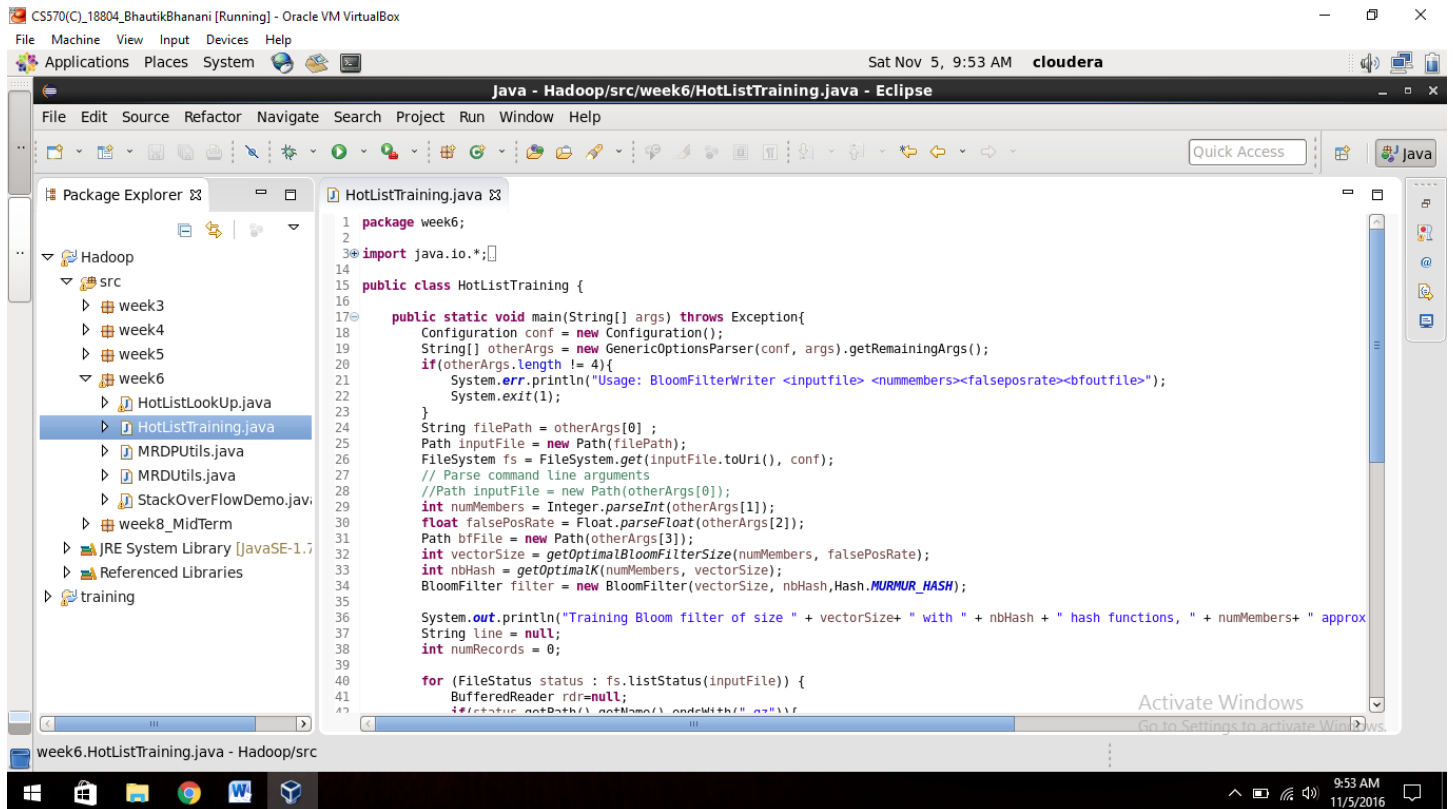
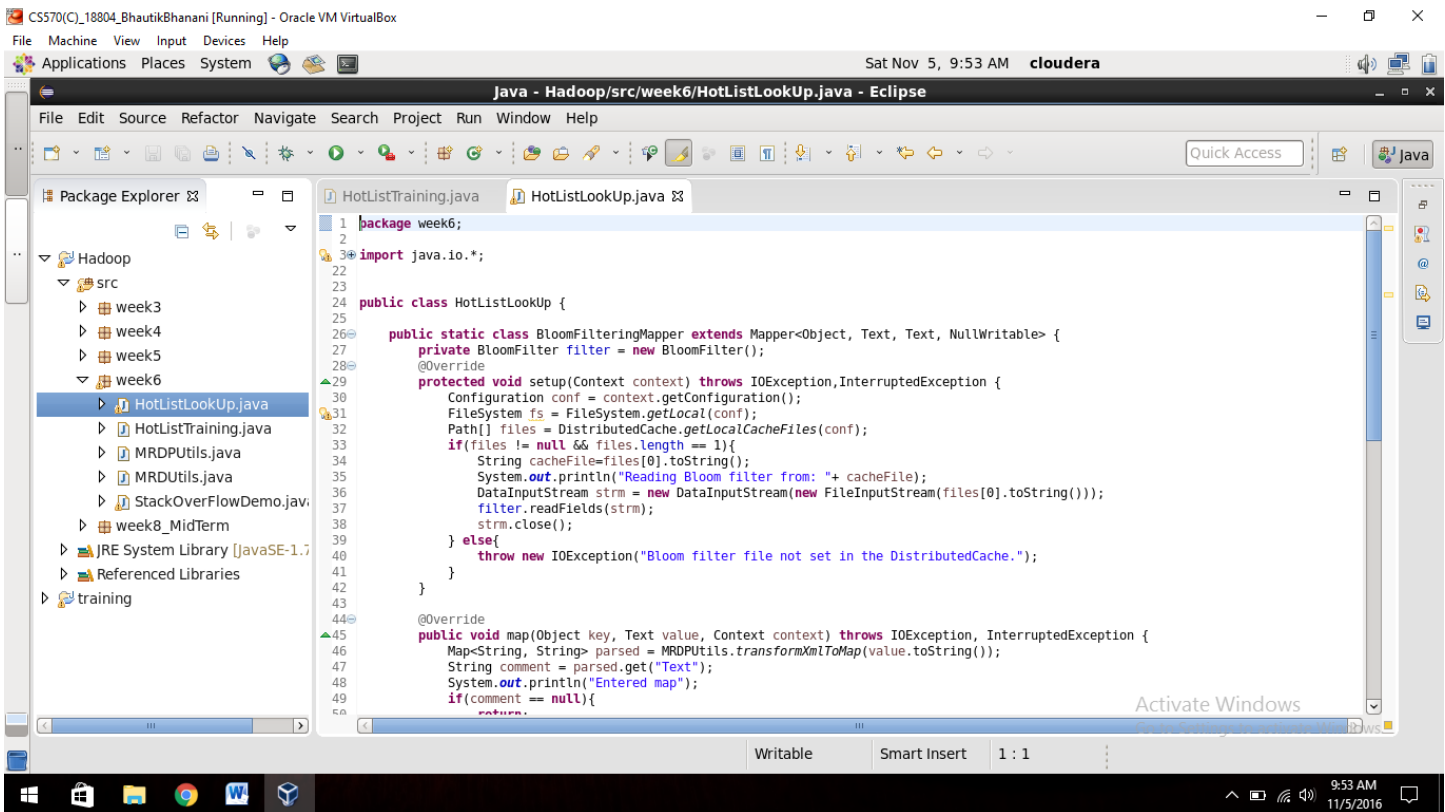


MapReduce to implement Bloom Filter for Hot List

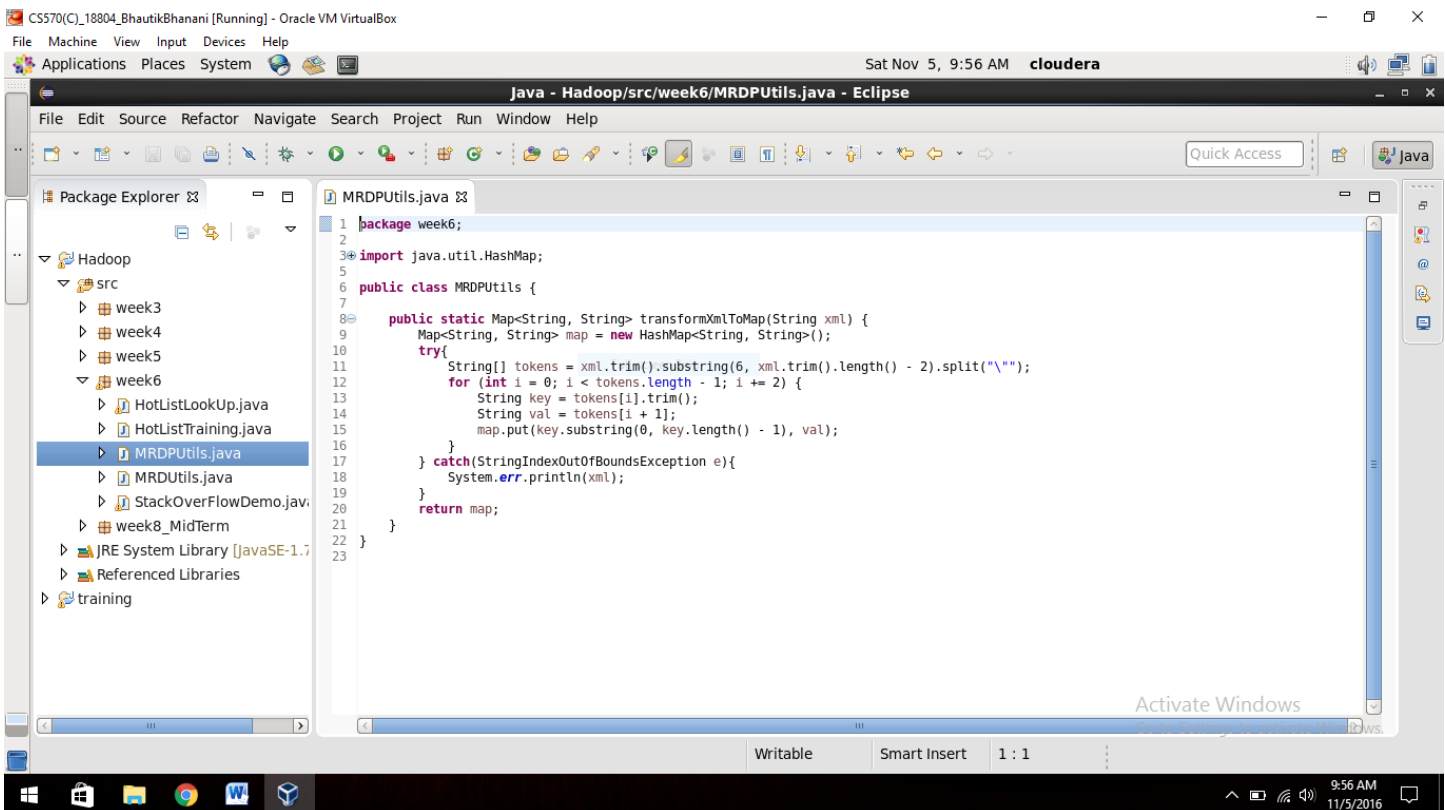
Solution:

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
```

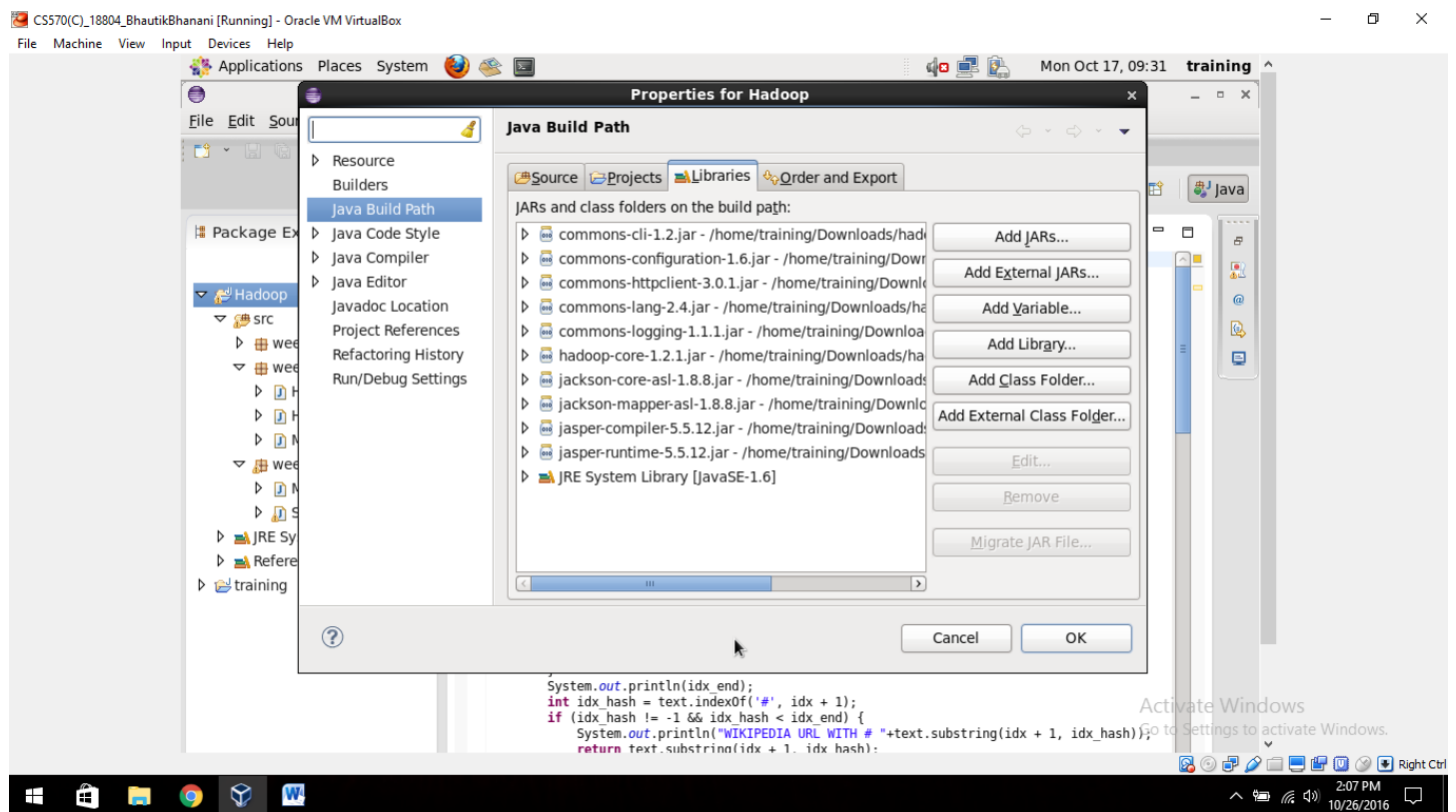
