CS6240, HW 1, Rushil Patel

Exploration Challenge

In Hadoop Distributed File System there is no provision for performing update operations on the File once written.
MapReduce works on the principle of write once and read many times.

The following is the evidence that HDFS does not support for updates to a file once written
Source : Hadoop the Definitive Guide Book

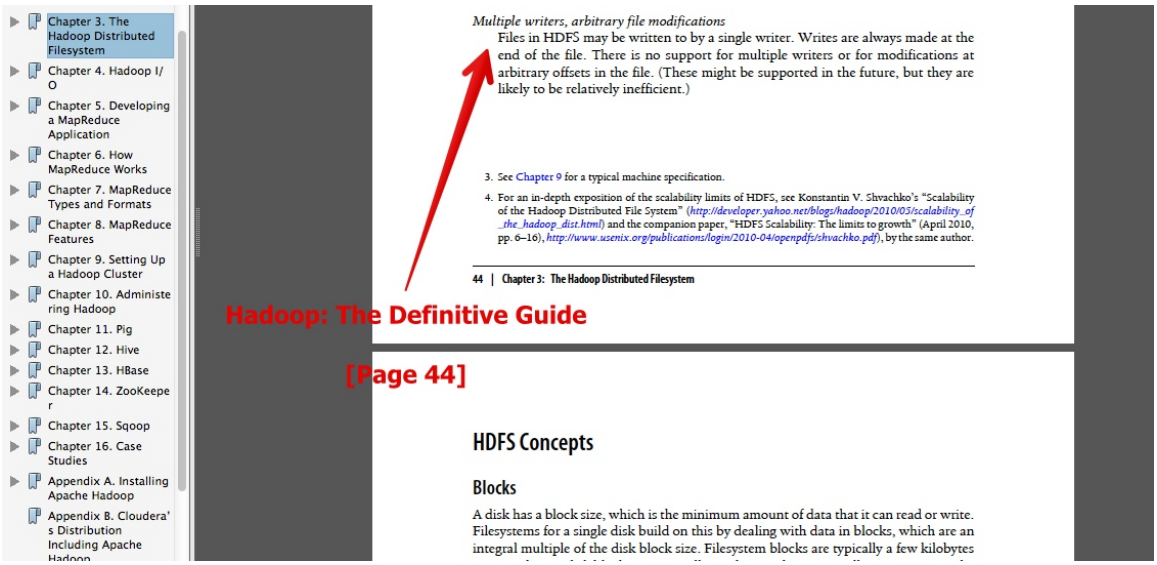
is less efficient than MapReduce, which uses Sort/Merge to rebuild the database.

In many ways, MapReduce can be seen as a complement to a Rational Database Management System (RDBMS). (The differences between the two systems are shown in Table 1-1.) MapReduce is a good fit for problems that need to analyze the whole dataset in a batch fashion, particularly for ad hoc analysis. An RDBMS is good for point queries or updates, where the dataset has been indexed to deliver low-latency retrieval and update times of a relatively small amount of data. MapReduce suits applications where the data is written once and read many times, whereas a relational database is good for datasets that are continually updated.

Table 1-1. RDBMS compared to MapReduce

	Traditional RDBMS	MapReduce
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear

Another difference between MapReduce and an RDBMS is the amount of structure in the datasets on which they operate. *Structured data* is data that is organized into entities that have a defined format, such as XML documents or database tables that conform to a particular predefined schema. This is the realm of the RDBMS. *Semi-structured data*, on the other hand, is looser, and though there may be a schema, it is often ignored, so it may be used only as a guide to the structure of the data: for example, a spreadsheet, in which the structure is the grid of cells, although the cells themselves may hold any form of data. *Unstructured data* does not have any particular internal structure: for example, plain text or image data. MapReduce works well on unstructured or semi-structured data because it is designed to interpret the data at processing time. In other words, the input keys and values for MapReduce are not intrinsic properties of the data, but they are chosen by the person analyzing the data.



