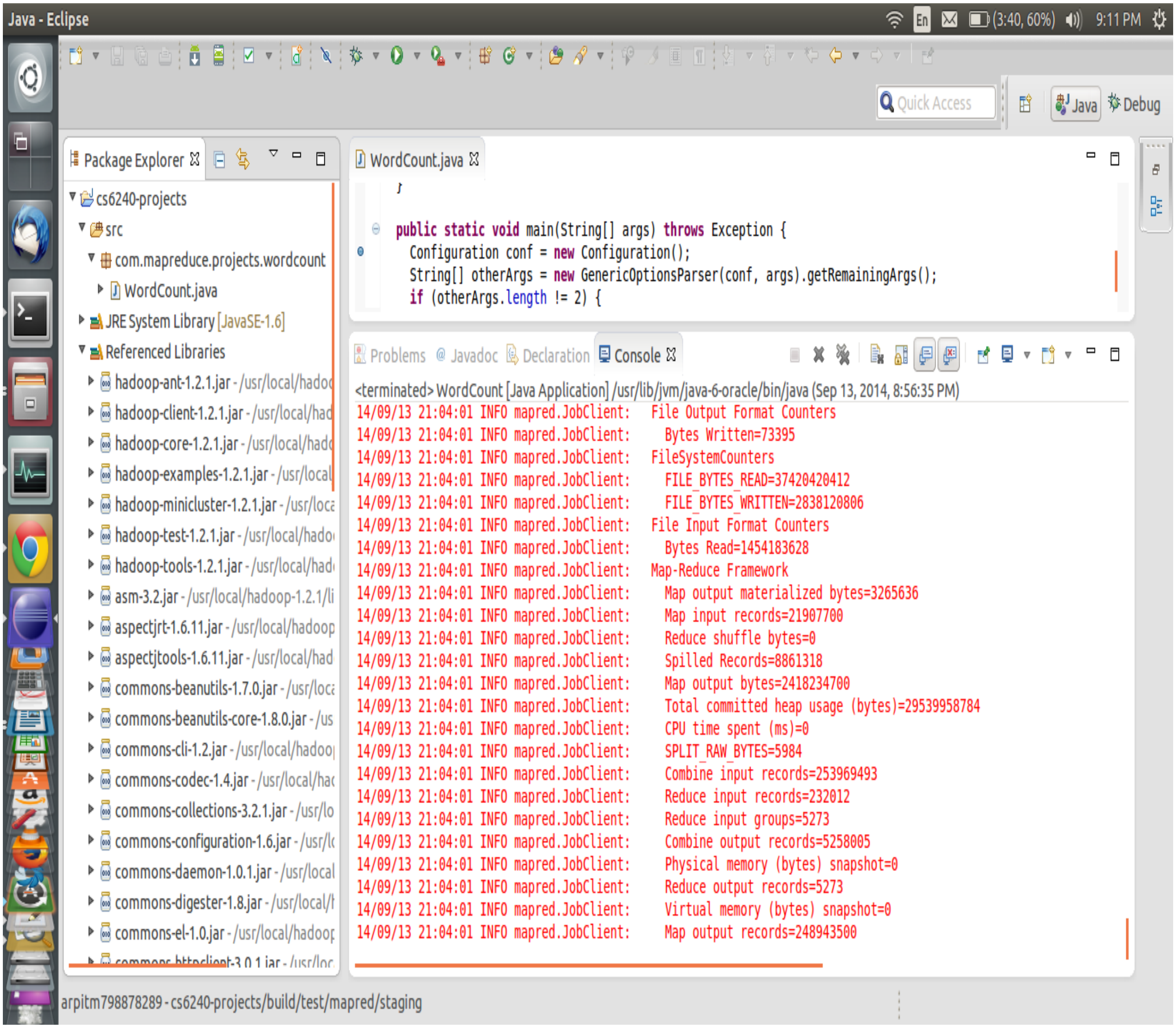
CS6240 – PARALLEL DATA PROCESSING IN MAP-REDUCE

Class: CS6240-02

Homework Number: 1

Name: Arpit Mehta

SCREENSHOT 1 – LOCAL EXECUTION: The below screen shot shows the project directory structure and console output of successful run of WordCount program inside the IDE:


Java - Eclipse
En
(3:40, 60%)
9:11 PM
Quick Access
Java
Debug
Package Explorer
cs6240-projects
src
com.mapreduce.projects.wordcount
WordCount.java
JRE System Library [JavaSE-1.6]
Referenced Libraries
hadoop-ant-1.2.1.jar - /usr/local/hadoo
hadoop-client-1.2.1.jar - /usr/local/had
hadoop-core-1.2.1.jar - /usr/local/hado
hadoop-examples-1.2.1.jar - /usr/local
hadoop-minicluster-1.2.1.jar - /usr/loca
hadoop-test-1.2.1.jar - /usr/local/hado
hadoop-tools-1.2.1.jar - /usr/local/had
asm-3.2.jar - /usr/local/hadoop-1.2.1/li
aspectjrt-1.6.11.jar - /usr/local/hadoop
aspectjtools-1.6.11.jar - /usr/local/had
commons-beanutils-1.7.0.jar - /usr/loca
commons-beanutils-core-1.8.0.jar - /us
commons-cli-1.2.jar - /usr/local/hadoo
commons-codec-1.4.jar - /usr/local/had
commons-collections-3.2.1.jar - /usr/lo
commons-configuration-1.6.jar - /usr/l
commons-daemon-1.0.1.jar - /usr/local
commons-digester-1.8.jar - /usr/local/h
commons-el-1.0.jar - /usr/local/hadoop
commons-httpclient-3.0.1.jar - /usr/loc
WordCount.java
Problems
Javadoc
Declaration
Console

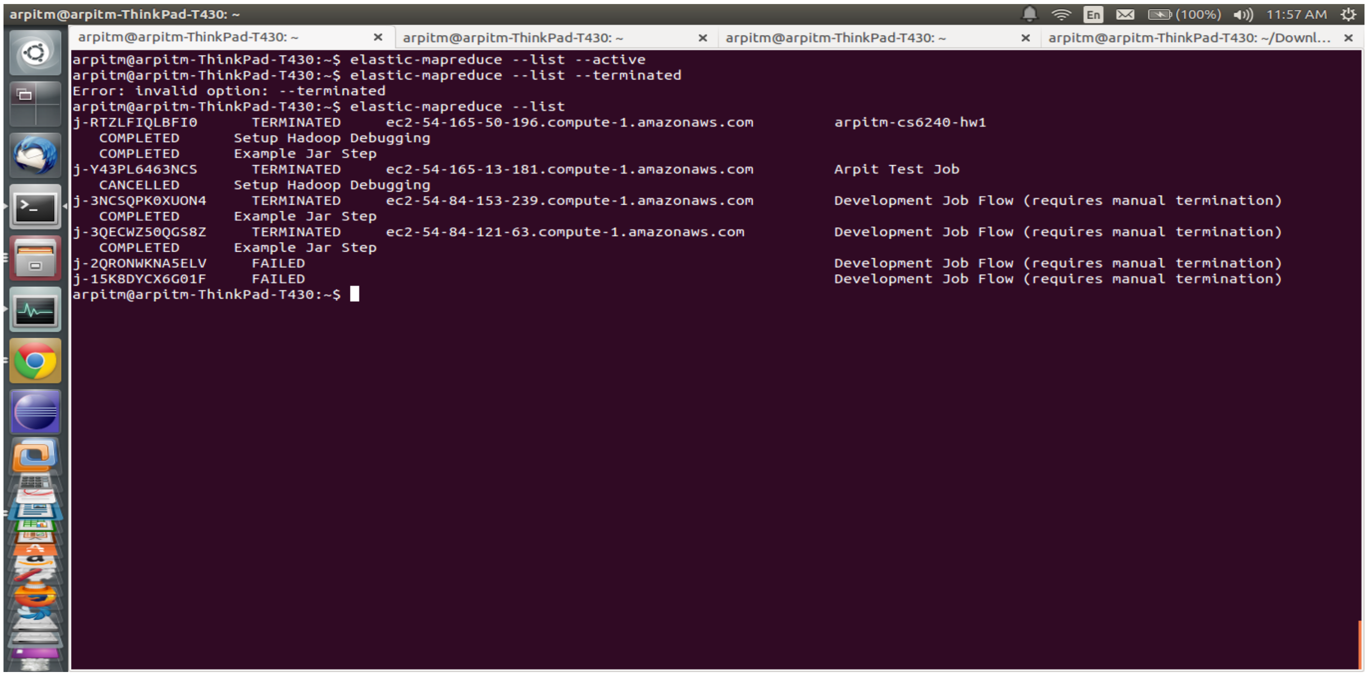
```
}

public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
  if (otherArgs.length != 2) {
```

```
<terminated> WordCount [Java Application] /usr/lib/jvm/java-6-oracle/bin/java (Sep 13, 2014, 8:56:35 PM)
14/09/13 21:04:01 INFO mapred.JobClient:   File Output Format Counters
14/09/13 21:04:01 INFO mapred.JobClient:     Bytes Written=73395
14/09/13 21:04:01 INFO mapred.JobClient:   FileSystemCounters
14/09/13 21:04:01 INFO mapred.JobClient:     FILE_BYTES_READ=37420420412
14/09/13 21:04:01 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=2838120806
14/09/13 21:04:01 INFO mapred.JobClient:   File Input Format Counters
14/09/13 21:04:01 INFO mapred.JobClient:     Bytes Read=1454183628
14/09/13 21:04:01 INFO mapred.JobClient:   Map-Reduce Framework
14/09/13 21:04:01 INFO mapred.JobClient:     Map output materialized bytes=3265636
14/09/13 21:04:01 INFO mapred.JobClient:     Map input records=21907700
14/09/13 21:04:01 INFO mapred.JobClient:     Reduce shuffle bytes=0
14/09/13 21:04:01 INFO mapred.JobClient:     Spilled Records=8861318
14/09/13 21:04:01 INFO mapred.JobClient:     Map output bytes=2418234700
14/09/13 21:04:01 INFO mapred.JobClient:     Total committed heap usage (bytes)=29539958784
14/09/13 21:04:01 INFO mapred.JobClient:     CPU time spent (ms)=0
14/09/13 21:04:01 INFO mapred.JobClient:     SPLIT_RAW_BYTES=5984
14/09/13 21:04:01 INFO mapred.JobClient:     Combine input records=253969493
14/09/13 21:04:01 INFO mapred.JobClient:     Reduce input records=232012
14/09/13 21:04:01 INFO mapred.JobClient:     Reduce input groups=5273
14/09/13 21:04:01 INFO mapred.JobClient:     Combine output records=5258005
14/09/13 21:04:01 INFO mapred.JobClient:     Physical memory (bytes) snapshot=0
14/09/13 21:04:01 INFO mapred.JobClient:     Reduce output records=5273
14/09/13 21:04:01 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
14/09/13 21:04:01 INFO mapred.JobClient:     Map output records=248943500
```