HotListLookUp.java

```
// copy from repository
```

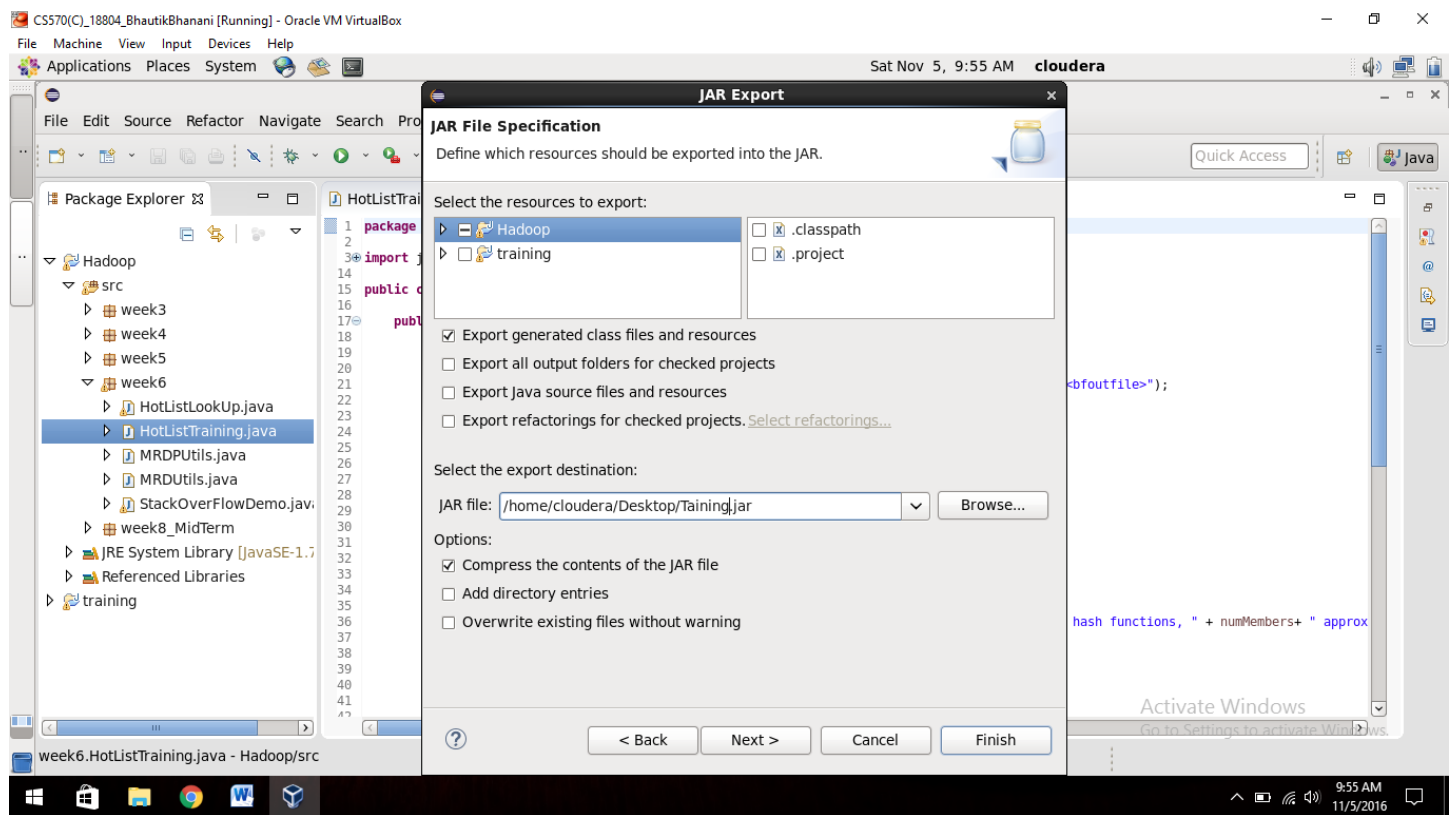
MRDUtils.java

```
// copy from repository
```

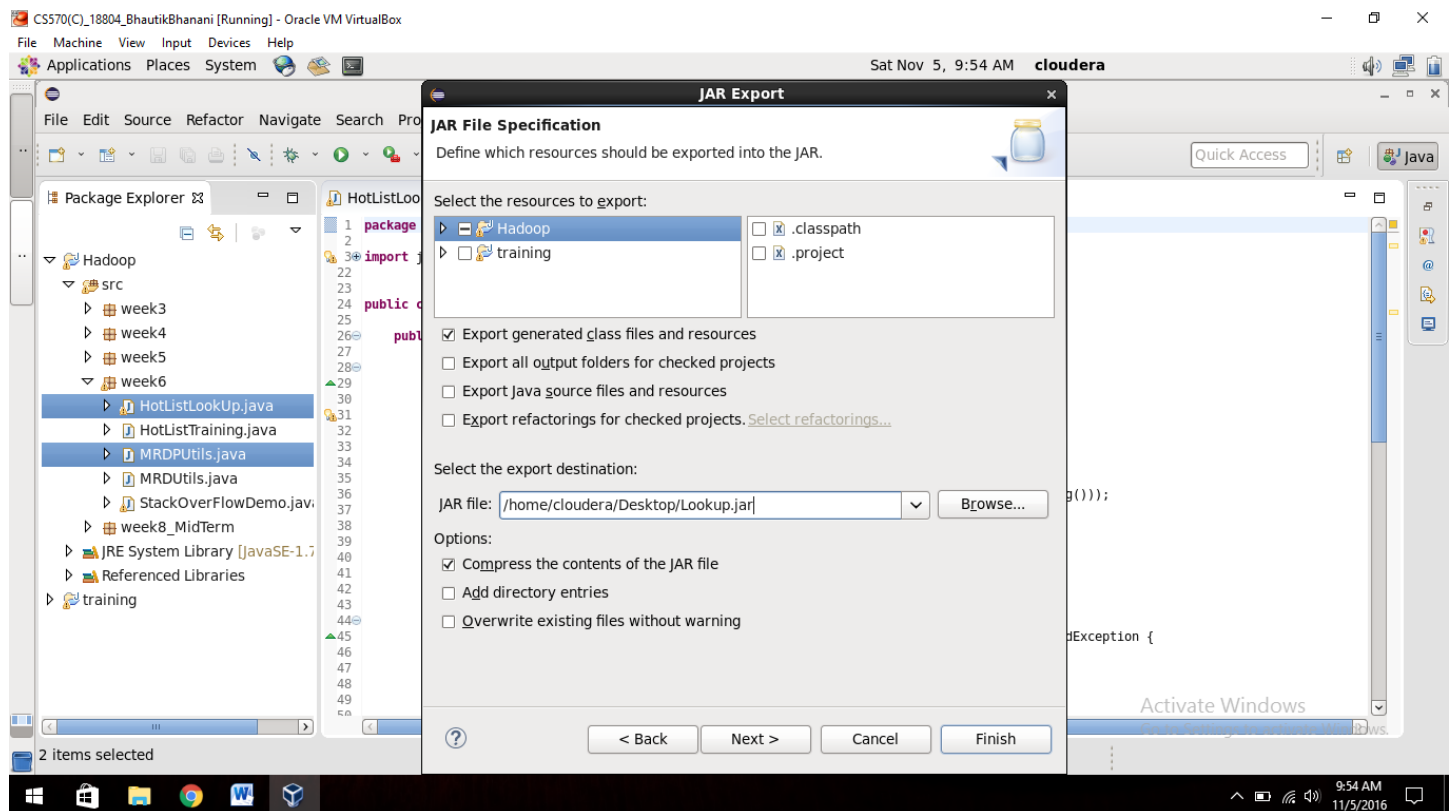
Step 4: Now import necessary hadoop jar files into “Java Build Path” of your project.