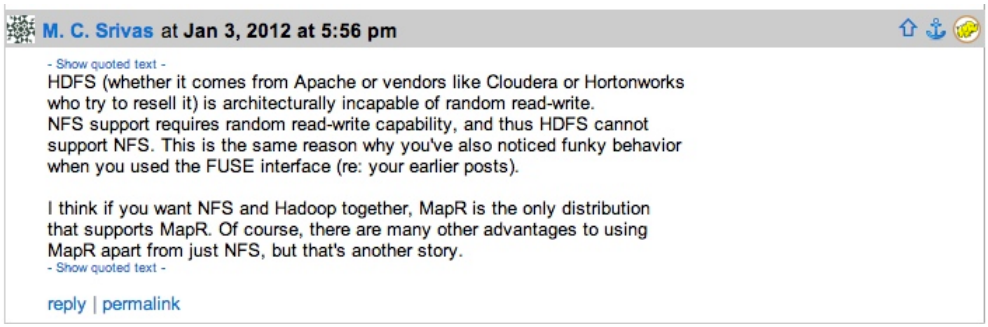
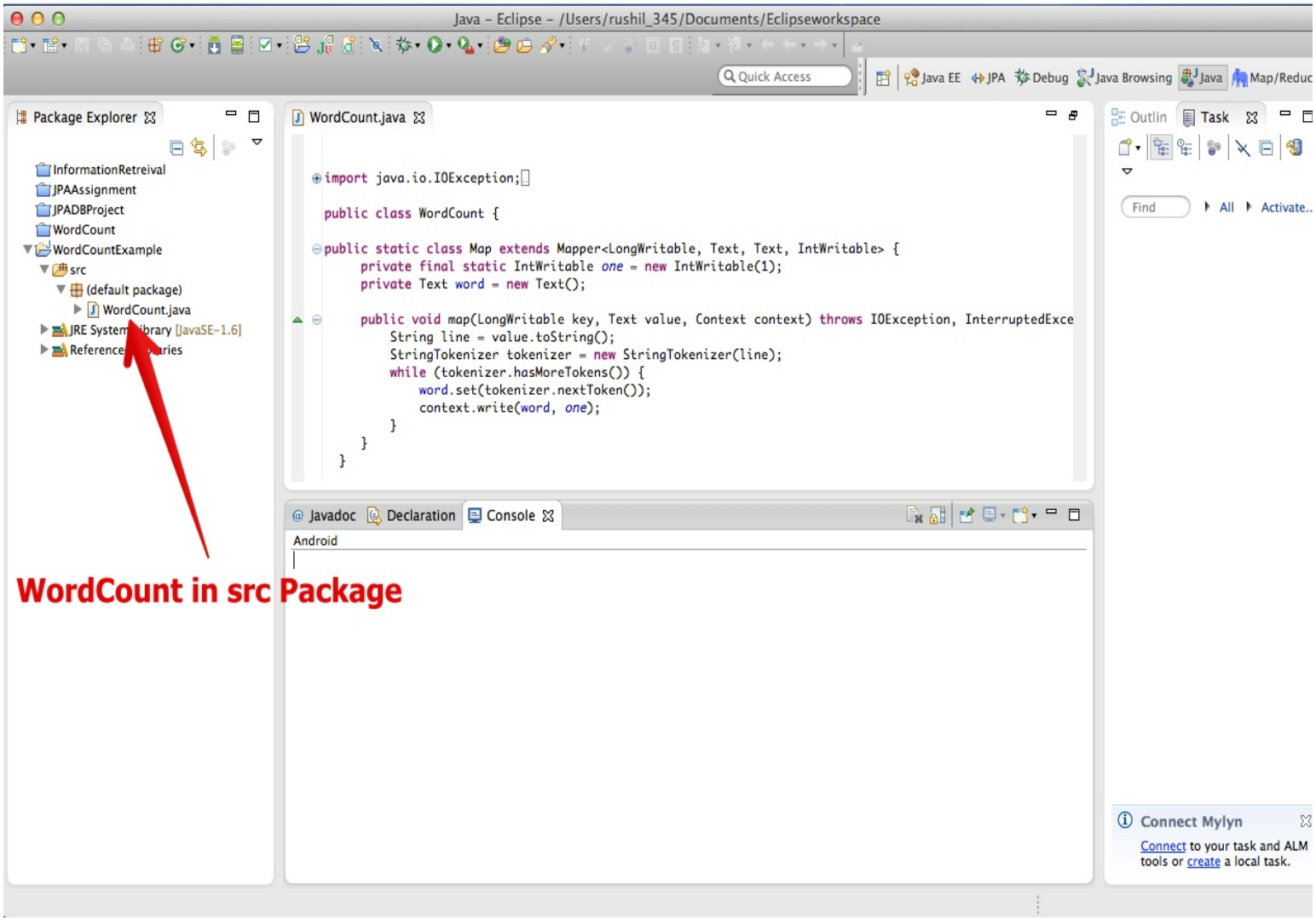
