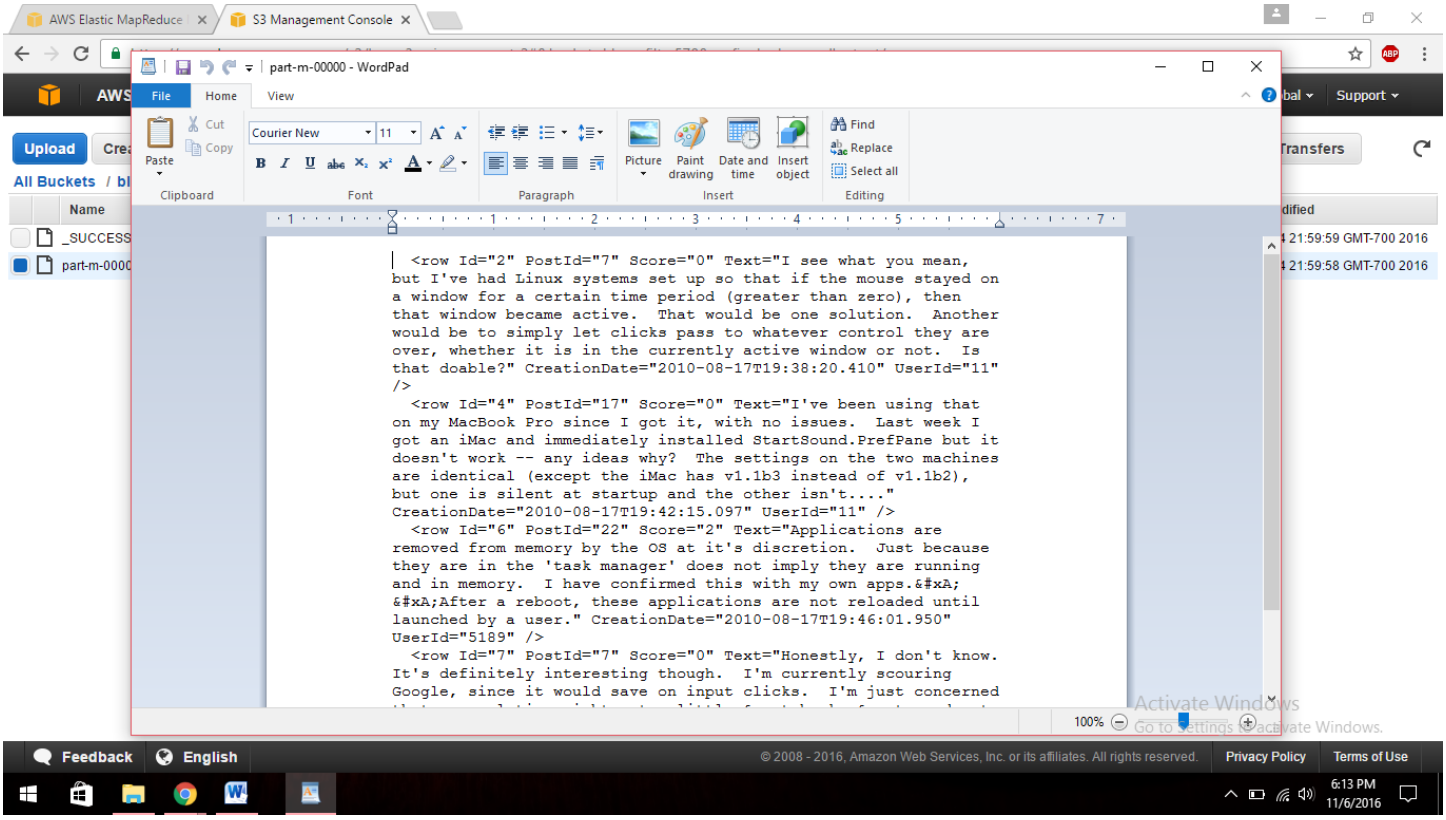


Argument: “/%PATH%/comments.xml /%PATH%/hotlist.blm /%OUTPUT_PATH%/”

