HW1
Name: Chandrima Ghosh
MapReduce-CS6240-Sec01

Report:

Weather Data Results:
For each of the versions of your sequential and multithreaded program detailed in B and C, report the minimum, average, and maximum running time observed over the 10 runs. (5 points)

Sequential: Without Fib(B)

min execution time: 2860
max execution time: 3657
avg execution time: 3194

Sequential: With Fib(C)

min execution time: 2988
max execution time: 7644
avg execution time: 4473
.....
No Lock: Without Fib(B)

min execution time: 1514
max execution time: 2157
avg execution time: 1634

No Lock: With Fib(C)

min execution time: 1565
max execution time: 2537
avg execution time: 1776
.....

Coarse Lock: Without Fib(B)

min execution time: 1774
max execution time: 3202
avg execution time: 2004

Coarse Lock: With Fib(C)

min execution time: 1978
max execution time: 8977
avg execution time: 2913
.....

Fine Lock: Without Fib(B)
min execution time: 1514
max execution time: 4819
avg execution time: 1635

Fine Lock: With Fib(C)
min execution time: 1566
max execution time: 3023
avg execution time: 1984

.....
No Sharing: Without Fib(B)
min execution time: 2920
max execution time: 3523
avg execution time: 3104

No Sharing: With Fib(C)

min execution time: 2874
max execution time: 7851
avg execution time: 3467
.....

Report the number of worker threads used and the speedup of the multithreaded versions based on the corresponding average running times. (5 points)
The number of worker threads used is :4
Speed Up:

No Lock: Without Fib(B)
1.9
No Lock: With Fib(C)
2.5

Coarse Lock: Without Fib(B)
1.5

Coarse Lock: With Fib(C)
1.5

Fine Lock: Without Fib(B)
1.9

Fine Lock: With Fib(C)
2.254

No Sharing: Without Fib(B)
1.02

No Sharing: With Fib(C)
1.28

Answer the following questions in a brief and concise manner: (4 points each)

1. Which program version (SEQ, NO-LOCK, COARSE-LOCK, FINE-LOCK, NO-SHARING) would you normally expect to finish fastest and why? Do the experiments confirm your expectation? If not, try to explain the reasons.

I would expect no-lock to work the fastest, maximum parallelism can be achieved as none of the worker threads work wait for other threads. Although this is the fastest this can return erroneous values. My experiment confirms to this.

2.Which program version (SEQ, NO-LOCK, COARSE-LOCK, FINE-LOCK, NO-SHARING) would you normally expect to finish slowest and why? Do the experiments confirm your expectation? If not, try to explain the reasons.

The Sequential version is the slowest as there is only the main thread doing all the work. There is no parallelism. My experiments confirm to this .

3.Compare the temperature averages returned by each program version. Report if any of them is incorrect or if any of the programs crashed because of concurrent accesses

All the program results are consistent except the No Lock version.
The program throws a Null Pointer Exception or Array Index Out Of Bounds and crashes once in a while. Even if it executes the values are mostly inconsistent.
Since no locks are involved one thread may overwrite another threads computation or trying to access a value that has not been written yet which leads to Null Pointer Exception.

4.Compare the running times of SEQ and COARSE-LOCK. Try to explain why one is slower than the other. (Make sure to consider the results of both B and C—this might support or refute a possible hypothesis.)

Version B
Sequential is slower than Coarse lock version.
In sequential the processing of the files is sequential while in coarse lock the processing is done by multiple processes speeding up the execution .