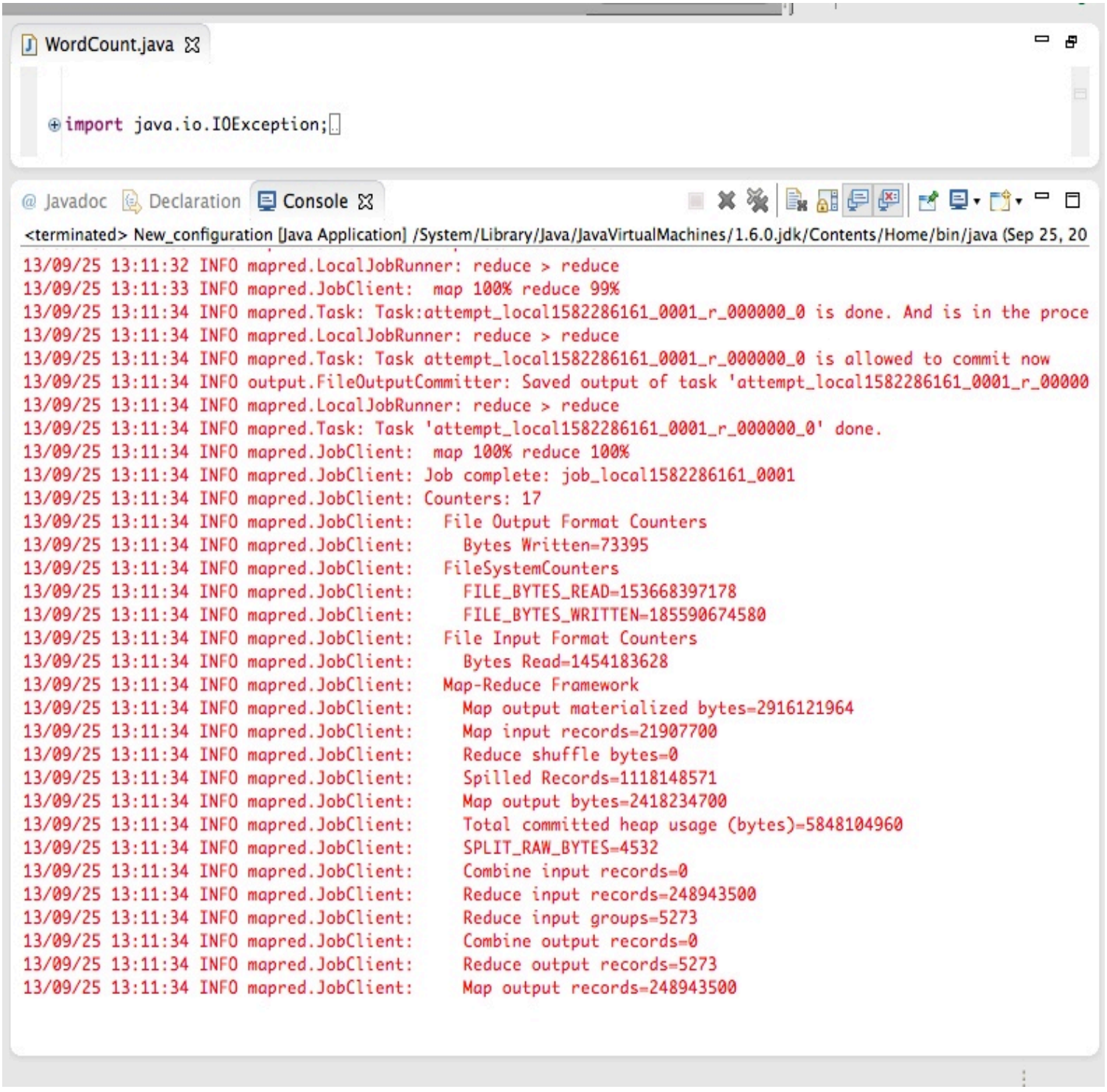


