**Log Aggregation MRv2 Job – Version 1**

We will create a new MR job to aggregate NASA logs.

The job will use **JDK 1.7 and we will run it on Cloudera's QuickStart VM**

Why JDK 1.7? Because this is what you get pre-installed on Cloudera QuickStart VM.

Unless you upgrade the VM to JDK8 - you need to use JDK1.7 to compile your MR job.

We will be working with a public set of data - NASA web server logs:

<http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html>

In particular - this archive:

achine generated alternative text:
Distribution 
Available from the Archive in Jul 01 to .Jul 31, ASCII format, 20.7 MB gzip compressed, 205.2 MB uncompressed, 
uncompressed. 

After you unzip it:

achine generated alternative text:
input_logs 
.DS Store 
NASA access_log_ 
Ju195 
Feb 4, 2016, 10:36 PM 
Jan 31, 2016, 4:37 PM 
Jan 30, 2016, 3:31 PM 
6 KB 
205.2 MB 

Create new Eclipse Java Project: **CSCIE63\_MRv2**

Add the created CDH Lib to your project Build Path:

Right click on the project -> Build Path -> Configure Build Path …

achine generated alternative text:
Package Explorer 
CSClE63_M 
cscie6 
Log 
cscie6 
Ecli 
Log 
cscie6 
cscie6 
edu.hu 
test 
cscie6 
Log 
*JRE syst 
* CDH5.5_ 
Binput 
Binput_cite 
Binput_log 
Binput_log 
MRv2_JDK8 
v e src 
cscie6 
Ecli 
Log 
Log 
edu.hu 
Inv 
test 
cscie6 
Log 
*JRE syst 
*CDH5.5_ 
CSClE63 MRv2 
JU JUnit 
JU LogParsi 
JU LogParsing 
OhJ JUnit 
New 
Go Into 
Open in New Window 
Open Type Hierarchy 
Show In 
[C Copy 
Copy Qualified Name 
Paste 
Delete 
Remove from Context 
Build Path 
Source 
Refactor 
Import... 
Export... 
Refresh 
Close Project 
Close Unrelated Projects 
Assign Working Sets... 
Run As 
Debug As 
Validate 
Team 
Compare With 
Restore from Local History... 
Configure 
Properties 
LogParsingUtiI. 
LogParsingUtilT B 
package cscie63.section1.agg; 
ort static org.junit.Assert.*• 
Unit tests for NASA Log Parsing 
@author mgr.inqp.pppy.q 
lic class LogParsingUtilTest { 
@Test 
public void testGetDateStrFroml 
String testLine — 
" 199 . 72 A 
String expectedDateHour 
Link Source... 
New Source Folder... 
Use as Source Folder 
Add External Archives... 
Add Libraries... 
O: Configure Build Path... 
File 
File 
Input Format Counters 
Bytes Read=12Ø2 
Output Format Counters 
Bytes Written=8Ø 

In the "Libraries" tab:

"Add Library…" -> "User Library " -> select the CDH5.5\_mrv2\_full lib

achine generated alternative text:
Add Library 
User Library 
Select a library. 
User libraries: 
User Libraries... 

Final config:

achine generated alternative text:
type filter text 
Resource 
Builders 
Java Build Path 
Java Code Style 
Java Compiler 
Java Editor 
Javadoc Location 
Project References 
Refactoring History 
Run/Debug Settings 
Task Repository 
Task Tags 
Validation 
WikiText 
Properties for CSClE63_MRv2 
Java Build Path 
Source 
Projects 
JARs and class folders on the build path: 
Libraries 
*JRE System Library [Java SE 7 [1.7.0_71]] 
Order and Export 
Add JARs... 
Add External JARs... 
Add Variable... 
Add Library... 
Add Class Folder... 
Add External Class Folder... 

Get the provided code:

achine generated alternative text:
CSClE63_MRv2 
.classpath 
.DS Store 
.project 
input_logs_short 
src 
.DS Store 
cscie63 
.DS Store 
sectionl 
.DS Store 
agg 
EclipseLogAggDriver.java 
LogAggregatorByHour.java 
LogAggregatorFirst.java 
LogParsingUtil.java 

Create new Run Configuration:

achine generated alternative text:
Create, manage, and run configurations 
Run a Java application 
Name: 
O 
EclipseLogAggDriver 
Arguments 
Run Configurations 
Classpath 
type filter text 
Gradle Project 
gjj Java Applet 
Java Application 
EclipseLogAggDriver 
EclipseLogAggDriver Date 
InverterCounterMRv2 
InverterCounterMRv2_bgc 
InverterList 
InverterList_JDK8 
LogAggregatorByHourAnc 
v JuJUnit 
JU LogParsingUtilTest 
JU LogParsingUtilTest.testGe 
Ju LogParsingUtilTest.testGe 
rrü Maven Build 
Ju Task Context Test 
Environment 
Common 
Project: 
CSClE63 MRv2 
Main class: 
cscie63 .sectionl . agg. EclipseLogAggDriver 
n 
Include system libraries when searching for a main class 
n 
Include inherited mains when searching for a main class 
n 
Stop in main 
Browse... 
Search... 