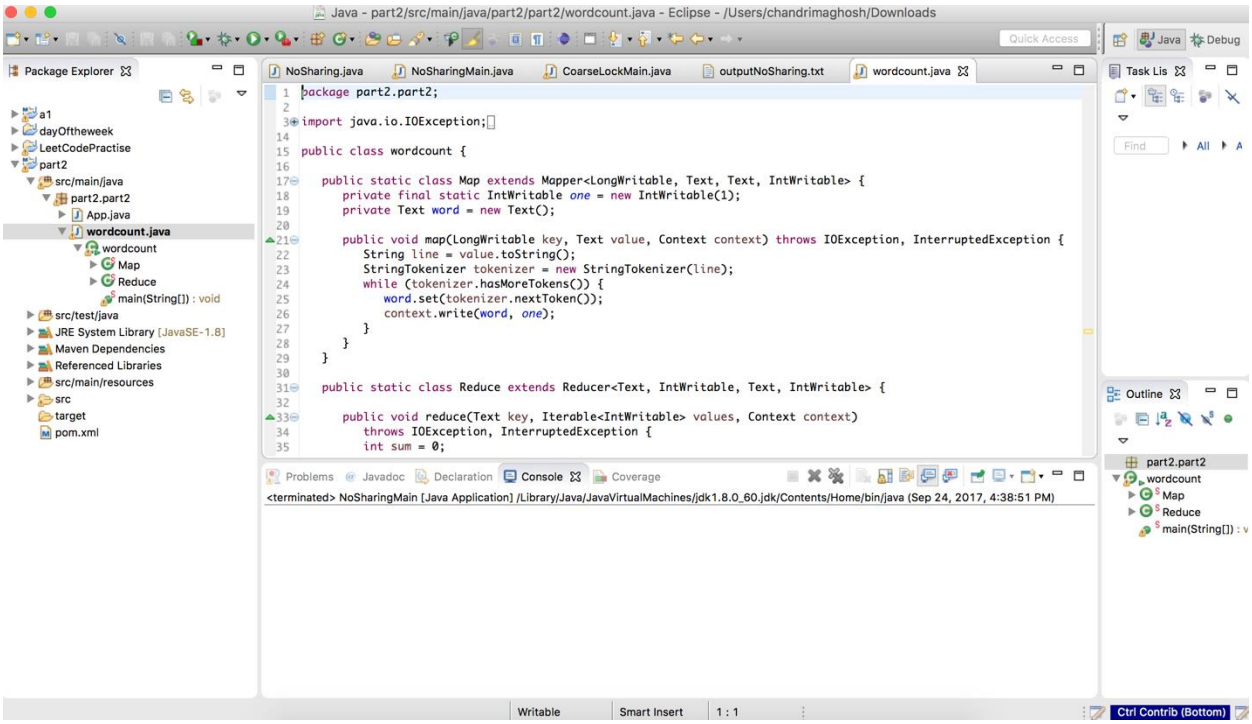
Version C
In Sequential and Coarse lock version in version C , we see that there is no significant Difference.
Since the lock is applied to the entire data structure , the other threads need to wait on the locks and for longer wait a bottleneck is created which is almost same as sequential execution and that's maybe the reason why there is no significant improvement .

5.How does the higher computation cost in part C (additional Fibonacci computation) affect the difference between COARSE-LOCK and FINE-LOCK? Try to explain the reason.

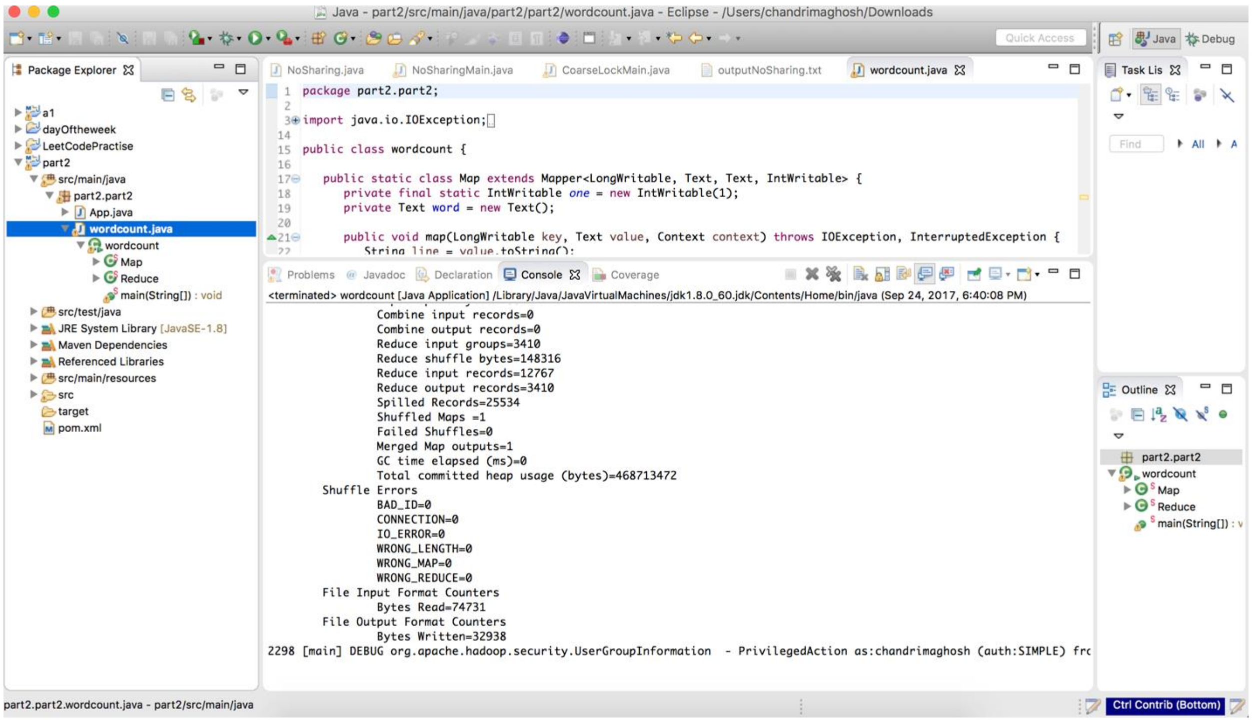
In Coarse lock version of the program the lock is applied on the entire data structure and the fib() call is inside that thus it cannot run parallel.
In fine lock there is a chance that fib() can run in parallel as the lock is only on the value object .Thus,
Fine-Lock outperforms Coarse lock program.