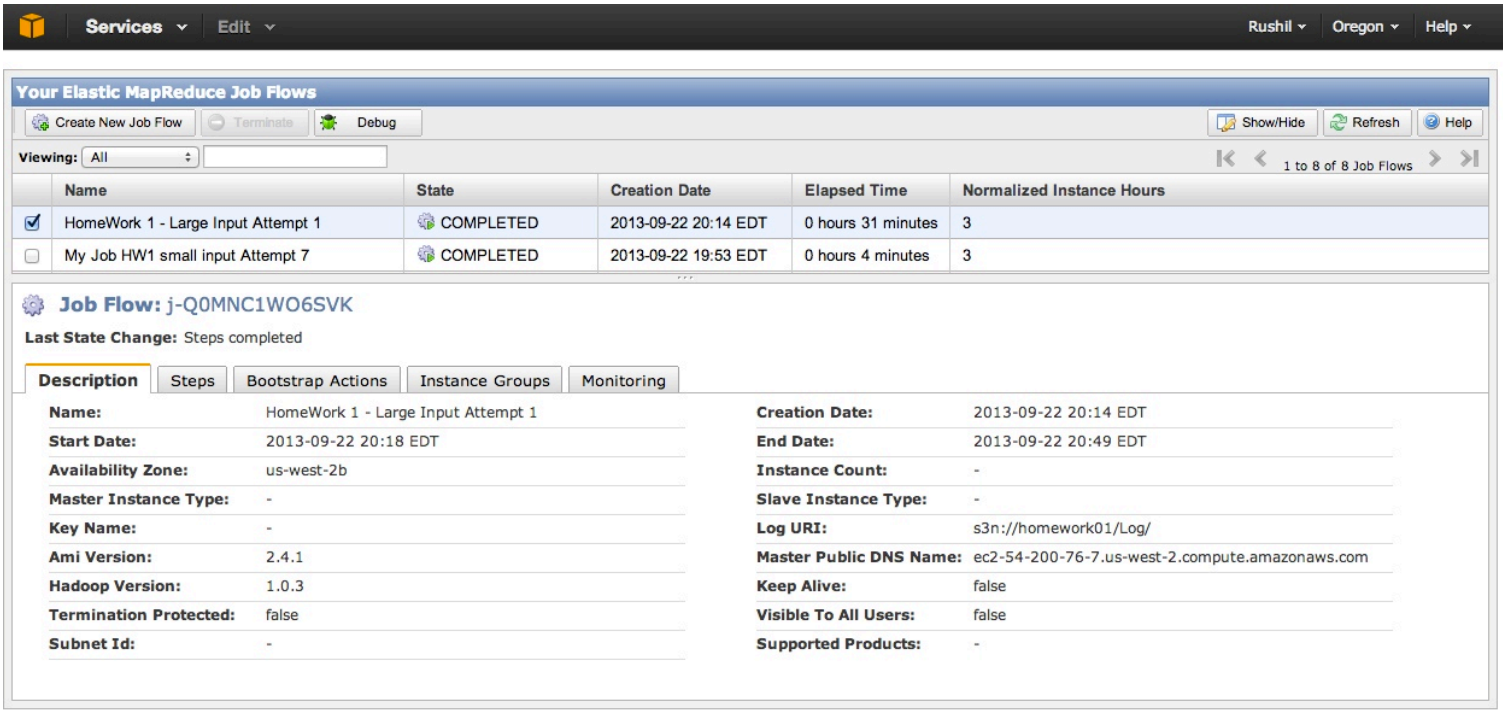
Figure 1http://grokbase.com/t/hadoop/hdfs-user/1213gt4fvz/file-edit-in-hadoop