Setup Arguments:

achine generated alternative text:
Create, manage, and run configurations 
Run a Java application 
Name: 
type filter text 
EclipseLogAggDriver 
Arguments 
Run Configurations 
Classpath 
Environment 
Common 
Gradle Project 
gjjl Java Applet 
Java Application 
EclipseLogAggDriver 
EclipseLogAggDriver Date 
InverterCounterMRv2 
InverterCounterMRv2_bgc 
InverterList 
InverterList_JDK8 
LogAggregatorByHourAnc 
v JUJUnit 
JU LogParsingUtilTest 
JU LogParsingUtilTest .testGe 
JU LogParsingUtilTest .testGe 
rtü Maven Build 
Ju Task Context Test 
Filter matched 16 of 16 items 
Program arguments: 
input_logs_short output_aggl 
VM arguments: 
o 
Use the -XstartOnFirstThread argument when launching with SWT 
Working directory: 
O 
Default: 
C) 
Other: 
Workspace. 
Variables... 
Variables... 
Variables... 
File System... 
Revert 

Run the Job in Eclipse:

achine generated alternative text:
@ Javadoc 
Problems 
Declaration = Console B 
<terminated> EclipseLogAggDriver [Java Application] /Library/Java/JavaVirtualMachines/jdk1.7.0_71 .jdk/Contents/Home/bin/java (Feb 1 3, 201 6, 12:08:20 AM) 
-JPL L L CU 
Shuffled Maps =1 
Failed Shuffles=Ø 
Merged Map outputs—I 
GC time elapsed 
Total committed heap usage (bytes)=51485Ø816 
Shuffle Errors 
CONNECTION=Ø 
File Input Format Counters 
Bytes Read=12Ø2 
File Output Format Counters 
Bytes Written=8Ø 

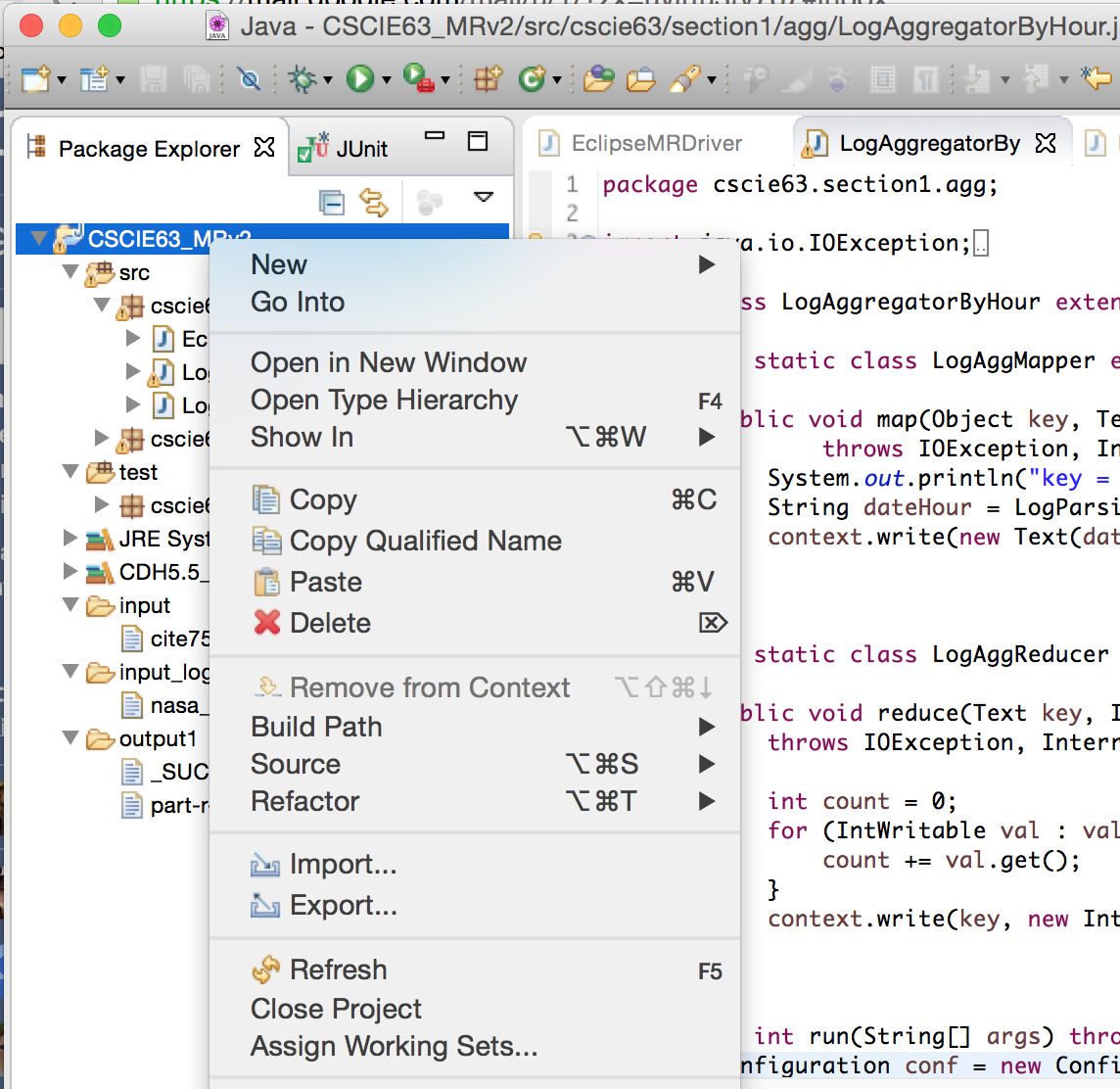
Refresh Eclipse project (to see new created output directory) - and Inspect output :