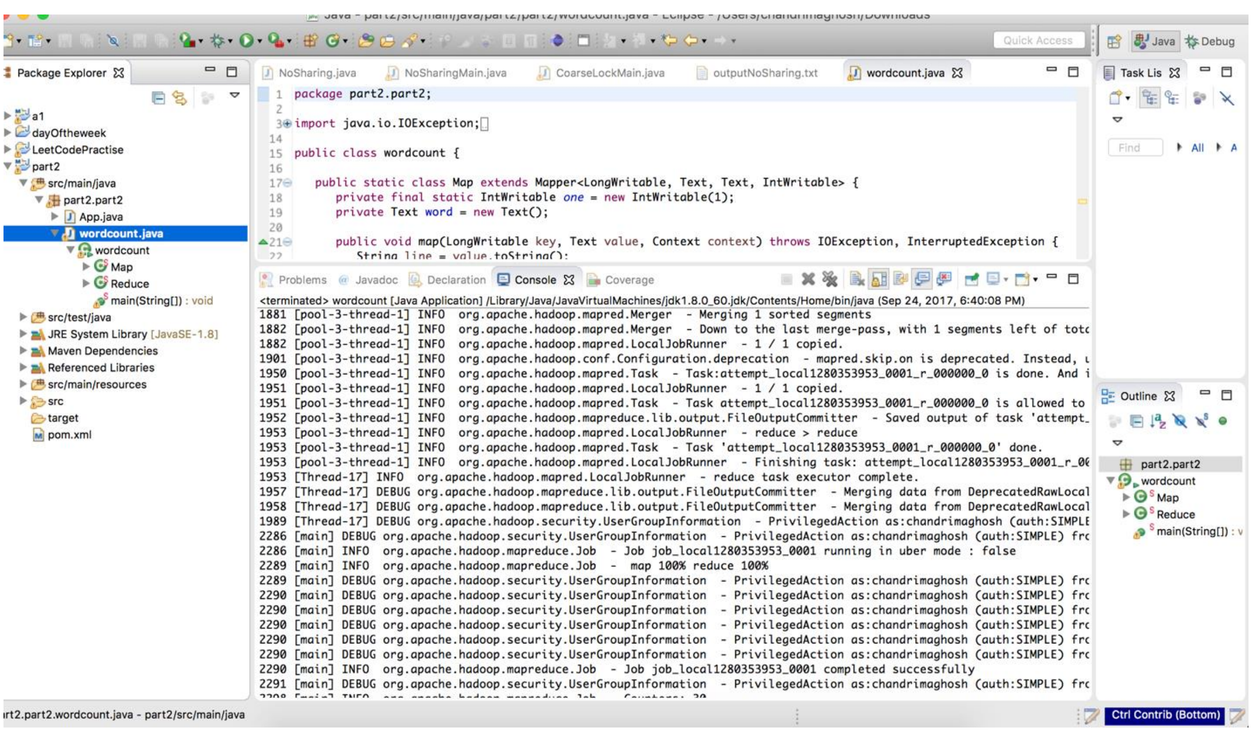
Word Count Local Execution

- *Project directory structure, showing that the WordCount.java file is somewhere in the src directory. (10 points)*



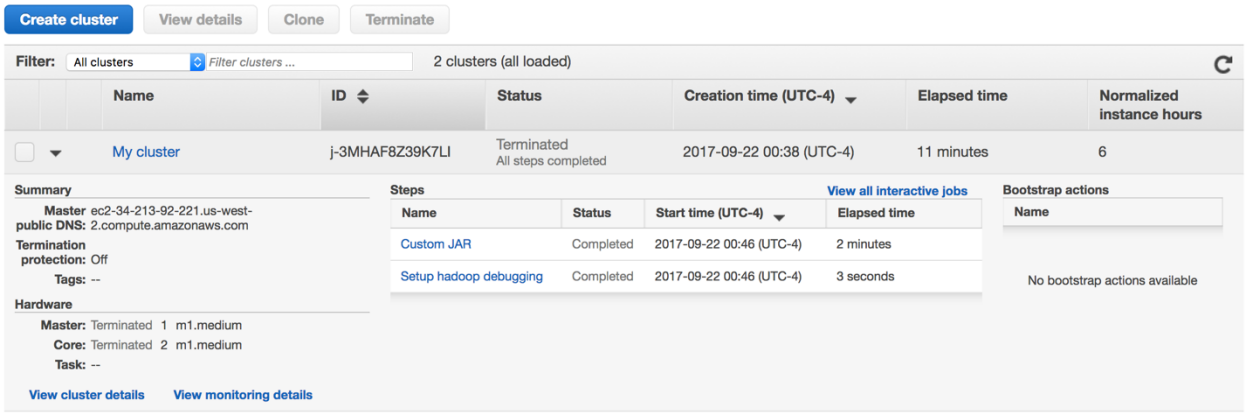
The console output for a successful run of the WordCount program inside the IDE. The console output refers to the job summary information Hadoop produces, not the output your job emits. Show at least the last 20 lines of the console output. (10 points)