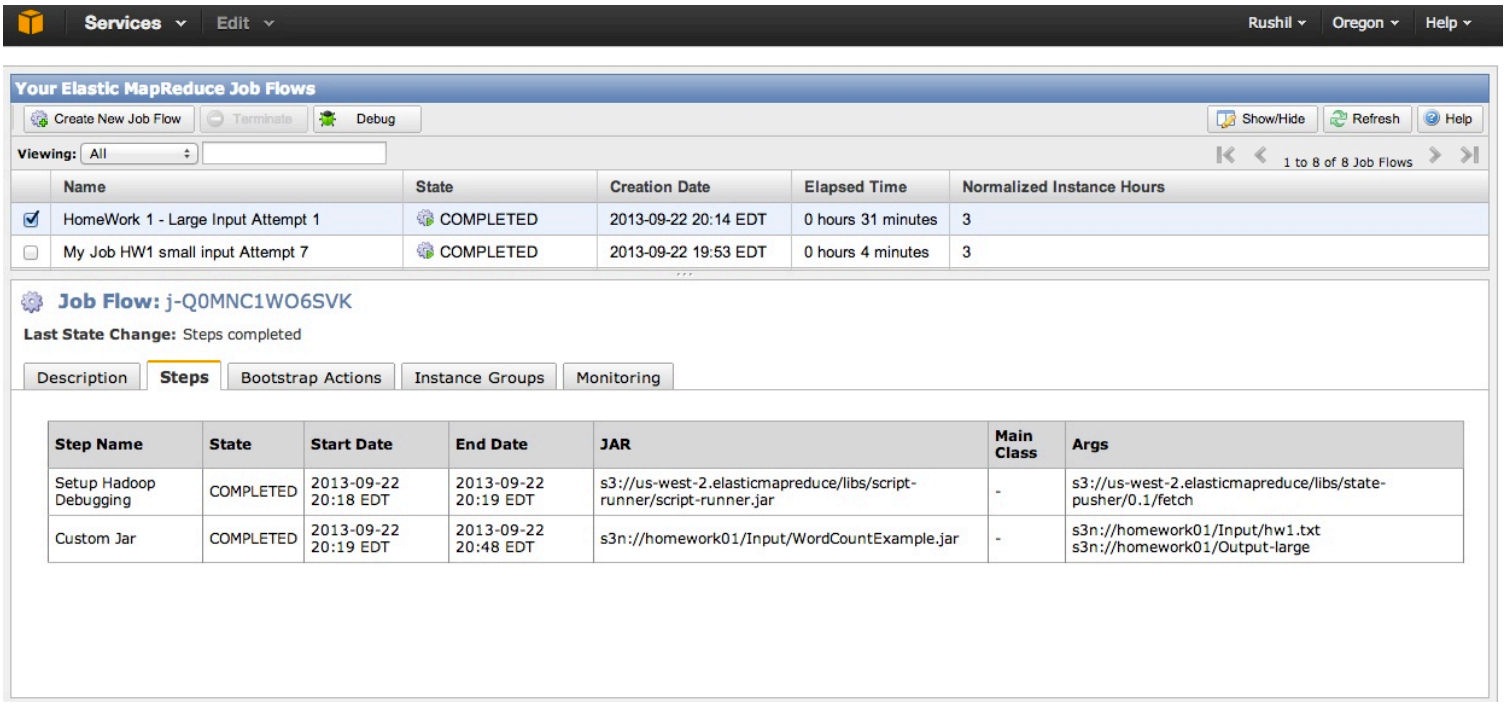
Local Execution

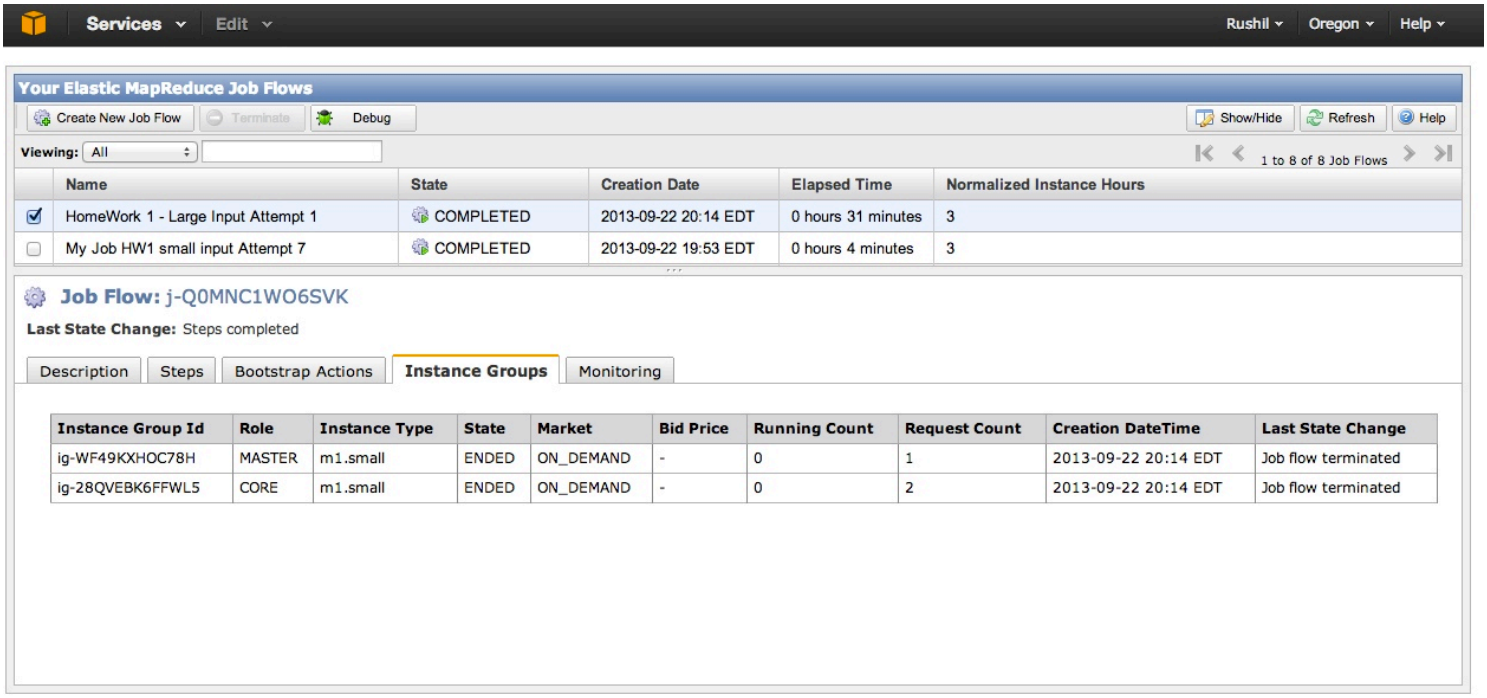


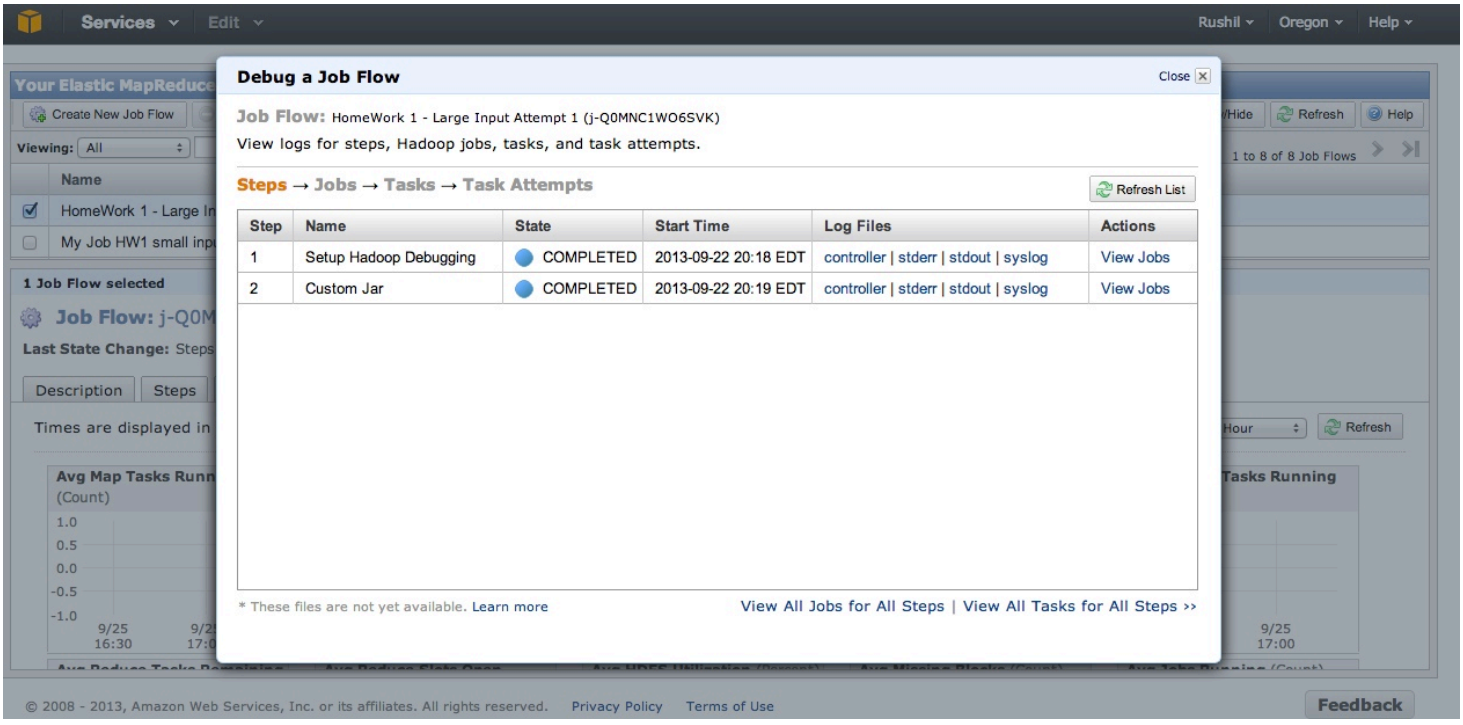

WordCount.java
import java.io.IOException;
Javadoc
Declaration
Console
<terminated> New_configuration [Java Application] /System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Home/bin/java (Sep 25, 20
13/09/25 13:11:32 INFO mapred.LocalJobRunner: reduce > reduce
13/09/25 13:11:33 INFO mapred.JobClient: map 100% reduce 99%
13/09/25 13:11:34 INFO mapred.Task: Task:attempt_local1582286161_0001_r_000000_0 is done. And is in the proce
13/09/25 13:11:34 INFO mapred.LocalJobRunner: reduce > reduce
13/09/25 13:11:34 INFO mapred.Task: Task attempt_local1582286161_0001_r_000000_0 is allowed to commit now
13/09/25 13:11:34 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1582286161_0001_r_00000
13/09/25 13:11:34 INFO mapred.LocalJobRunner: reduce > reduce
13/09/25 13:11:34 INFO mapred.Task: Task 'attempt_local1582286161_0001_r_000000_0' done.
13/09/25 13:11:34 INFO mapred.JobClient: map 100% reduce 100%
13/09/25 13:11:34 INFO mapred.JobClient: Job complete: job_local1582286161_0001