arpitm798878289 - cs6240-projects/build/test/mapred/staging

SCREENSHOT 2 – AWS EXECUTION:

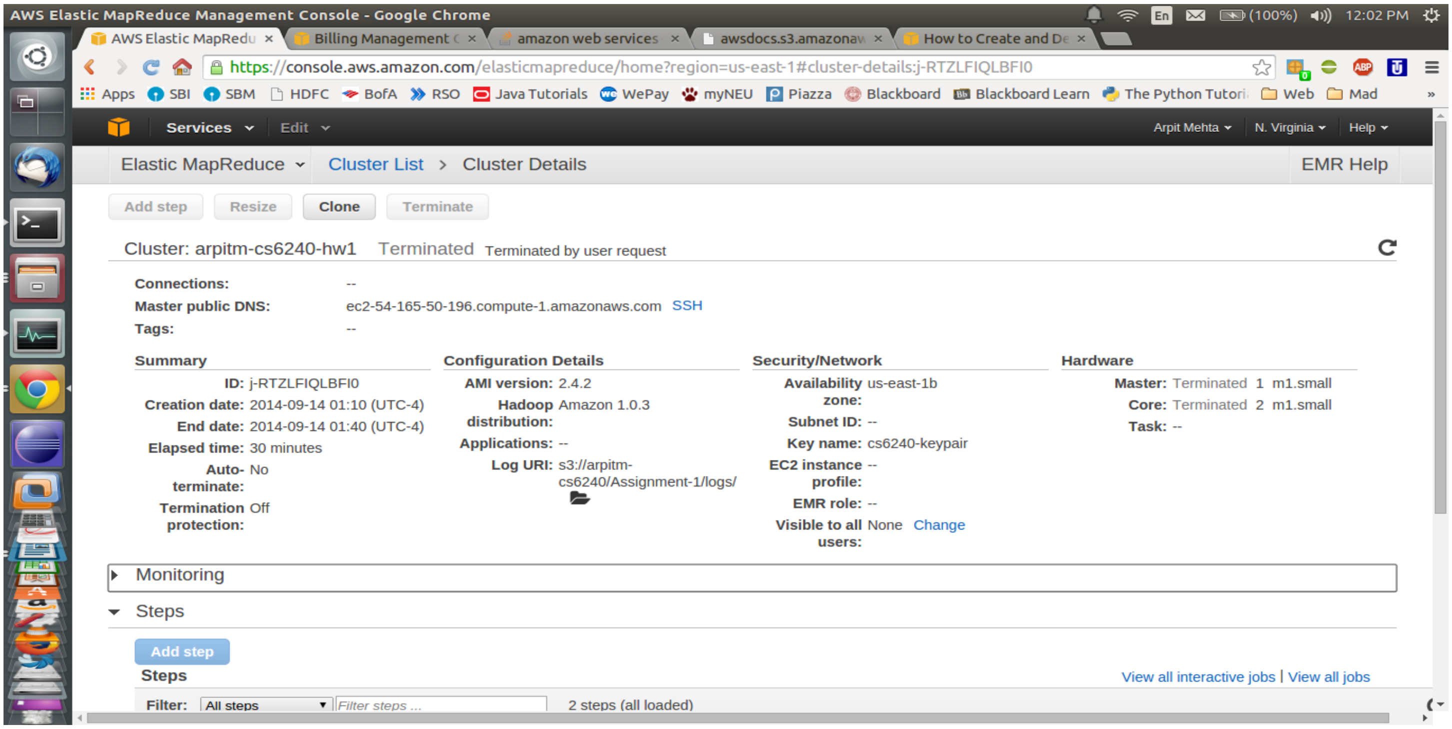
The below screen shot shows the list of jobflows that I created on AWS for running the custom wordcount.jar. The job id 'j-RTZLFIQLBFI0' with job name 'arpitm-cs6240-hw1' represents the jobflow of run using 3 small machines.