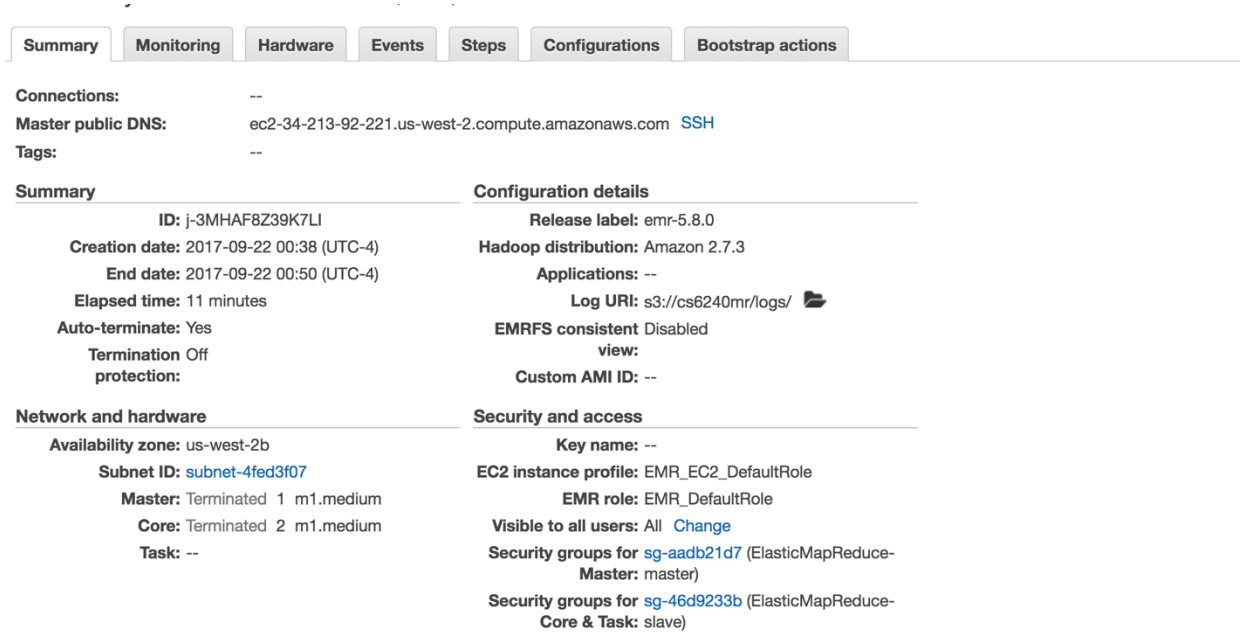

Java - part2/src/main/java/part2/part2/wordcount.java - Eclipse - /Users/chandrimaghosh/Downloads
Package Explorer
a1
dayOftheweek
LeetCodePractise
part2
src/main/java
part2.part2
App.java
wordcount.java
wordcount
Map
Reduce
main(String[]) : void
src/test/java
JRE System Library [JavaSE-1.8]
Maven Dependencies
Referenced Libraries
src/main/resources
src
target
pom.xml
NoSharing.java
NoSharingMain.java
CoarseLockMain.java
outputNoSharing.txt
wordcount.java
package part2.part2;
import java.io.IOException;
public class wordcount {
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
private final static IntWritable one = new IntWritable(1);
private Text word = new Text();
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
String line = value.toString();
Problems
Javadoc
Declaration
Console
Coverage
<terminated> wordcount [Java Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_60.jdk/Contents/Home/bin/java (Sep 24, 2017, 6:40:08 PM)
Combine input records=0
Combine output records=0
Reduce input groups=3410
Reduce shuffle bytes=148316
Reduce input records=12767
Reduce output records=3410
Spilled Records=25534
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=468713472
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=74731
File Output Format Counters
Bytes Written=32938
2298 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
Task Lis
Outline
part2.part2
wordcount
Map
Reduce
main(String[]) : v
part2.part2.wordcount.java - part2/src/main/java
Ctrl Contrib (Bottom)