13/09/25 13:11:34 INFO mapred.JobClient: Counters: 17
13/09/25 13:11:34 INFO mapred.JobClient: File Output Format Counters
13/09/25 13:11:34 INFO mapred.JobClient: Bytes Written=73395
13/09/25 13:11:34 INFO mapred.JobClient: FileSystemCounters
13/09/25 13:11:34 INFO mapred.JobClient: FILE_BYTES_READ=153668397178
13/09/25 13:11:34 INFO mapred.JobClient: FILE_BYTES_WRITTEN=185590674580
13/09/25 13:11:34 INFO mapred.JobClient: File Input Format Counters
13/09/25 13:11:34 INFO mapred.JobClient: Bytes Read=1454183628
13/09/25 13:11:34 INFO mapred.JobClient: Map-Reduce Framework
13/09/25 13:11:34 INFO mapred.JobClient: Map output materialized bytes=2916121964
13/09/25 13:11:34 INFO mapred.JobClient: Map input records=21907700
13/09/25 13:11:34 INFO mapred.JobClient: Reduce shuffle bytes=0
13/09/25 13:11:34 INFO mapred.JobClient: Spilled Records=1118148571
13/09/25 13:11:34 INFO mapred.JobClient: Map output bytes=2418234700
13/09/25 13:11:34 INFO mapred.JobClient: Total committed heap usage (bytes)=5848104960
13/09/25 13:11:34 INFO mapred.JobClient: SPLIT_RAW_BYTES=4532
13/09/25 13:11:34 INFO mapred.JobClient: Combine input records=0