arpitm@arpitm-ThinkPad-T430: ~
En
(100%)
11:57 AM
arpitm@arpitm-ThinkPad-T430: ~
arpitm@arpitm-ThinkPad-T430: ~
arpitm@arpitm-ThinkPad-T430: ~
arpitm@arpitm-ThinkPad-T430: ~/Downl...

```
arpitm@arpitm-ThinkPad-T430:~$ elastic-mapreduce --list --active
arpitm@arpitm-ThinkPad-T430:~$ elastic-mapreduce --list --terminated
Error: invalid option: --terminated
arpitm@arpitm-ThinkPad-T430:~$ elastic-mapreduce --list
j-RTZLFIQLBFI0      TERMINATED     ec2-54-165-50-196.compute-1.amazonaws.com         arpitm-cs6240-hw1
   COMPLETED      Setup Hadoop Debugging
   COMPLETED      Example Jar Step
j-Y43PL6463NCS      TERMINATED     ec2-54-165-13-181.compute-1.amazonaws.com         Arpit Test Job
   CANCELLED      Setup Hadoop Debugging
j-3NCSQPK0XUON4     TERMINATED     ec2-54-84-153-239.compute-1.amazonaws.com         Development Job Flow (requires manual termination)
   COMPLETED      Example Jar Step
j-3QECWZ50QGS8Z     TERMINATED     ec2-54-84-121-63.compute-1.amazonaws.com          Development Job Flow (requires manual termination)
   COMPLETED      Example Jar Step
j-2QRONWKNA5ELV     FAILED                                                           Development Job Flow (requires manual termination)
j-15K8DYCX6G01F     FAILED                                                           Development Job Flow (requires manual termination)
arpitm@arpitm-ThinkPad-T430:~$ █
```

The below screenshot shows the AWS EMR 'Cluster Details' console detailing the job run.


AWS Elastic MapReduce Management Console - Google Chrome
12:02 PM
AWS Elastic MapRedu
Billing Management C
amazon web services
awsdocs.s3.amazonaw
How to Create and De
https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-RTZLFIQLBFI0
Apps SBI SBM HDFC BofA RSO Java Tutorials WePay myNEU Piazza Blackboard Blackboard Learn The Python Tutori Web Mad
Services
Edit
Arpit Mehta
N. Virginia
Help
Elastic MapReduce
Cluster List
Cluster Details
EMR Help
Add step
Resize
Clone
Terminate

Cluster: arpitm-cs6240-hw1 Terminated Terminated by user request

Connections: --
Master public DNS: ec2-54-165-50-196.compute-1.amazonaws.com SSH
Tags: --