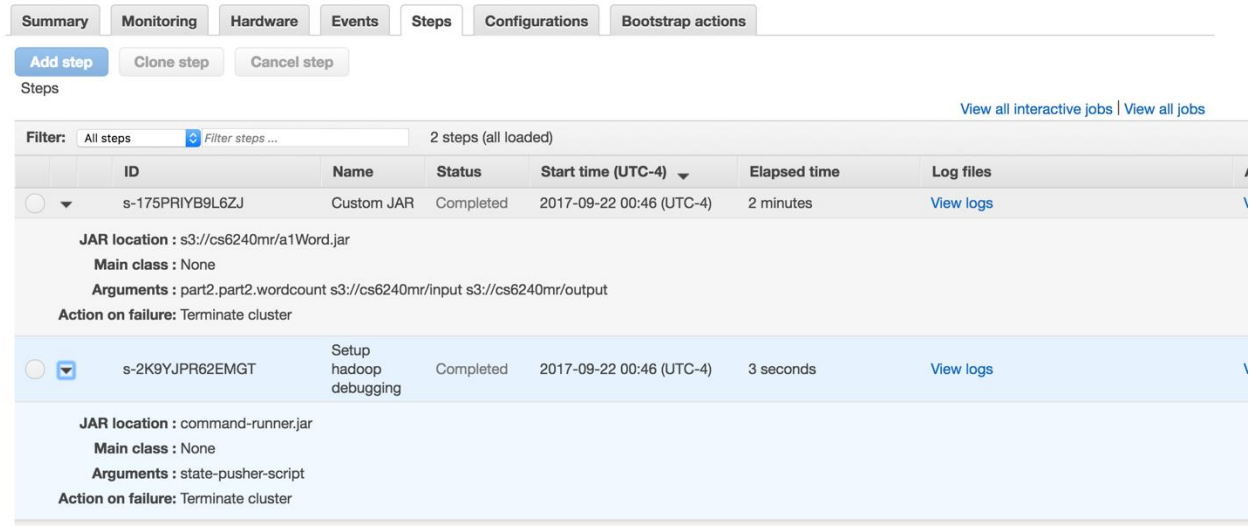

Package Explorer
a1
dayOftheweek
LeetCodePractise
part2
src/main/java
part2.part2
App.java
wordcount.java
wordcount
Map
Reduce
main(String[]) : void
src/test/java
JRE System Library [JavaSE-1.8]
Maven Dependencies
Referenced Libraries
src/main/resources
src
target
pom.xml
NoSharing.java
NoSharingMain.java
CoarseLockMain.java
outputNoSharing.txt
wordcount.java
package part2.part2;
import java.io.IOException;
public class wordcount {
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
private final static IntWritable one = new IntWritable(1);
private Text word = new Text();
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
Problems
Javadoc
Declaration
Console
Coverage
<terminated> wordcount [Java Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_60.jdk/Contents/Home/bin/java (Sep 24, 2017, 6:40:08 PM)
1881 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Merger - Merging 1 sorted segments
1882 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Merger - Down to the last merge-pass, with 1 segments left of tota
1882 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - 1 / 1 copied.
1901 [pool-3-thread-1] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.skip.on is deprecated. Instead, u
1950 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Task - Task:attempt_local1280353953_0001_r_000000_0 is done. And i
1951 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - 1 / 1 copied.
1951 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Task - Task attempt_local1280353953_0001_r_000000_0 is allowed to
1952 [pool-3-thread-1] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_
1953 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce > reduce
1953 [pool-3-thread-1] INFO org.apache.hadoop.mapred.Task - Task 'attempt_local1280353953_0001_r_000000_0' done.
1953 [pool-3-thread-1] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local1280353953_0001_r_00
1953 [Thread-17] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce task executor complete.
1957 [Thread-17] DEBUG org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Merging data from DeprecatedRawLocal
1958 [Thread-17] DEBUG org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Merging data from DeprecatedRawLocal
1989 [Thread-17] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE
2286 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
2286 [main] INFO org.apache.hadoop.mapreduce.Job - Job job_local1280353953_0001 running in uber mode : false
2289 [main] INFO org.apache.hadoop.mapreduce.Job - map 100% reduce 100%
2289 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
2290 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
2290 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
2290 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
2290 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
2290 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
2290 [main] INFO org.apache.hadoop.mapreduce.Job - Job job_local1280353953_0001 completed successfully
2291 [main] DEBUG org.apache.hadoop.security.UserGroupInformation - PrivilegedAction as:chandrimaghosh (auth:SIMPLE) frc
Task Lis
Outline
part2.part2
wordcount
Map
Reduce
main(String[]) : v
rt2.part2.wordcount.java - part2/src/main/java
Ctrl Contrib (Bottom)