13/09/25 13:11:34 INFO mapred.JobClient: Reduce input records=248943500
13/09/25 13:11:34 INFO mapred.JobClient: Reduce input groups=5273
13/09/25 13:11:34 INFO mapred.JobClient: Combine output records=0
13/09/25 13:11:34 INFO mapred.JobClient: Reduce output records=5273
13/09/25 13:11:34 INFO mapred.JobClient: Map output records=248943500

AWS Execution






Services
Edit
Rushil
Oregon
Help
Your Elastic MapReduce Job Flows
Create New Job Flow
Terminate
Debug
Show/Hide
Refresh
Help
Viewing: All
1 to 8 of 8 Job Flows
Name
State
Creation Date
Elapsed Time
Normalized Instance Hours
HomeWork 1 - Large Input Attempt 1
COMPLETED
2013-09-22 20:14 EDT
0 hours 31 minutes
3
My Job HW1 small input Attempt 7
COMPLETED
2013-09-22 19:53 EDT
0 hours 4 minutes
3
Job Flow: j-Q0MNC1WO6SVK
Last State Change: Steps completed
Description
Steps
Bootstrap Actions
Instance Groups
Monitoring
Instance Group Id
Role
Instance Type
State
Market
Bid Price
Running Count
Request Count
Creation DateTime
Last State Change
ig-WF49KXHOC78H
MASTER
m1.small
ENDED
ON_DEMAND
-
0
1
2013-09-22 20:14 EDT
Job flow terminated
ig-28QVEBK6FFWL5
CORE
m1.small
ENDED
ON_DEMAND
-
0
2
2013-09-22 20:14 EDT
Job flow terminated