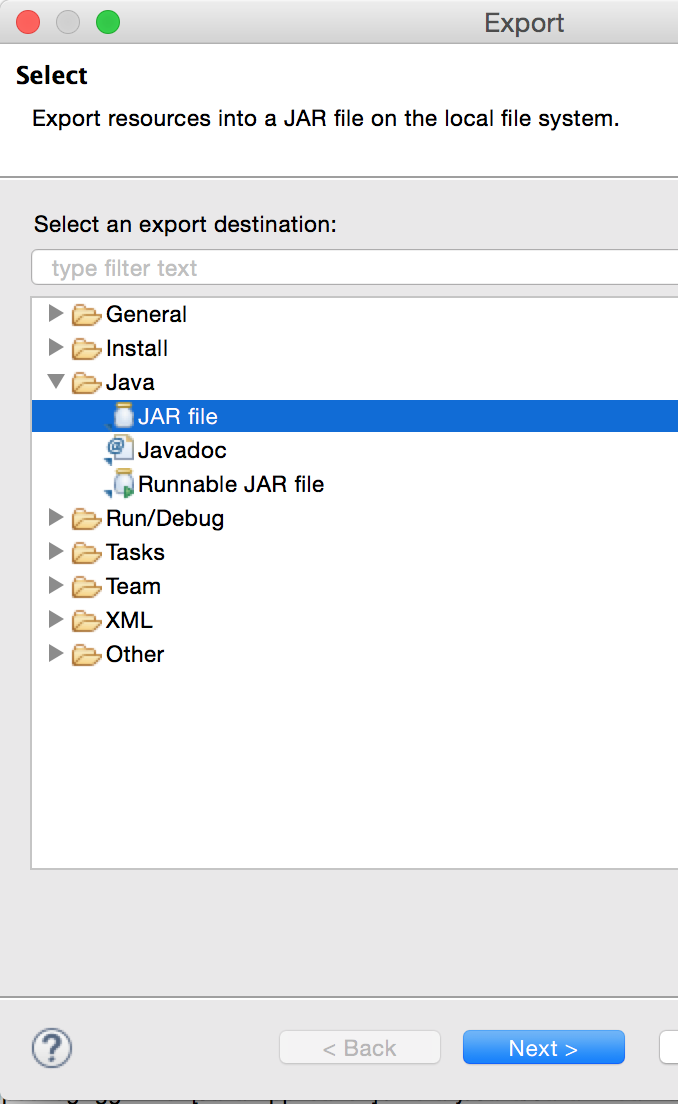
achine generated alternative text:
Java - /part-r-00000 - Eclipse - /Users/marinapopova/Marina/GoogleDrive[ 
Package Explorer 
CSClE63_MRv2 
test 
JUnit 
LogParsingUtilT 
maul 71995 : 00 
01/Au1/1995 : 02 
Ø2/Auq/1995 : 02 
: 02 
@ LogAggregatorBy 
9 
1 
1 
1 
EclipseLogAggDr 
part-r-OOOOO B 
*JRE System Library 
[Java SE 7 [1.7.0_71]] 
Binput 
Binput_cited_short 
Binput_logs 
v Boutput_aggl 
SUCCESS 
part-r-OOOOO 
MRv2_JDK8 
@ Javadoc 
Problems 
Declaration = Console B 
<terminated> EclipseLogAggDriver [Java Application] /Library/Java/JavaVirtualMachines/jdk1.7.0_71. 
Shuffled Maps =1 
Failed Shuffles=Ø 
Merged Map outputs—I 
GC time elapsed 
Total committed heap usage (bytes)=51485Ø816 