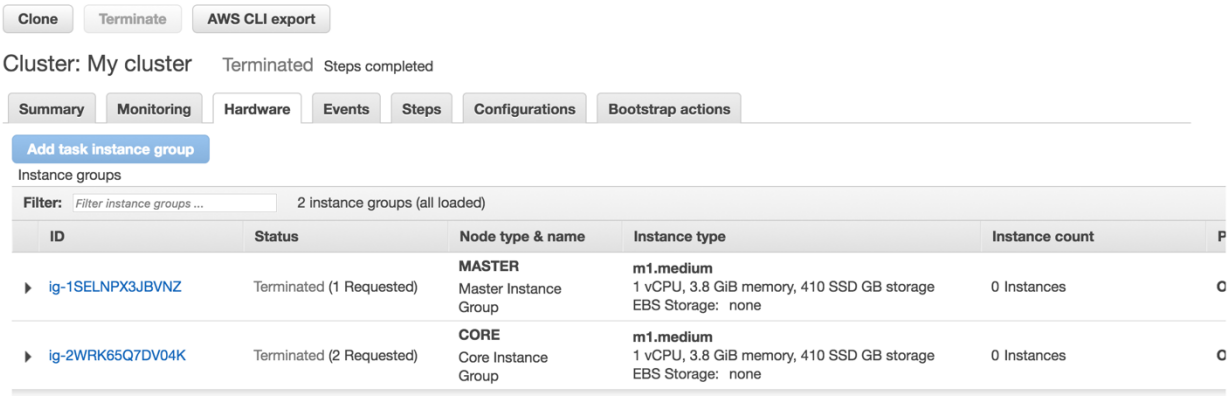
Word Count AWS Execution

Show a similar screenshot that provides convincing evidence of a successful run of the Word Count program on AWS. Make sure you run the program using at least three machines, i.e., one master node and two workers. (10 points) Once the execution is completed, look for the corresponding log files, in particular controller and syslog, and save them.

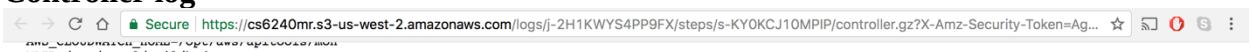





Summary
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Add step
Clone step
Cancel step
Steps
View all interactive jobs | View all jobs
Filter: All steps Filter steps ...
2 steps (all loaded)
ID
Name
Status
Start time (UTC-4)
Elapsed time
Log files
s-175PRIYB9L6ZJ
Custom JAR
Completed
2017-09-22 00:46 (UTC-4)
2 minutes
View logs
JAR location : s3://cs6240mr/a1Word.jar
Main class : None
Arguments : part2.part2.wordcount s3://cs6240mr/input s3://cs6240mr/output
Action on failure: Terminate cluster
s-2K9YJPR62EMGT
Setup hadoop debugging
Completed
2017-09-22 00:46 (UTC-4)
3 seconds
View logs
JAR location : command-runner.jar
Main class : None
Arguments : state-pusher-script
Action on failure: Terminate cluster


Clone
Terminate
AWS CLI export
Cluster: My cluster
Terminated
Steps completed
Summary
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Add task instance group
Instance groups
Filter: Filter instance groups ...
2 instance groups (all loaded)
ID
Status
Node type & name
Instance type
Instance count
ig-1SELNPX3JBVNZ
Terminated (1 Requested)
MASTER
Master Instance Group
m1.medium
1 vCPU, 3.8 GiB memory, 410 SSD GB storage
EBS Storage: none
0 Instances
ig-2WRK65Q7DV04K
Terminated (2 Requested)
CORE
Core Instance Group
m1.medium
1 vCPU, 3.8 GiB memory, 410 SSD GB storage
EBS Storage: none
0 Instances