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

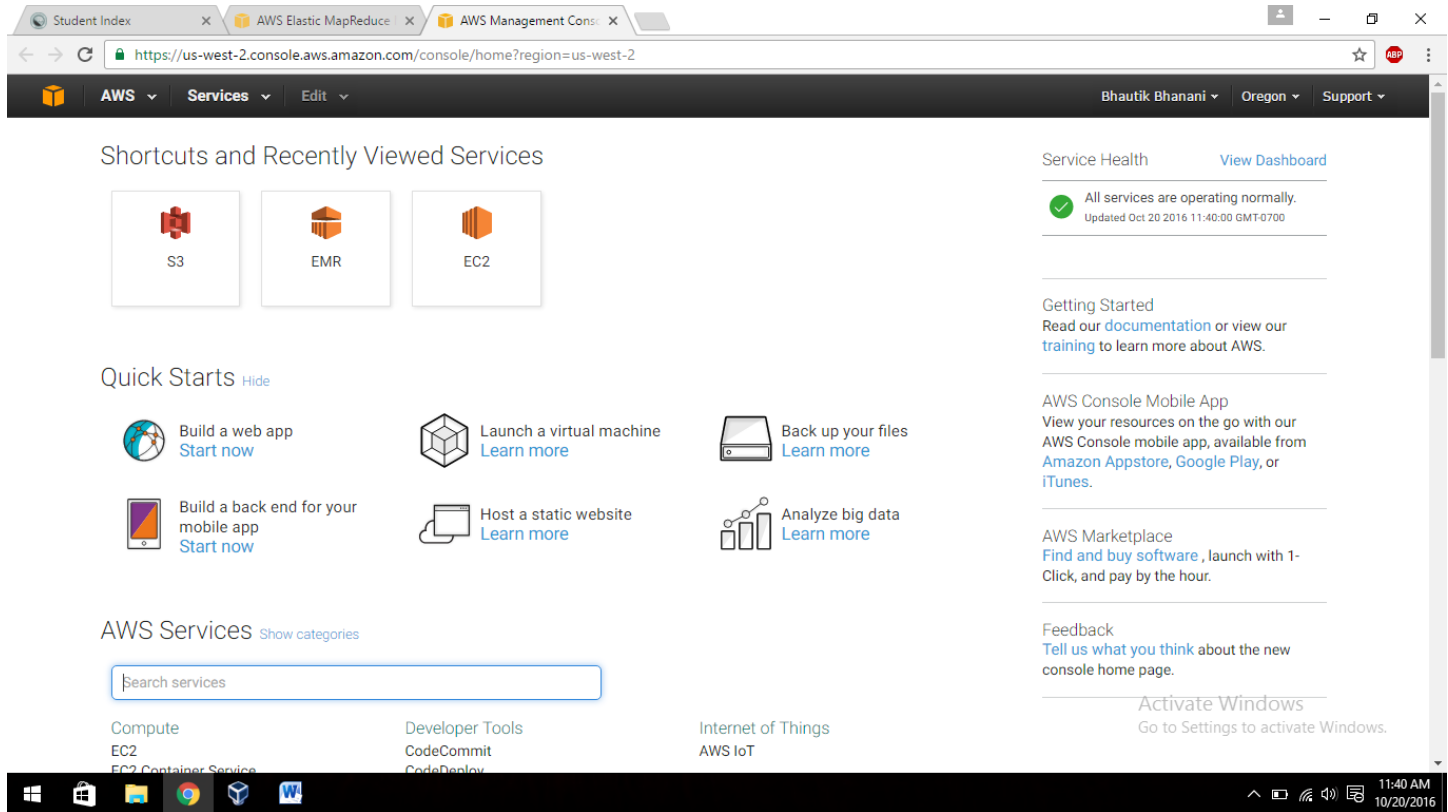


Step 6: Now export HotListLookUp.class and MRDPUtills.class, both into jar file.

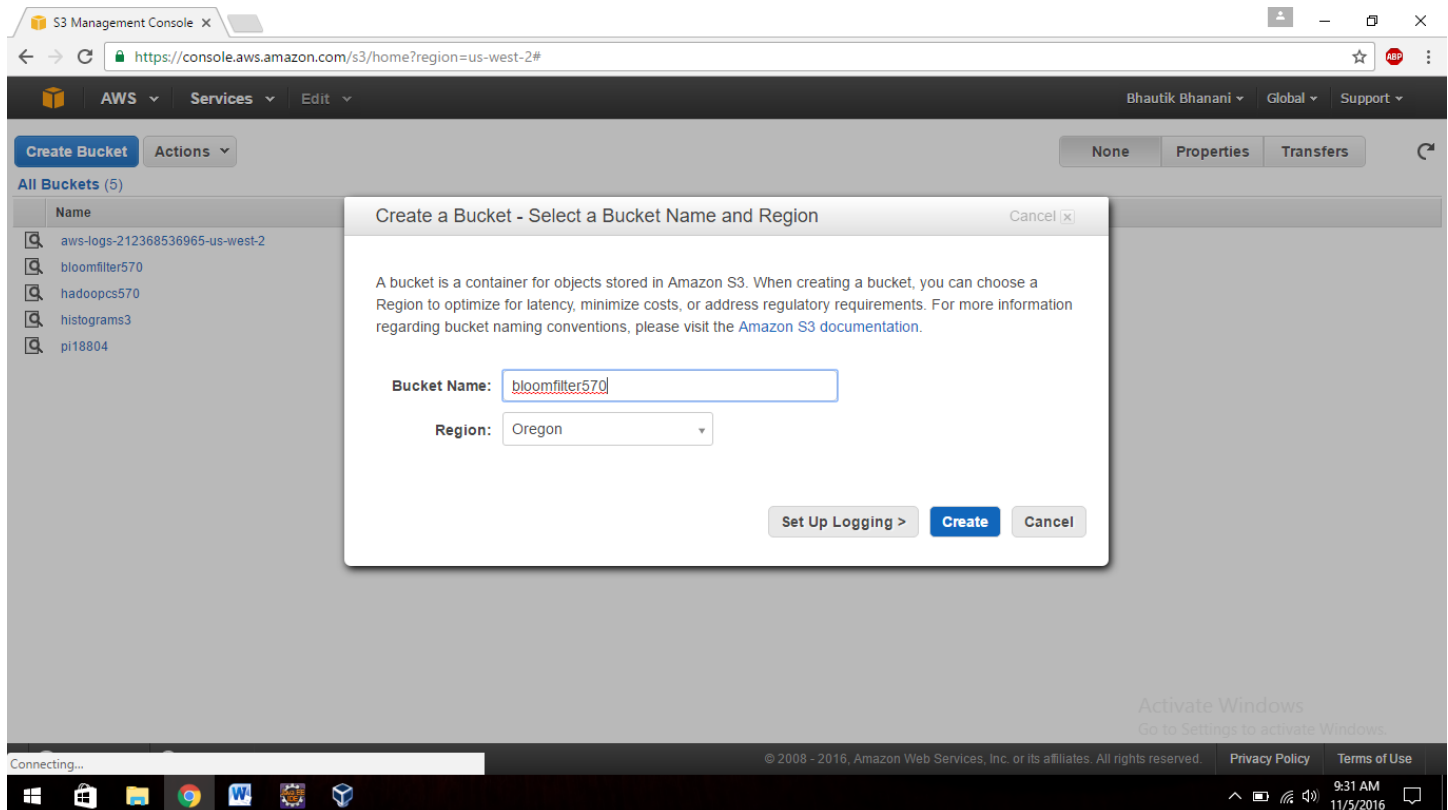


Now run StackOverFlow Indexing project on AWS

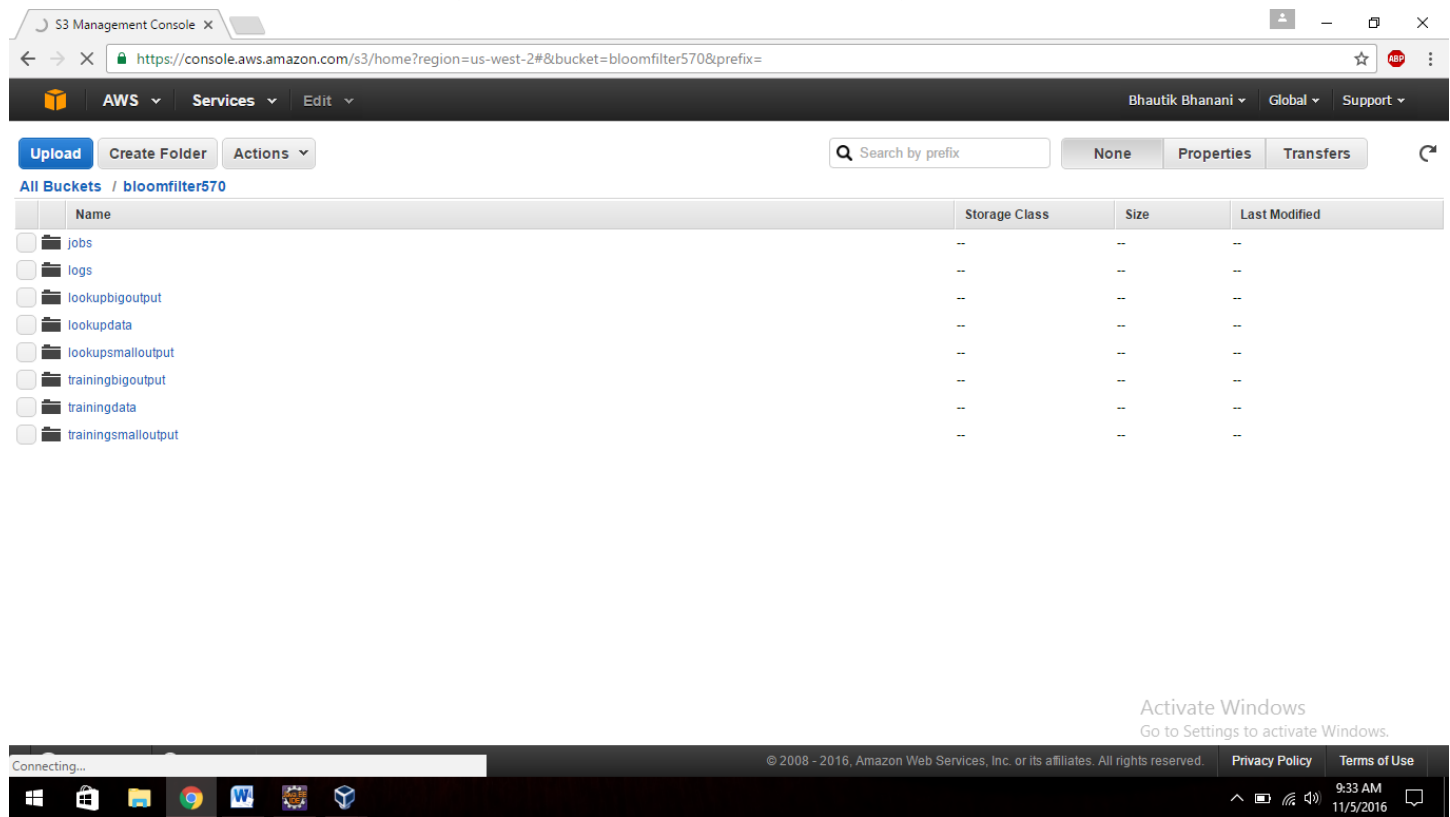
Step 1: Login to your AWS account and AWS dashboard.



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.

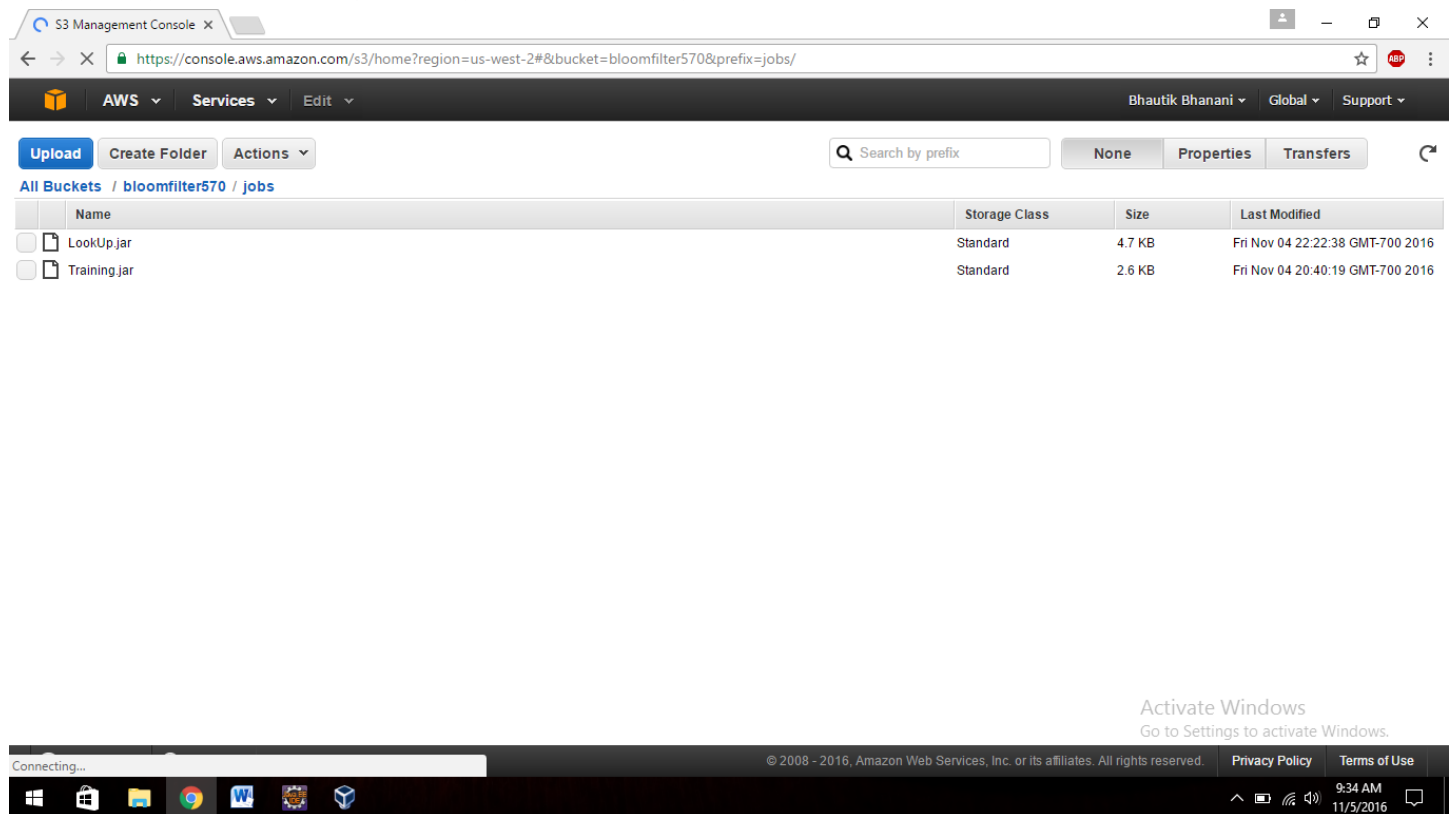


The screenshot shows the AWS S3 Management Console interface. The browser address bar displays the URL: `https://console.aws.amazon.com/s3/home?region=us-west-2#&bucket=bloomfilter570&prefix=`. The console header includes the AWS logo, navigation tabs (AWS, Services, Edit), and user information (Bhautik Bhanani, Global, Support). Below the header, there are buttons for 'Upload', 'Create Folder', and 'Actions', along with a search bar labeled 'Search by prefix'. The main content area shows the bucket 'bloomfilter570' with a table of its contents:

| Name | Storage Class | Size | Last Modified |
|-------------------|---------------|------|---------------|
| jobs | -- | -- | -- |
| logs | -- | -- | -- |
| lookupbigoutput | -- | -- | -- |
| lookupdata | -- | -- | -- |
| lookupsmloutput | -- | -- | -- |
| trainingbigoutput | -- | -- | -- |
| trainingdata | -- | -- | -- |
| trainingsmloutput | -- | -- | -- |

At the bottom of the console, there is a Windows taskbar with various application icons and a system tray showing the time as 9:33 AM on 11/5/2016. An 'Activate Windows' watermark is visible in the bottom right corner.

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