Summary
ID: j-RTZLFIQLBFI0
Creation date: 2014-09-14 01:10 (UTC-4)
End date: 2014-09-14 01:40 (UTC-4)
Elapsed time: 30 minutes
Auto-terminate: No
Termination protection: Off

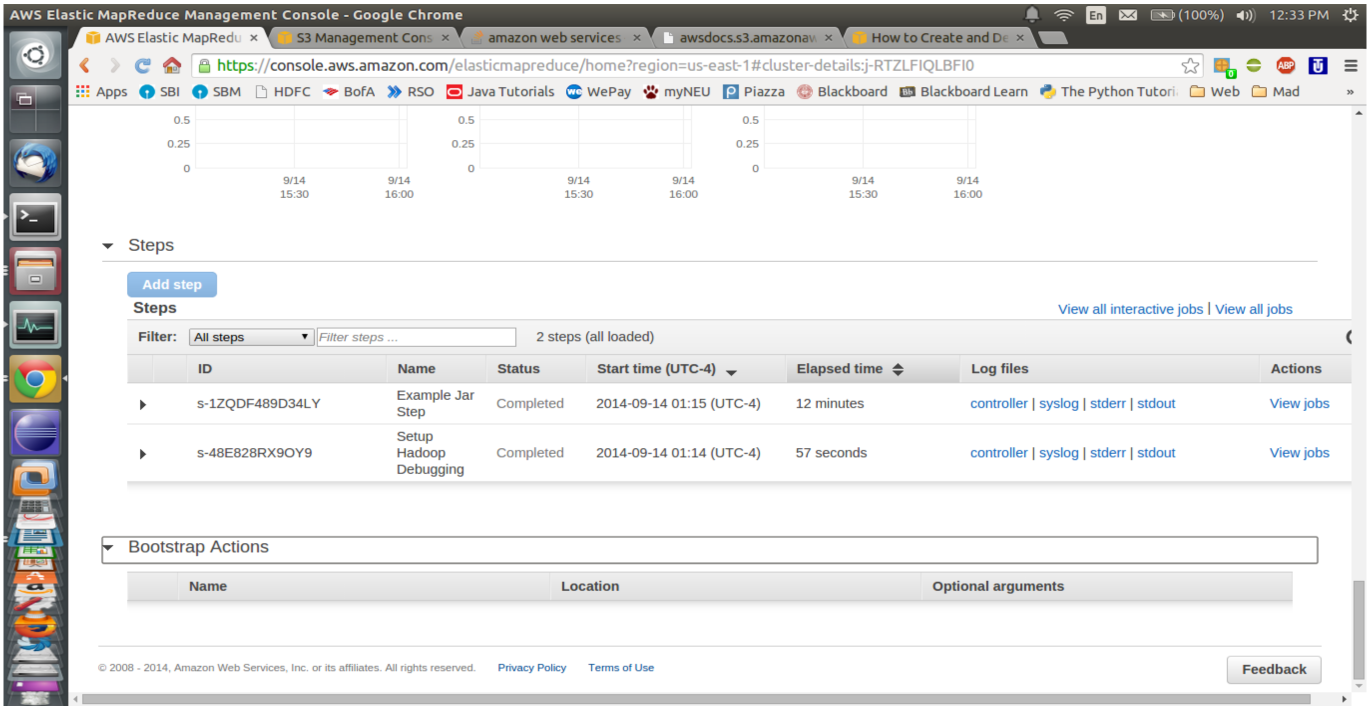
Configuration Details
AMI version: 2.4.2
Hadoop distribution: Amazon 1.0.3
Applications: --
Log URI: s3://arpitm-cs6240/Assignment-1/logs/

Security/Network
Availability zone: us-east-1b
Subnet ID: --
Key name: cs6240-keypair
EC2 instance profile: --
EMR role: --
Visible to all users: None Change

Hardware
Master: Terminated 1 m1.small
Core: Terminated 2 m1.small
Task: --

2 steps (all loaded)

The below screenshot shows the jobflow steps:


AWS Elastic MapReduce Management Console - Google Chrome
12:33 PM
AWS Elastic MapRedu
S3 Management Cons
amazon web services
awsdocs.s3.amazonaw
How to Create and De
https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-RTZLFIQLBFI0
Apps SBI SBM HDFC BofA RSO Java Tutorials WePay myNEU Piazza Blackboard Blackboard Learn The Python Tutori Web Mad
0.5
0.25
0
9/14 15:30
9/14 16:00
Steps
Add step
Steps
View all interactive jobs | View all jobs
Filter: All steps
Filter steps ...
2 steps (all loaded)