**Now we want to run the job on our VM**

Create a JAR from your Eclipse Project:

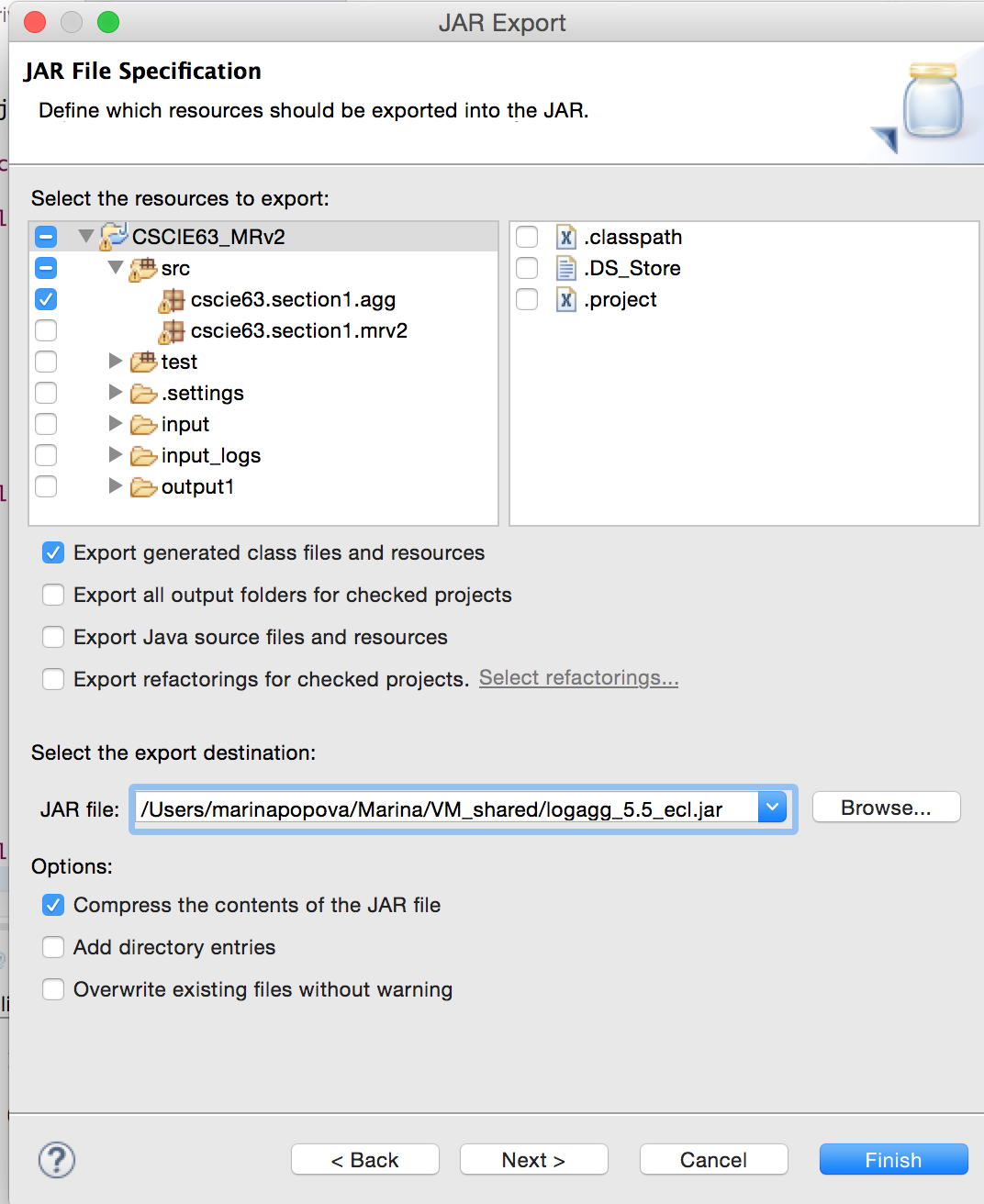
Right-click on project name -> Export :