The screenshot shows the AWS S3 Management Console interface with the 'jobs' folder selected. The browser address bar displays the URL: `https://console.aws.amazon.com/s3/home?region=us-west-2#&bucket=bloomfilter570&prefix=jobs/`. The console header and navigation tabs are the same as in the previous screenshot. The main content area shows the bucket 'bloomfilter570' with the 'jobs' folder selected. Below the folder name, there is a table of its contents:

| Name | Storage Class | Size | Last Modified |
|--------------|---------------|--------|----------------------------------|
| LookUp.jar | Standard | 4.7 KB | Fri Nov 04 22:22:38 GMT-700 2016 |
| Training.jar | Standard | 2.6 KB | Fri Nov 04 20:40:19 GMT-700 2016 |

At the bottom of the console, there is a Windows taskbar with various application icons and a system tray showing the time as 9:34 AM on 11/5/2016. An 'Activate Windows' watermark is visible in the bottom right corner.

Step 5: Upload input file in input folder.

Training Data:

S3 Management Console

https://console.aws.amazon.com/s3/home?region=us-west-2#&bucket=bloomfilter570&prefix=trainingdata/

AWS Services Edit

Bhautik Bhanani Global Support

Upload Create Folder Actions

Search by prefix

None Properties Transfers

All Buckets / bloomfilter570 / trainingdata

| | Name | Storage Class | Size | Last Modified |
|--------------------------|------------------|---------------|-----------|----------------------------------|
| <input type="checkbox"/> | hotlistbig.txt | Standard | 921 bytes | Fri Nov 04 21:46:39 GMT-700 2016 |
| <input type="checkbox"/> | hotlistsmall.txt | Standard | 20 bytes | Fri Nov 04 20:44:03 GMT-700 2016 |

Activate Windows
Go to Settings to activate Windows.