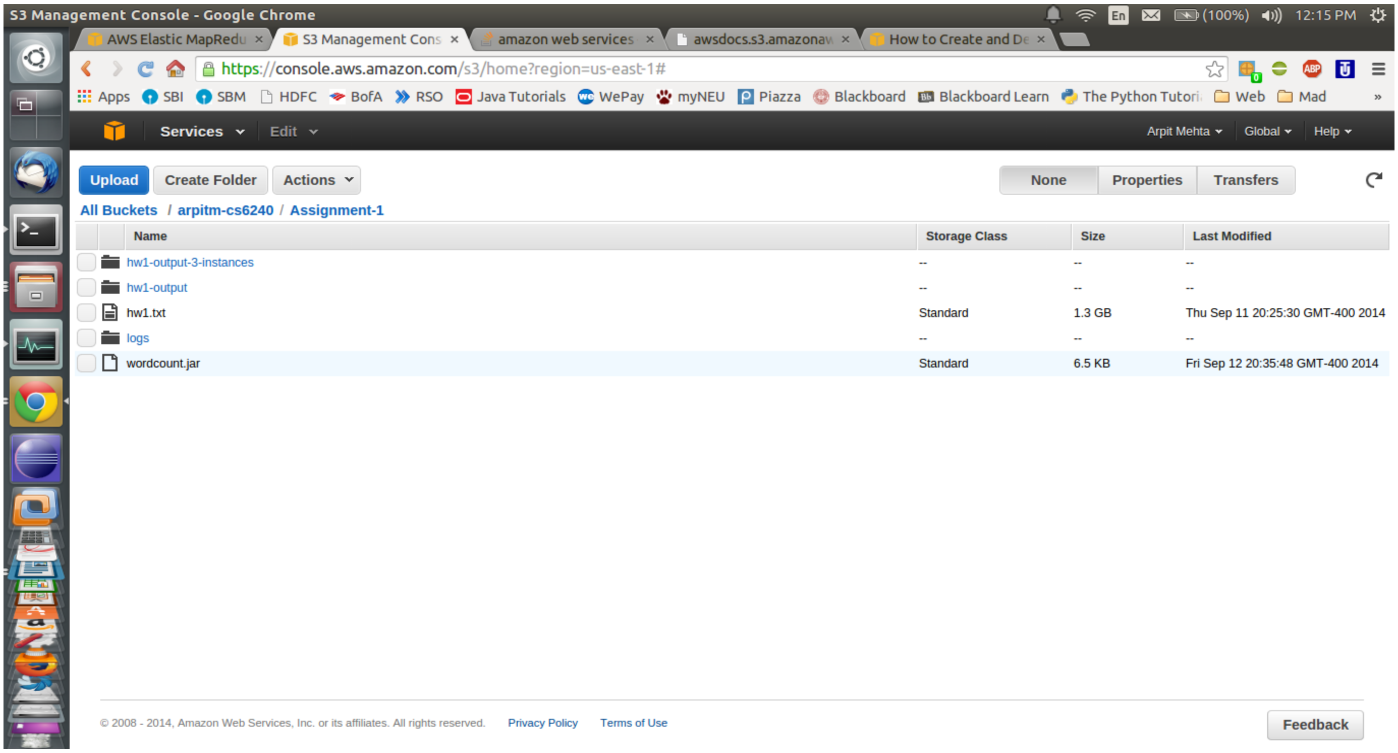
ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
s-1ZQDF489D34LY	Example Jar Step	Completed	2014-09-14 01:15 (UTC-4)	12 minutes	controller | syslog | stderr | stdout	View jobs
s-48E828RX9OY9	Setup Hadoop Debugging	Completed	2014-09-14 01:14 (UTC-4)	57 seconds	controller | syslog | stderr | stdout	View jobs

Bootstrap Actions

Name	Location	Optional arguments

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Feedback

The below screenshot shows my s3 bucket 'arpitm-cs6240' that contains the 'wordcount.jar', the input file 'hw1.txt', the 'logs' directory, the 'hw1-output-3instances' directory contains the output of the jobflow on 3 small machines.