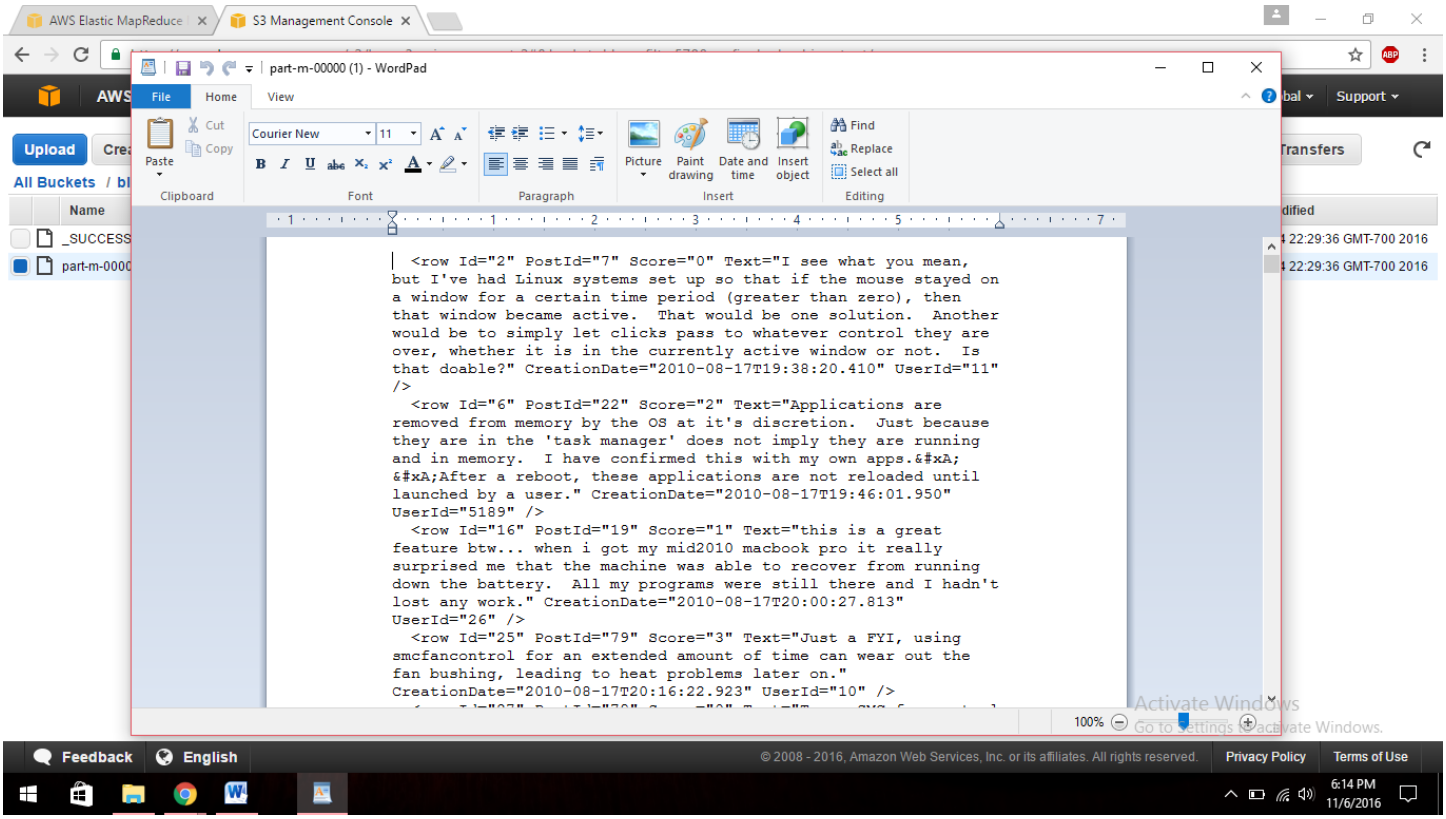


Step 13: Once both step complete, check output folder and open part-r-00000 file.

Small data output:



Big data output:



Step 14: After this, terminate cluster.