Feedback English © 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

9:34 AM 11/5/2016

Lookup data:

S3 Management Console

https://console.aws.amazon.com/s3/home?region=us-west-2#&bucket=bloomfilter570&prefix=lookupdata/

AWS Services Edit

Bhautik Bhanani Global Support

Upload Create Folder Actions

Search by prefix

None Properties Transfers

All Buckets / bloomfilter570 / lookupdata

| | Name | Storage Class | Size | Last Modified |
|--------------------------|------------------------------|---------------|-------|----------------------------------|
| <input type="checkbox"/> | Comments.xml | Standard | 31 MB | Fri Nov 04 20:43:39 GMT-700 2016 |
| <input type="checkbox"/> | Comments_10_no_wikipedia.xml | Standard | 3 KB | Fri Nov 04 20:43:39 GMT-700 2016 |

Activate Windows
Go to Settings to activate Windows.

Feedback English © 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

9:34 AM 11/5/2016

Step 6: Now go to EC2 from AWS dashboard.

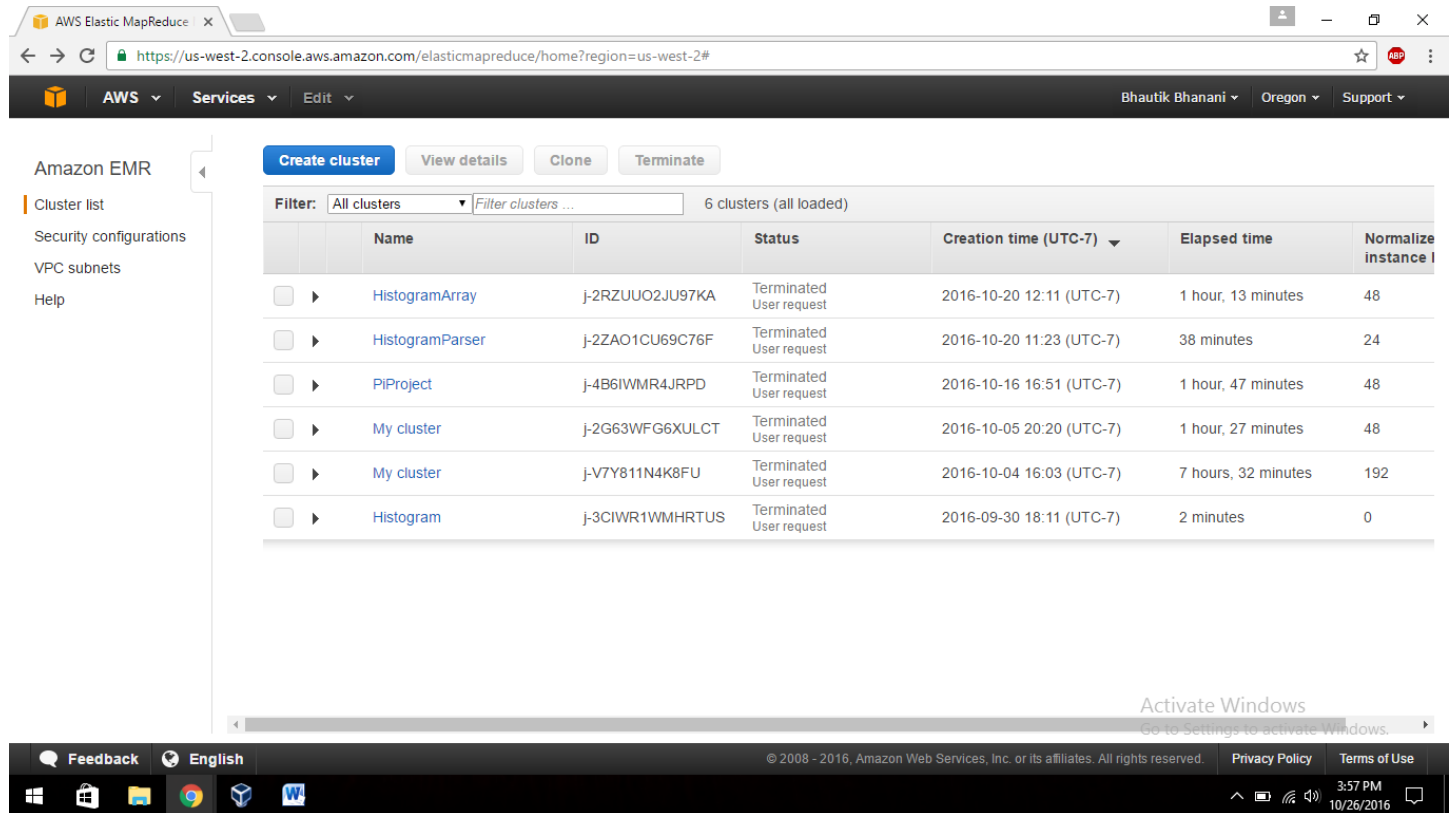
The screenshot shows the AWS Management Console for the 'us-west-2' region. The left-hand navigation pane is expanded, showing the 'Key Pairs' section under 'NETWORK & SECURITY'. The main content area displays the 'Key Pairs' page with buttons for 'Create Key Pair', 'Import Key Pair', and 'Delete'. Below these buttons is a search bar and a table of existing key pairs. The table has two columns: 'Key pair name' and 'Fingerprint'. There are three key pairs listed: 'HadoopCS570', 'histogramKeyPair', and 'piProjectKeyPair'. At the bottom of the page, there is an 'Activate Windows' watermark and a system taskbar at the very bottom showing the time as 11:41 AM on 10/20/2016.

| Key pair name | Fingerprint |
|------------------|---|
| HadoopCS570 | da:f3:b8:4c:4b:3d:28:b2:8c:65:c6:fe:fa:83:1e:78:ef:96:21:e0 |
| histogramKeyPair | c4:fa:7b:a9:65:40:66:85:e5:ca:38:e7:ab:ca:eb:71:58:d9:5c:cc |
| piProjectKeyPair | 34:83:0f:3c:d7:1b:4a:77:99:57:3e:f8:28:7f:42:ed:2f:b7:b7:83 |

Step 7: Create key pair for your project.

This screenshot shows the same AWS Management Console interface as the previous one, but with a 'Create Key Pair' dialog box open in the center. The dialog box has a title bar with a close button (X). Inside, there is a label 'Key pair name:' followed by a text input field containing the text 'HadoopCS570'. Below the input field are two buttons: 'Cancel' and 'Create'. The background content of the console is dimmed. The system taskbar at the bottom remains the same, showing the time as 11:41 AM on 10/20/2016.

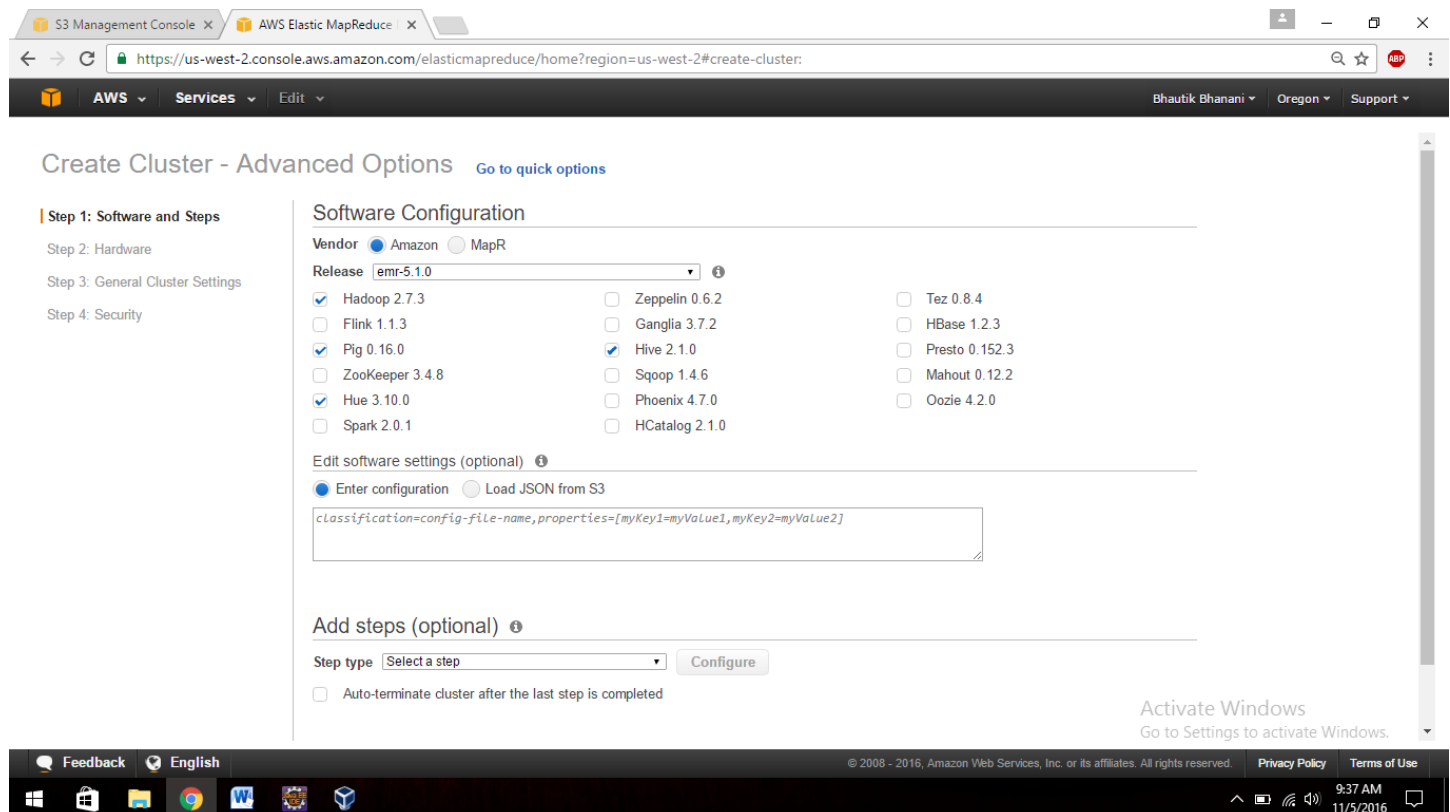
Step 8: Go to EMR from AWS dashboard.