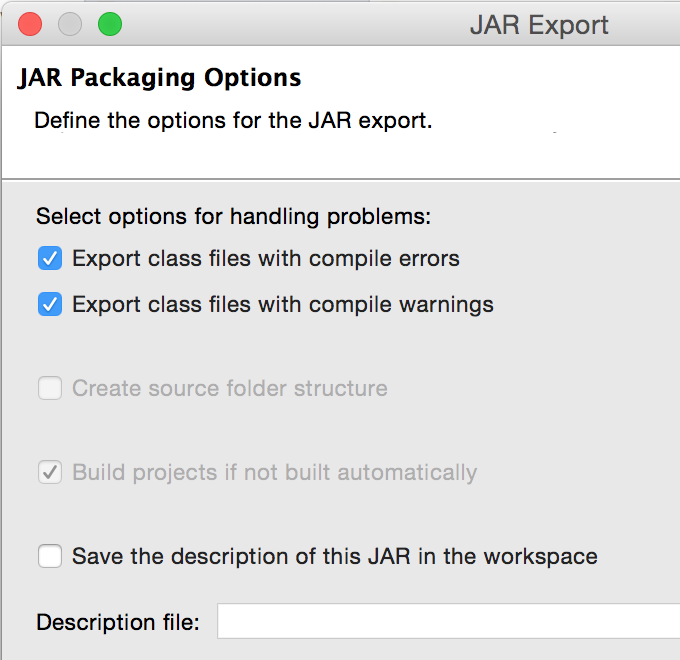


uncheck all un-necessary items (like test, input, output folders, Eclipse artifacts, etc.) - only leave your classes