Controller log




```
  MAIL=/var/spool/mail/hadoop
  LESS_TERMCAP_ue=[0m
  LOGNAME=hadoop
  PWD=/
  LANGSH_SOURCED=1
  HADOOP_CLIENT_OPTS=-Djava.io.tmpdir=/mnt/var/lib/hadoop/steps/s-KY0KCJ10MPIP/tmp
  _=/etc/alternatives/jre/bin/java
  CONSOLETYPE=serial
  RUNLEVEL=3
  LESSOPEN=||/usr/bin/lesspipe.sh %s
  previous=N
  UPSTART_EVENTS=runlevel
  AWS_PATH=/opt/aws
  USER=hadoop
  UPSTART_INSTANCE=
  PREVLEVEL=N
  HADOOP_LOGFILE=syslog
  PYTHON_INSTALL_LAYOUT=amzn
  HOSTNAME=ip-172-31-29-50
  NLSPATH=/usr/dt/lib/nls/msg/%L/%N.cat
  HADOOP_LOG_DIR=/mnt/var/log/hadoop/steps/s-KY0KCJ10MPIP
  EC2_AMITOOL_HOME=/opt/aws/amitools/ec2
  SHLVL=5
  HOME=/home/hadoop
  HADOOP_IDENT_STRING=hadoop
INFO redirectOutput to /mnt/var/log/hadoop/steps/s-KY0KCJ10MPIP/stdout
INFO redirectError to /mnt/var/log/hadoop/steps/s-KY0KCJ10MPIP/stderr
INFO Working dir /mnt/var/lib/hadoop/steps/s-KY0KCJ10MPIP
INFO ProcessRunner started child process 8400 :
hadoop    8400  4058  2 00:01 ?        00:00:00 /etc/alternatives/jre/bin/java -Xmx1000m -server -XX:OnOutOfMemoryError=kill -9 %p -
Dhadoop.log.dir=/mnt/var/log/hadoop/steps/s-KY0KCJ10MPIP -Dhadoop.log.file=syslog -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str=hadoop -
Dhadoop.root.logger=INFO,DRFA -Djava.library.path=:/usr/lib/hadoop-lzo/lib/native:/usr/lib/hadoop/lib/native -Dhadoop.policy.file=hadoop-policy.xml -
Djava.net.preferIPv4Stack=true -Djava.io.tmpdir=/mnt/var/lib/hadoop/steps/s-KY0KCJ10MPIP/tmp -Dhadoop.security.logger=INFO,NullAppender -Dsun.net.inetaddr.ttl=30
org.apache.hadoop.util.RunJar /mnt/var/lib/hadoop/steps/s-KY0KCJ10MPIP/a1Word.jar part2.part2.wordcount s3://cs6240mr/input s3://cs6240mr/output
2017-09-25T00:02:01.057Z INFO HadoopJarStepRunner.Runner: startRun() called for s-KY0KCJ10MPIP Child Pid: 8400
INFO Synchronously wait child process to complete : hadoop jar /mnt/var/lib/hadoop/steps/s-KY0KCJ10...
INFO waitProcessCompletion ended with exit code 0 : hadoop jar /mnt/var/lib/hadoop/steps/s-KY0KCJ10...
INFO total process run time: 528 seconds
2017-09-25T00:10:47.196Z INFO Step created jobs: job_1506297597294_0001
2017-09-25T00:10:47.197Z INFO Step succeeded with exitCode 0 and took 528 seconds
```