The screenshot shows the AWS Elastic MapReduce console. The left sidebar contains links for Amazon EMR, Cluster list, Security configurations, VPC subnets, and Help. The main area displays a table of clusters with columns: Name, ID, Status, Creation time (UTC-7), Elapsed time, and Normalize instance I. The table lists six clusters, all with a status of 'Terminated User request'. The bottom of the screen shows a Windows taskbar with various application icons and a system clock indicating 3:57 PM on 10/26/2016.

| Name | ID | Status | Creation time (UTC-7) | Elapsed time | Normalize instance I |
|-----------------|-----------------|-------------------------|--------------------------|---------------------|----------------------|
| HistogramArray | j-2RZUUO2JU97KA | Terminated User request | 2016-10-20 12:11 (UTC-7) | 1 hour, 13 minutes | 48 |
| HistogramParser | j-2ZAO1CU69C76F | Terminated User request | 2016-10-20 11:23 (UTC-7) | 38 minutes | 24 |
| PiProject | j-4B6IWMR4JRPD | Terminated User request | 2016-10-16 16:51 (UTC-7) | 1 hour, 47 minutes | 48 |
| My cluster | j-2G63WFG6XULCT | Terminated User request | 2016-10-05 20:20 (UTC-7) | 1 hour, 27 minutes | 48 |
| My cluster | j-V7Y811N4K8FU | Terminated User request | 2016-10-04 16:03 (UTC-7) | 7 hours, 32 minutes | 192 |
| Histogram | j-3CIWR1WMHRTUS | Terminated User request | 2016-09-30 18:11 (UTC-7) | 2 minutes | 0 |

Step 9: Create new cluster of Bloom Filter project.



The screenshot shows the 'Create Cluster - Advanced Options' page in the AWS Elastic MapReduce console. The left sidebar lists steps: Step 1: Software and Steps, Step 2: Hardware, Step 3: General Cluster Settings, and Step 4: Security. The main area is titled 'Software Configuration' and includes options for Vendor (Amazon), Release (emr-5.1.0), and a list of software packages (Hadoop, Flink, Pig, ZooKeeper, Hue, Spark, Zeppelin, Ganglia, Hive, Sqoop, Phoenix, HCatalog, Tez, HBase, Presto, Mahout, Oozie). The 'Add steps (optional)' section is also visible. The bottom of the screen shows a Windows taskbar with various application icons and a system clock indicating 9:37 AM on 11/5/2016.

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:

☒ Hadoop 2.7.3 ☐ Zeppelin 0.6.2 ☐ Tez 0.8.4

☐ Flink 1.1.3 ☐ Ganglia 3.7.2 ☐ HBase 1.2.3

☒ Pig 0.16.0 ☒ Hive 2.1.0 ☐ Presto 0.152.3

☐ ZooKeeper 3.4.8 ☐ Sqoop 1.4.6 ☐ Mahout 0.12.2

☒ Hue 3.10.0 ☐ Phoenix 4.7.0 ☐ Oozie 4.2.0

☐ Spark 2.0.1 ☐ HCatalog 2.1.0

Edit software settings (optional) ⓘ

☒ Enter configuration ☐ Load JSON from S3

Add steps (optional) ⓘ

Step type:

☐ Auto-terminate cluster after the last step is completed

Add one training small data face as custom jar.

S3 Management Console x AWS Elastic MapReduce x AWS Elastic MapReduce x

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

Software Configuration

Vendor ☒ Amazon ☐ MapR

Add Step

Step type Custom JAR

Name* Training Face

JAR location* s3://bloomfilter570/jobs/Training.jar

Arguments s3://bloomfilter570/trainingdata/hotlistsmall.txt 3 0.01
s3://bloomfilter570/trainingsmalloutput/hotlist.blm

Action on failure Continue What to do if the step fails.

Cancel Add

JAR location maybe a path into S3 or a fully qualified java class in the classpath.

These are passed to the main function in the JAR. If the JAR does not specify a main class in its manifest file you can specify another class name as the first argument.

Add steps (optional) ⓘ

Step type Custom JAR Configure

☐ Auto-terminate cluster after the last step is completed

Feedback English

© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

9:38 AM 11/5/2016

S3 Management Console x AWS Elastic MapReduce x AWS Elastic MapReduce x

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

Hardware Configuration ⓘ

If you need more than 20 EC2 instances, [complete this form](#).

Network vpc-2dabba49 (172.31.0.0/16) (default) [Create a VPC](#) ⓘ

EC2 Subnet subnet-68f6230 | Default in us-west-2c

| Type | Name | EC2 instance type | Instance count | Storage per instance | Request spot | Bid price |
|--------|---------------------------|-------------------|----------------|---|--------------------------|-----------|
| Master | Master instance group - 1 | m3.xlarge | 1 | 80 GiB Add EBS volumes | <input type="checkbox"/> | |
| Core | Core instance group - 2 | m3.xlarge | 2 | 80 GiB Add EBS volumes | <input type="checkbox"/> | |
| Task | Task instance group - 3 | m3.xlarge | 0 | 80 GiB Add EBS volumes | <input type="checkbox"/> | |

[Add task instance group](#)

Cancel Previous Next

Activate Windows
[Go to Settings to activate Windows.](#)

Feedback English

© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

9:39 AM 11/5/2016

S3 Management Console x AWS Elastic MapReduce x AWS Elastic MapReduce x

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

Search Star AWS

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

☒ Logging ⓘ
S3 folder

☒ Debugging ⓘ

☒ Termination protection ⓘ

Tags ⓘ

| Key | Value (optional) |
|--|----------------------|
| <input type="text" value="Add a key to create a tag"/> | <input type="text"/> |

Additional Options

☐ EMRFS consistent view ⓘ

▶ Bootstrap Actions

Cancel Previous Next

Activate Windows
Go to Settings to activate Windows.

S3 Management Console x AWS Elastic MapReduce x AWS Elastic MapReduce x

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

Search Star AWS

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

☒ 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

EC2 instance profile

▶ Encryption Options

▶ EC2 Security Groups

Cancel Previous Create cluster

Activate Windows
Go to Settings to activate Windows.

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

Amazon EMR

- Cluster list
- Security configurations
- VPC subnets
- Help

Monitoring

Hardware

Steps

Add step Clone step

Steps

Filter: All steps Filter steps ... 6 steps (all loaded)

| | ID | Name | Status | Start time (UTC-8) | Elapsed time | Log files | Action |
|--|-----------------|------------------------|-----------|--------------------------|--------------|--|------------------------|
| | s-3T8YTFIS43JJC | Lookup big | Completed | 2016-11-04 22:29 (UTC-8) | 34 seconds | View logs | View j |
| | s-2PWY1KT92HVDZ | LookupSmall | Completed | 2016-11-04 21:59 (UTC-8) | 28 seconds | View logs | View j |
| | s-3MQ4TLIKAGTBZ | Training Big | Completed | 2016-11-04 21:47 (UTC-8) | 6 seconds | View logs | View j |
| JAR location: s3://bloomfilter570/jobs/Training.jar | | | | | | | |
| Main class: None | | | | | | | |
| Arguments: s3://bloomfilter570/trainingdata/hotlistbig.txt 99 0.01 s3://bloomfilter570/trainingbigoutput/hotlist.blm | | | | | | | |
| Action on failure: Continue | | | | | | | |
| | s-1CEU1HISQ1RAQ | Training Big | Failed | 2016-11-04 21:36 (UTC-8) | 6 seconds | controller syslog* stderr stdout | View j |
| | s-29RQGPZ2Q68J | Training Jar | Completed | 2016-11-04 21:22 (UTC-8) | 6 seconds | View logs | View j |
| | s-JH4O38TMD8N | Setup hadoop debugging | Completed | 2016-11-04 21:21 (UTC-8) | 2 seconds | View logs | View j |

Activate Windows
Go to Settings to activate Windows.

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

Training Bloom filter of size 28 with 6 hash functions, 3 approximate number of records, and 0.01 false positive rate

Reading s3://bloomfilter570/trainingdata/hotlistsmall.txt

line window

line Linux

line memory

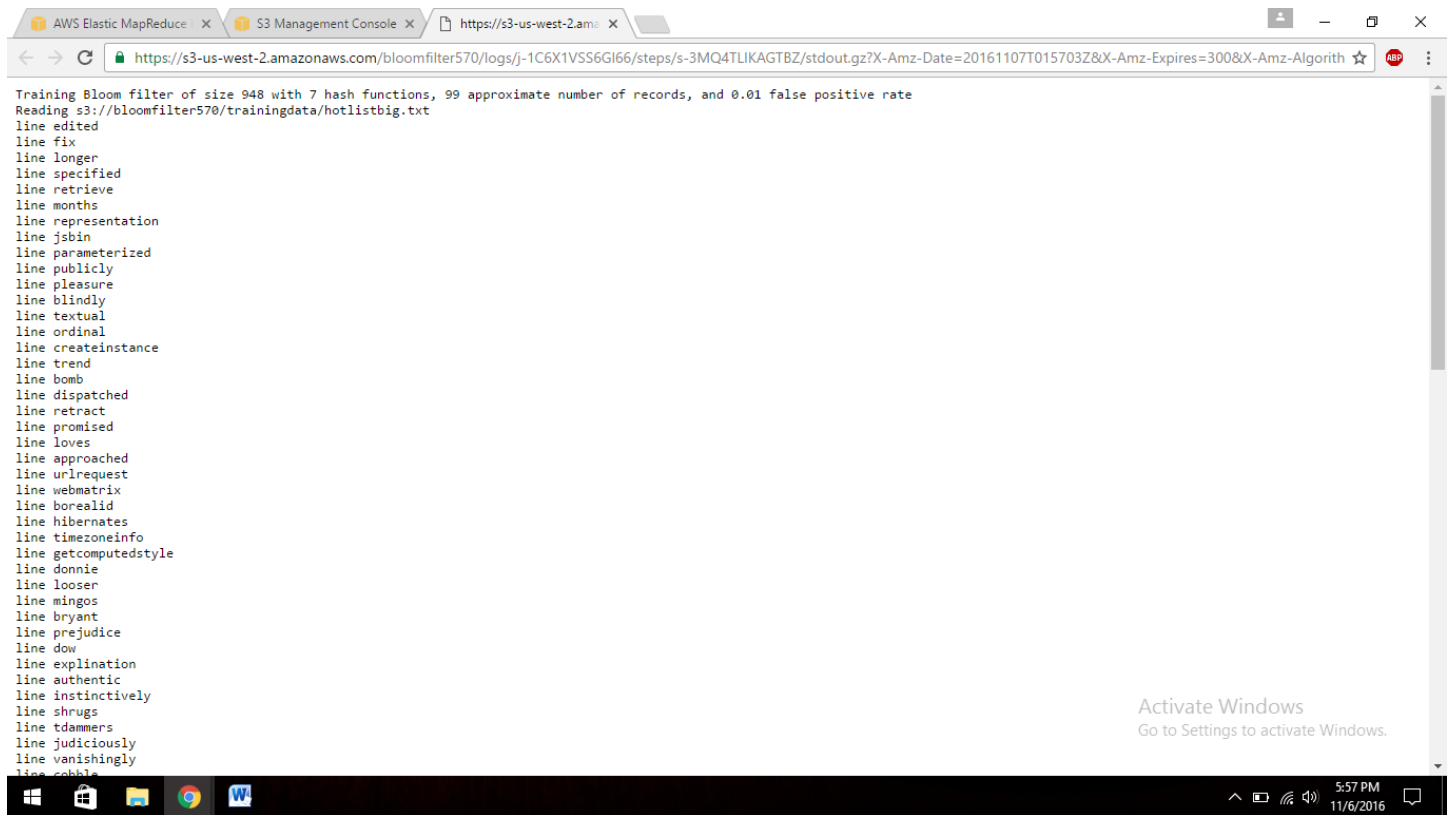
Trained Bloom filter with 3 entries.