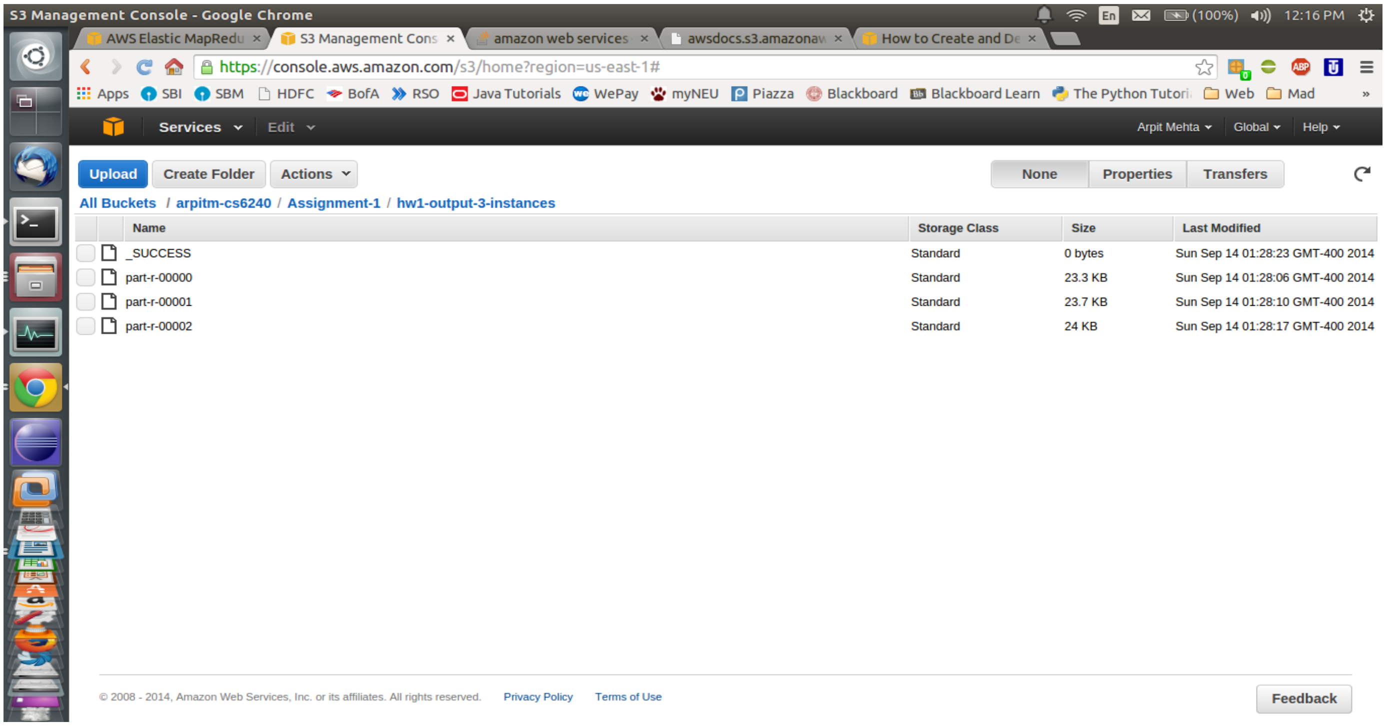
S3 Management Console - Google Chrome
https://console.aws.amazon.com/s3/home?region=us-east-1#
Services
Edit
Arpit Mehta
Global
Help
Upload
Create Folder
Actions
None
Properties
Transfers
Feedback

All Buckets / arpitm-cs6240 / Assignment-1

Name	Storage Class	Size	Last Modified
hw1-output-3-instances	--	--	--
hw1-output	--	--	--
hw1.txt	Standard	1.3 GB	Thu Sep 11 20:25:30 GMT-400 2014
logs	--	--	--
wordcount.jar	Standard	6.5 KB	Fri Sep 12 20:35:48 GMT-400 2014

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

The below screenshot shows 'hw1-output-3instances' containing the output files:



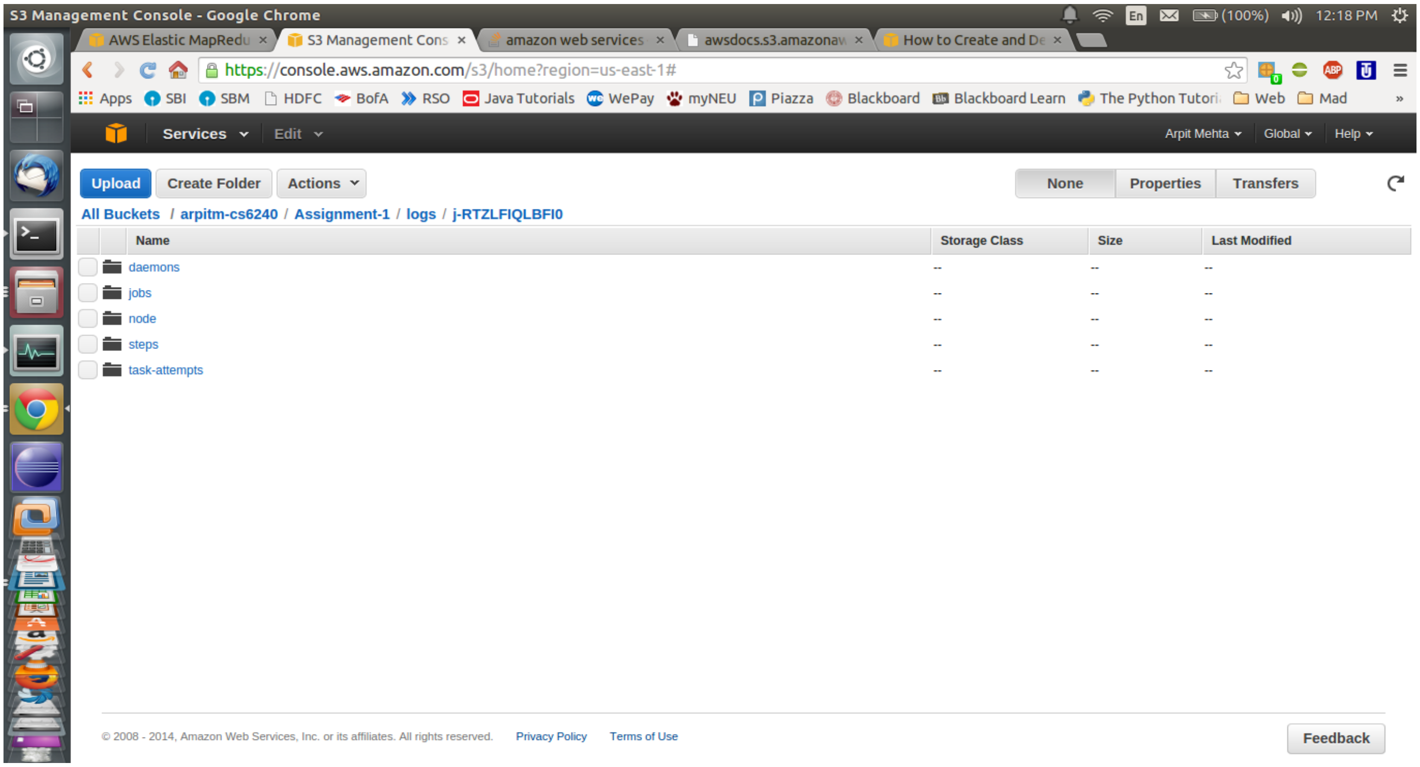
S3 Management Console - Google Chrome
https://console.aws.amazon.com/s3/home?region=us-east-1#
Services
Edit
Arpit Mehta
Global
Help
Upload
Create Folder
Actions
None
Properties
Transfers
Feedback

All Buckets / arpitm-cs6240 / Assignment-1 / hw1-output-3-instances

Name	Storage Class	Size	Last Modified
_SUCCESS	Standard	0 bytes	Sun Sep 14 01:28:23 GMT-400 2014
part-r-00000	Standard	23.3 KB	Sun Sep 14 01:28:06 GMT-400 2014
part-r-00001	Standard	23.7 KB	Sun Sep 14 01:28:10 GMT-400 2014
part-r-00002	Standard	24 KB	Sun Sep 14 01:28:17 GMT-400 2014

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

The below screenshot shows the directory content of jobflow 'j-RTZLFIQLBFI0' in the logs directory.



https://console.aws.amazon.com/s3/home?region=us-east-1#

All Buckets / arpitm-cs6240 / Assignment-1 / logs / j-RTZLFIQLBFI0

Name	Storage Class	Size	Last Modified
daemons	--	--	--
jobs	--	--	--
node	--	--	--
steps	--	--	--
task-attempts	--	--	--

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Please find attached with the submission the controller & syslog logs and the final output files part-r-00000, part-r-00001 & part-r-00002.