Services
Edit
Rushil
Oregon
Help
Debug a Job Flow
Close
Job Flow: HomeWork 1 - Large Input Attempt 1 (j-Q0MNC1WO6SVK)
View logs for steps, Hadoop jobs, tasks, and task attempts.
Steps → Jobs → Tasks → Task Attempts
Refresh List
Step
Name
State
Start Time
Log Files
Actions
1
Setup Hadoop Debugging
COMPLETED
2013-09-22 20:18 EDT
controller | stderr | stdout | syslog
View Jobs
2
Custom Jar
COMPLETED
2013-09-22 20:19 EDT
controller | stderr | stdout | syslog
View Jobs
* These files are not yet available. Learn more
View All Jobs for All Steps | View All Tasks for All Steps ››
© 2008 - 2013, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Feedback

```
2013-09-23T00:19:03.632Z INFO Fetching jar file.
2013-09-23T00:19:20.890Z INFO Working dir /mnt/var/lib/hadoop/steps/2
2013-09-23T00:19:20.890Z INFO Executing /usr/java/latest/bin/java -cp /home/hadoop/conf:/usr/java/latest/lib/tools.jar:/home/hadoop:/home/hadoop/hadoop-core-
1.0.3.jar:/home/hadoop/hadoop-core.jar:/home/hadoop/hadoop-tools-1.0.3.jar:/home/hadoop/hadoop-tools.jar:/home/hadoop/lib/*:/home/hadoop/lib/jetty-ext/* -Xmx1000m -
Dhadoop.log.dir=/mnt/var/log/hadoop/steps/2 -Dhadoop.log.file=syslog -Dhadoop.home.dir=/home/hadoop -Dhadoop.id.str=hadoop -Dhadoop.root.logger=INFO,DRFA -
Djava.io.tmpdir=/mnt/var/lib/hadoop/steps/2/tmp -Djava.library.path=/home/hadoop/native/Linux-i386-32 org.apache.hadoop.util.RunJar /mnt/var/lib/hadoop/steps/2/WordCountExample.jar
s3n://homework01/Input/hw1.txt s3n://homework01/Output-large
2013-09-23T00:48:21.337Z INFO Execution ended with ret val 0
2013-09-23T00:48:28.914Z INFO Step created jobs: job_201309230016_0001
2013-09-23T00:48:39.090Z INFO Step succeeded
```

```
2013-09-23 00:47:11,008 INFO org.apache.hadoop.mapred.JobClient (main):  map 100% reduce 94%
2013-09-23 00:47:26,029 INFO org.apache.hadoop.mapred.JobClient (main):  map 100% reduce 95%
2013-09-23 00:47:35,042 INFO org.apache.hadoop.mapred.JobClient (main):  map 100% reduce 96%
2013-09-23 00:47:44,055 INFO org.apache.hadoop.mapred.JobClient (main):  map 100% reduce 97%
2013-09-23 00:47:53,068 INFO org.apache.hadoop.mapred.JobClient (main):  map 100% reduce 98%
2013-09-23 00:48:02,081 INFO org.apache.hadoop.mapred.JobClient (main):  map 100% reduce 99%
2013-09-23 00:48:11,094 INFO org.apache.hadoop.mapred.JobClient (main):  map 100% reduce 100%
2013-09-23 00:48:19,110 INFO org.apache.hadoop.mapred.JobClient (main): Job complete: job_201309230016_0001
2013-09-23 00:48:19,126 INFO org.apache.hadoop.mapred.JobClient (main): Counters: 30
2013-09-23 00:48:19,126 INFO org.apache.hadoop.mapred.JobClient (main):   Job Counters
2013-09-23 00:48:19,126 INFO org.apache.hadoop.mapred.JobClient (main):     Launched reduce tasks=4
2013-09-23 00:48:19,126 INFO org.apache.hadoop.mapred.JobClient (main):     SLOTS_MILLIS_MAPS=4958150
2013-09-23 00:48:19,126 INFO org.apache.hadoop.mapred.JobClient (main):     Total time spent by all reduces waiting after reserving slots (ms)=0
2013-09-23 00:48:19,126 INFO org.apache.hadoop.mapred.JobClient (main):     Total time spent by all maps waiting after reserving slots (ms)=0
2013-09-23 00:48:19,126 INFO org.apache.hadoop.mapred.JobClient (main):     Rack-local map tasks=23
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):     Launched map tasks=23
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):     SLOTS_MILLIS_REDUCES=2725453
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):   File Output Format Counters
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):     Bytes Written=72815
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):   FileSystemCounters
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):     S3N_BYTES_READ=1454089428
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):     FILE_BYTES_READ=391376467
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):     HDFS_BYTES_READ=2090
2013-09-23 00:48:19,127 INFO org.apache.hadoop.mapred.JobClient (main):     S3N_BYTES_WRITTEN=72815
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):     FILE_BYTES_WRITTEN=470175762
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):   File Input Format Counters
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):     Bytes Read=1454089428
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):   Map-Reduce Framework
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):     Map output materialized bytes=139613088
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):     Map input records=21907700
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):     Reduce shuffle bytes=139613088
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):     Spilled Records=746830500
2013-09-23 00:48:19,128 INFO org.apache.hadoop.mapred.JobClient (main):     Map output bytes=2418234700
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Total committed heap usage (bytes)=5268267008
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     CPU time spent (ms)=2702550
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Combine input records=0
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     SPLIT_RAW_BYTES=2090
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Reduce input records=248943500
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Reduce input groups=5273
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Combine output records=0
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Physical memory (bytes) snapshot=6348492800
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Reduce output records=5273
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Virtual memory (bytes) snapshot=11790819328
2013-09-23 00:48:19,129 INFO org.apache.hadoop.mapred.JobClient (main):     Map output records=248943500
```

Note:
All the above files are available in the Images folder of this supplied Zip file.

Credits: Hadoop Library for WordCount Example & AWS.