Serializing Bloom filter to HDFS at s3://bloomfilter570/trainingsmalloutput/hotlist.blm

Done training Bloom filter.

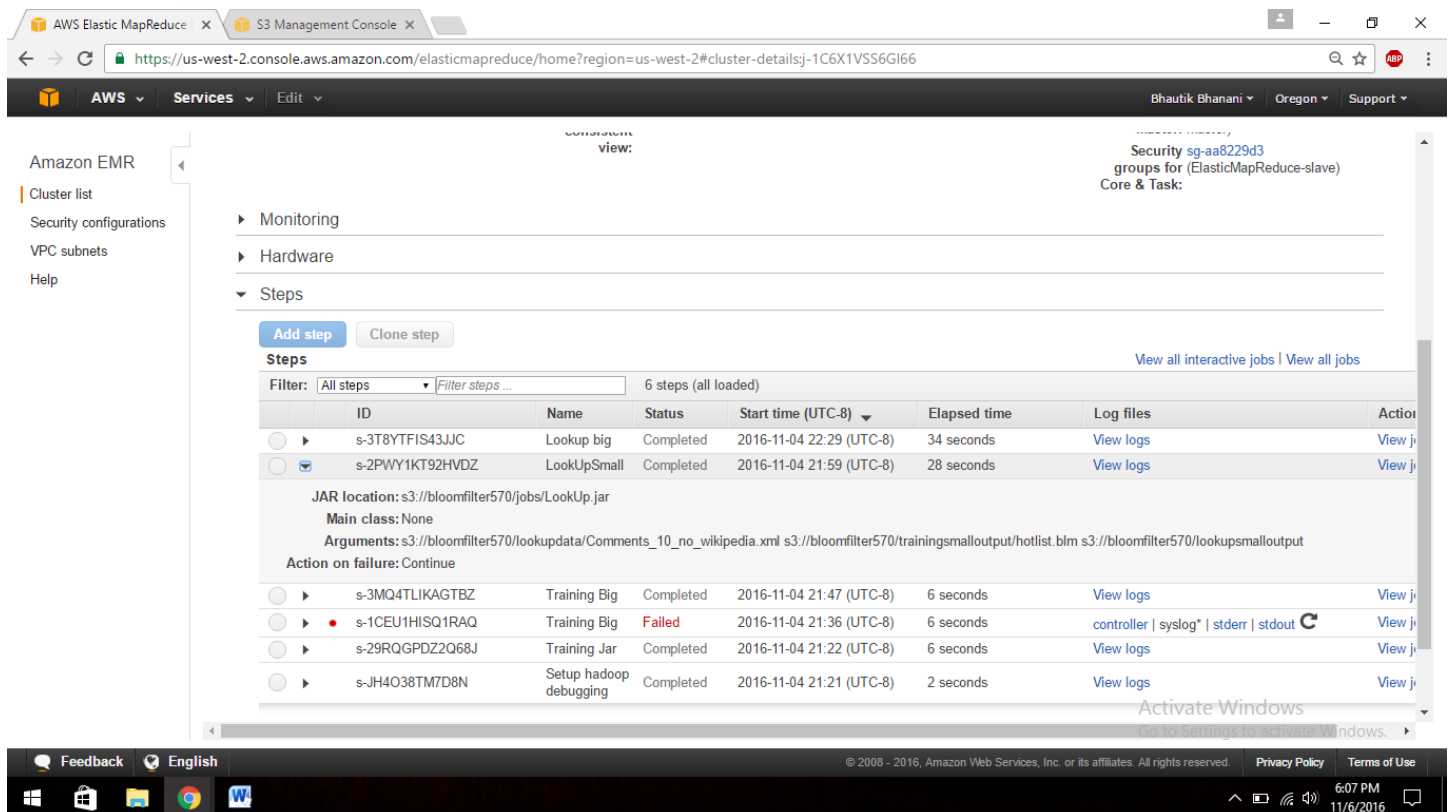
Activate Windows
Go to Settings to activate Windows.

Big data:

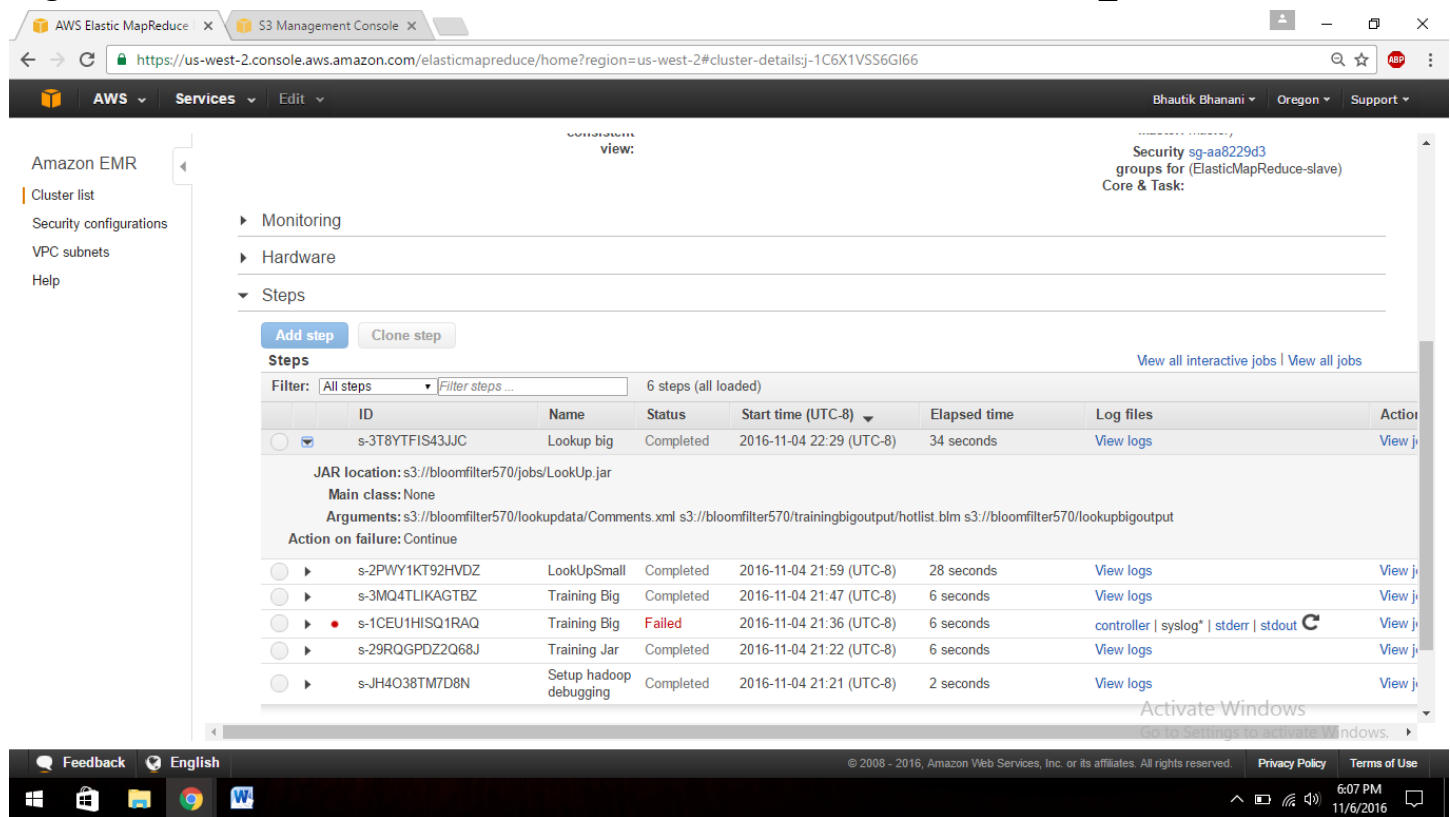


Step 12: After this, now create steps for lookup face for small and big data.