Syslog

Secure | https://cs6240mr.s3-us-west-2.amazonaws.com/logs/j-2H1KWYS4PP9FX/steps/s-KY0KCJ10MPIP/syslog.gz?X-Amz-Security-Token=AgoG...

```
2017-09-25 00:02:05,837 INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl (main): Timeline service address: http://ip-172-31-29-50.us-west-
2.compute.internal:8188/ws/v1/timeline/
2017-09-25 00:02:05,875 INFO org.apache.hadoop.yarn.client.RMProxy (main): Connecting to ResourceManager at ip-172-31-29-50.us-west-
2.compute.internal/172.31.29.50:8032
2017-09-25 00:02:07,181 WARN org.apache.hadoop.mapreduce.JobResourceUploader (main): Hadoop command-line option parsing not performed. Implement the Tool
interface and execute your application with ToolRunner to remedy this.
2017-09-25 00:02:08,354 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat (main): Total input paths to process : 1
2017-09-25 00:02:08,367 INFO com.hadoop.compression.lzo.GPLNativeCodeLoader (main): Loaded native gpl library
2017-09-25 00:02:08,369 INFO com.hadoop.compression.lzo.LzoCodec (main): Successfully loaded & initialized native-lzo library [hadoop-lzo rev
cb482944667f96f43c89932dcb66d61ee7e4ac1d]
2017-09-25 00:02:08,979 INFO org.apache.hadoop.mapreduce.JobSubmitter (main): number of splits:22
2017-09-25 00:02:09,418 INFO org.apache.hadoop.mapreduce.JobSubmitter (main): Submitting tokens for job: job_1506297597294_0001
2017-09-25 00:02:09,902 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl (main): Submitted application application_1506297597294_0001
2017-09-25 00:02:10,032 INFO org.apache.hadoop.mapreduce.Job (main): The url to track the job: http://ip-172-31-29-50.us-west-
2.compute.internal:20888/proxy/application_1506297597294_0001/
2017-09-25 00:02:10,033 INFO org.apache.hadoop.mapreduce.Job (main): Running job: job_1506297597294_0001
2017-09-25 00:02:21,171 INFO org.apache.hadoop.mapreduce.Job (main): Job job_1506297597294_0001 running in uber mode : false
2017-09-25 00:02:21,172 INFO org.apache.hadoop.mapreduce.Job (main):  map 0% reduce 0%
2017-09-25 00:02:38,286 INFO org.apache.hadoop.mapreduce.Job (main):  map 1% reduce 0%
2017-09-25 00:02:41,308 INFO org.apache.hadoop.mapreduce.Job (main):  map 2% reduce 0%
2017-09-25 00:02:42,318 INFO org.apache.hadoop.mapreduce.Job (main):  map 3% reduce 0%
2017-09-25 00:02:44,329 INFO org.apache.hadoop.mapreduce.Job (main):  map 4% reduce 0%
2017-09-25 00:02:51,361 INFO org.apache.hadoop.mapreduce.Job (main):  map 5% reduce 0%
2017-09-25 00:02:54,379 INFO org.apache.hadoop.mapreduce.Job (main):  map 7% reduce 0%
2017-09-25 00:02:56,388 INFO org.apache.hadoop.mapreduce.Job (main):  map 8% reduce 0%
2017-09-25 00:02:57,394 INFO org.apache.hadoop.mapreduce.Job (main):  map 10% reduce 0%
2017-09-25 00:02:59,403 INFO org.apache.hadoop.mapreduce.Job (main):  map 11% reduce 0%
2017-09-25 00:03:00,409 INFO org.apache.hadoop.mapreduce.Job (main):  map 12% reduce 0%
2017-09-25 00:03:01,414 INFO org.apache.hadoop.mapreduce.Job (main):  map 13% reduce 0%
2017-09-25 00:03:04,426 INFO org.apache.hadoop.mapreduce.Job (main):  map 14% reduce 0%
2017-09-25 00:03:06,434 INFO org.apache.hadoop.mapreduce.Job (main):  map 15% reduce 0%
2017-09-25 00:03:08,444 INFO org.apache.hadoop.mapreduce.Job (main):  map 16% reduce 0%
2017-09-25 00:03:11,454 INFO org.apache.hadoop.mapreduce.Job (main):  map 17% reduce 0%
2017-09-25 00:03:26,544 INFO org.apache.hadoop.mapreduce.Job (main):  map 18% reduce 0%
2017-09-25 00:03:29,565 INFO org.apache.hadoop.mapreduce.Job (main):  map 19% reduce 0%
2017-09-25 00:03:30,569 INFO org.apache.hadoop.mapreduce.Job (main):  map 21% reduce 0%
2017-09-25 00:03:31,573 INFO org.apache.hadoop.mapreduce.Job (main):  map 23% reduce 0%
2017-09-25 00:03:33,587 INFO org.apache.hadoop.mapreduce.Job (main):  map 25% reduce 0%
2017-09-25 00:03:45,648 INFO org.apache.hadoop.mapreduce.Job (main):  map 26% reduce 0%
2017-09-25 00:03:48,662 INFO org.apache.hadoop.mapreduce.Job (main):  map 27% reduce 0%
```

Secure | https://cs6240mr.s3-us-west-2.amazonaws.com/logs/j-2H1KWYS4PP9FX/steps/s-KY0KCJ10MPIP/syslog.gz?X-Amz-Security-Token=AgoG...

```
                Data-local map tasks=22
                Total time spent by all maps in occupied slots (ms)=87580992
                Total time spent by all reduces in occupied slots (ms)=48441696
                Total time spent by all map tasks (ms)=1824604
                Total time spent by all reduce tasks (ms)=504601
                Total vcore-milliseconds taken by all map tasks=1824604
                Total vcore-milliseconds taken by all reduce tasks=504601
                Total megabyte-milliseconds taken by all map tasks=2802591744
                Total megabyte-milliseconds taken by all reduce tasks=1550134272
        Map-Reduce Framework
                Map input records=21907700
                Map output records=248943500
                Map output bytes=2418234700
                Map output materialized bytes=139613162
                Input split bytes=2024
                Combine input records=0
                Combine output records=0
                Reduce input groups=5273
                Reduce shuffle bytes=139613162
                Reduce input records=248943500
                Reduce output records=5273
                Spilled Records=746830500
                Shuffled Maps =66
                Failed Shuffles=0
                Merged Map outputs=66
                GC time elapsed (ms)=20730
                CPU time spent (ms)=1069580
                Physical memory (bytes) snapshot=19995815936
                Virtual memory (bytes) snapshot=86531956736
                Total committed heap usage (bytes)=19304284160
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1454251378
        File Output Format Counters
                Bytes Written=72815
```