Argument: `"/%PATH%/comment_10_no_wikipedia.xml /%PATH%/hotlist.blm /%OUTPUT_PATH%/"`



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

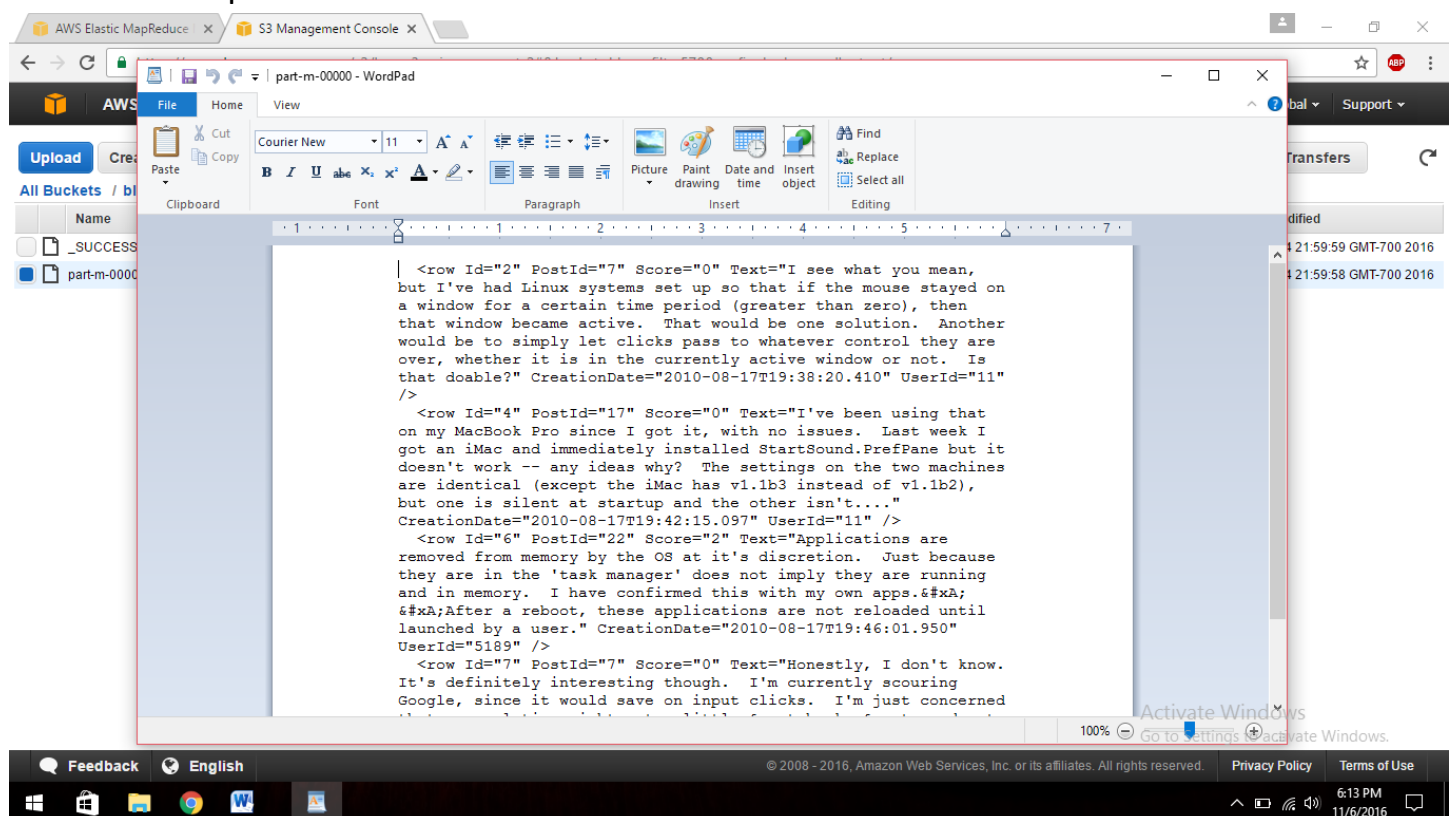


The screenshot shows the AWS Elastic MapReduce console. The left sidebar lists navigation options: Amazon EMR, Cluster list, Security configurations, VPC subnets, and Help. The main content area shows the 'Steps' section for a cluster. The 'Lookup big' step is completed, and the 'Training Big' step is failed. The 'Failed' status is highlighted in red. The console also shows the 'Add step' and 'Clone step' buttons, and a table of steps with columns for ID, Name, Status, Start time, Elapsed time, Log files, and Action.

| ID | Name | Status | Start time (UTC-8) | Elapsed time | Log files | Action |
|-----------------|------------------------|-----------|--------------------------|--------------|--|---------------------------|
| s-3T8YTFIS43JJC | Lookup big | Completed | 2016-11-04 22:29 (UTC-8) | 34 seconds | View logs | View jobs |
| s-2PWY1KT92HVDZ | LookUpSmall | Completed | 2016-11-04 21:59 (UTC-8) | 28 seconds | View logs | View jobs |
| s-3MQ4TLIKAGTBZ | Training Big | Completed | 2016-11-04 21:47 (UTC-8) | 6 seconds | View logs | View jobs |
| s-1CEU1HISQ1RAQ | Training Big | Failed | 2016-11-04 21:36 (UTC-8) | 6 seconds | controller syslog* stderr stdout | View jobs |
| s-29RQPDZ2Q68J | Training Jar | Completed | 2016-11-04 21:22 (UTC-8) | 6 seconds | View logs | View jobs |
| s-JH4O38TM7D8N | Setup hadoop debugging | Completed | 2016-11-04 21:21 (UTC-8) | 2 seconds | View logs | View jobs |

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

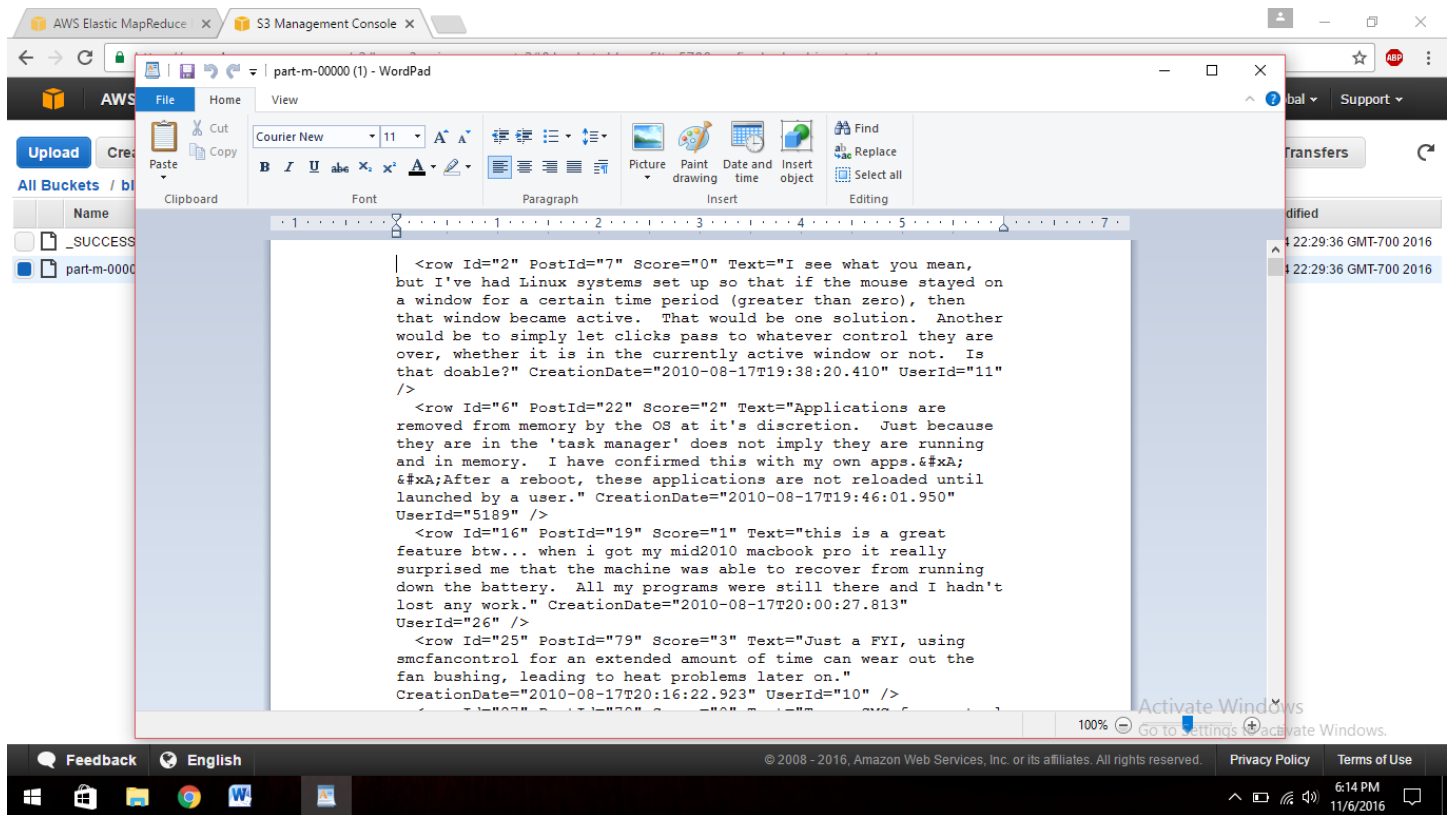
Small data output:



The screenshot shows a WordPad window titled 'part-m-00000 - WordPad'. The text content is a JSON array of objects, each representing a post. The text is as follows:

```
[{"row Id": "2" PostId="7" Score="0" Text="I see what you mean, but I've had Linux systems set up so that if the mouse stayed on a window for a certain time period (greater than zero), then that window became active. That would be one solution. Another would be to simply let clicks pass to whatever control they are over, whether it is in the currently active window or not. Is that doable?" CreationDate="2010-08-17T19:38:20.410" UserId="11" />}, {"row Id": "4" PostId="17" Score="0" Text="I've been using that on my MacBook Pro since I got it, with no issues. Last week I got an iMac and immediately installed StartSound.PrefPane but it doesn't work -- any ideas why? The settings on the two machines are identical (except the iMac has v1.1b3 instead of v1.1b2), but one is silent at startup and the other isn't...." CreationDate="2010-08-17T19:42:15.097" UserId="11" />}, {"row Id": "6" PostId="22" Score="2" Text="Applications are removed from memory by the OS at it's discretion. Just because they are in the 'task manager' does not imply they are running and in memory. I have confirmed this with my own apps.&#xA;After a reboot, these applications are not reloaded until launched by a user." CreationDate="2010-08-17T19:46:01.950" UserId="5189" />}, {"row Id": "7" PostId="7" Score="0" Text="Honestly, I don't know. It's definitely interesting though. I'm currently scouring Google, since it would save on input clicks. I'm just concerned"}]
```

Big data output:



Step 14: After this, terminate cluster.