select name and location of your target JAR

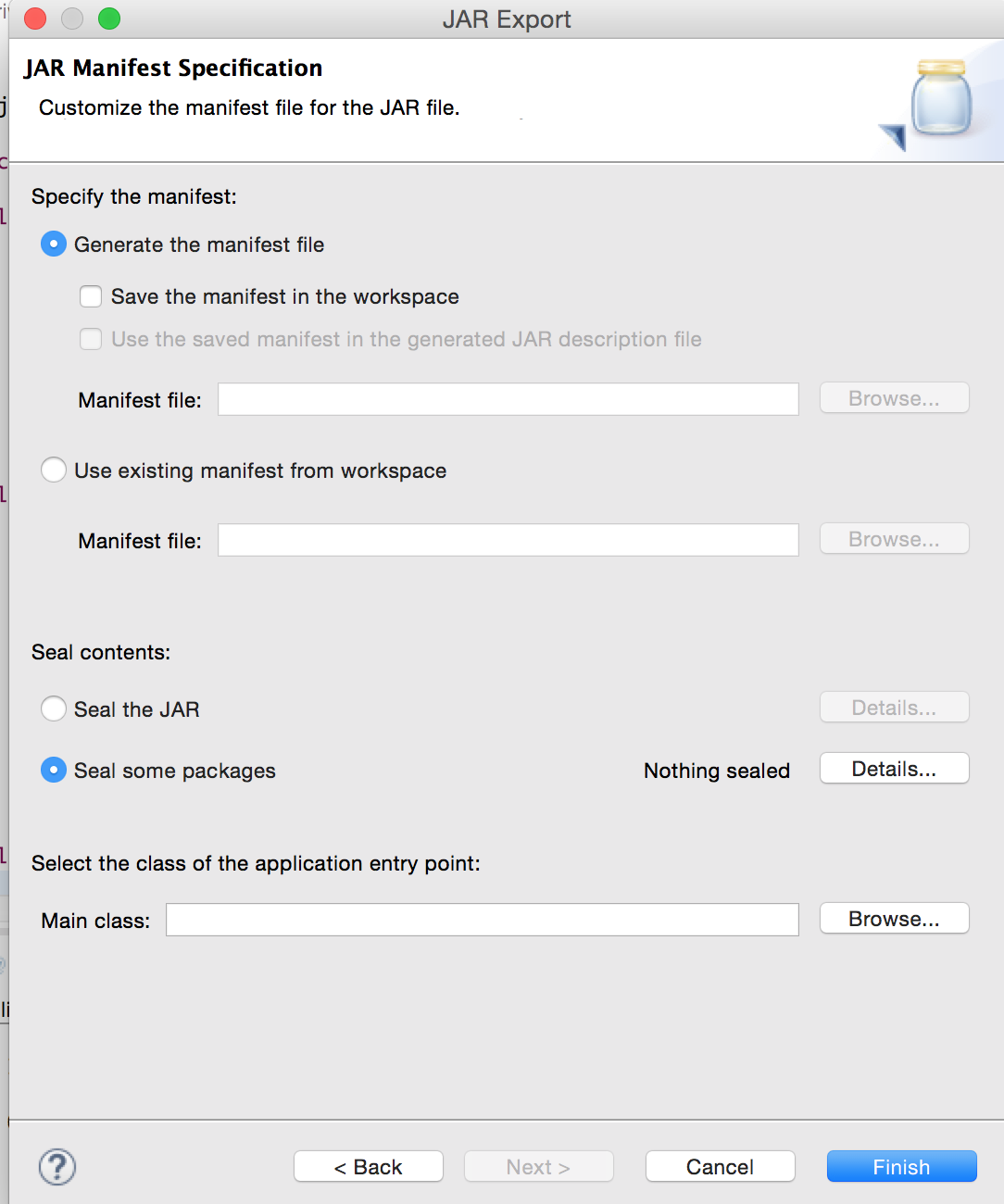


click Next



click Next

select “Generate manifest file “ and do not specify the Main class:



Click Finish:

you may get this warning:



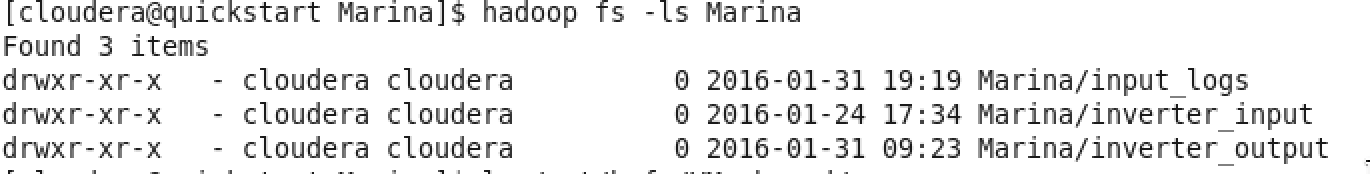
it’s Ok - click Ok

You’ve got your new jar created - copy it to a folder that is shared with your CDH VM

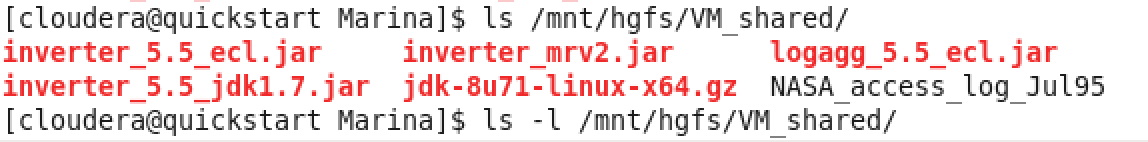
Start your **CDH5.5 QuickStart VM** and copy the jar into your working directory

Copy NASA log data into HDFS input\_logs dir (create it first):

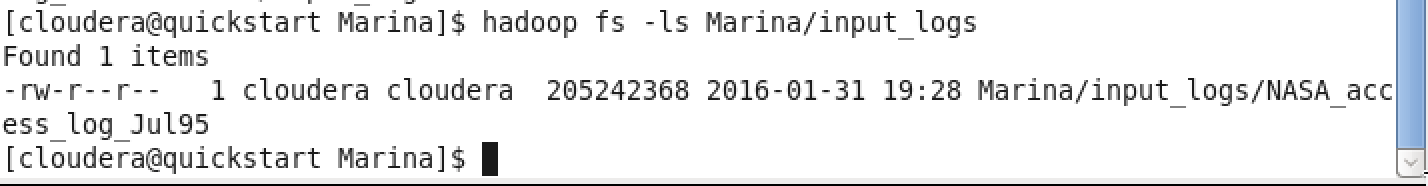
[cloudera@quickstart Marina]$ **hadoop fs -mkdir Marina/input\_logs**

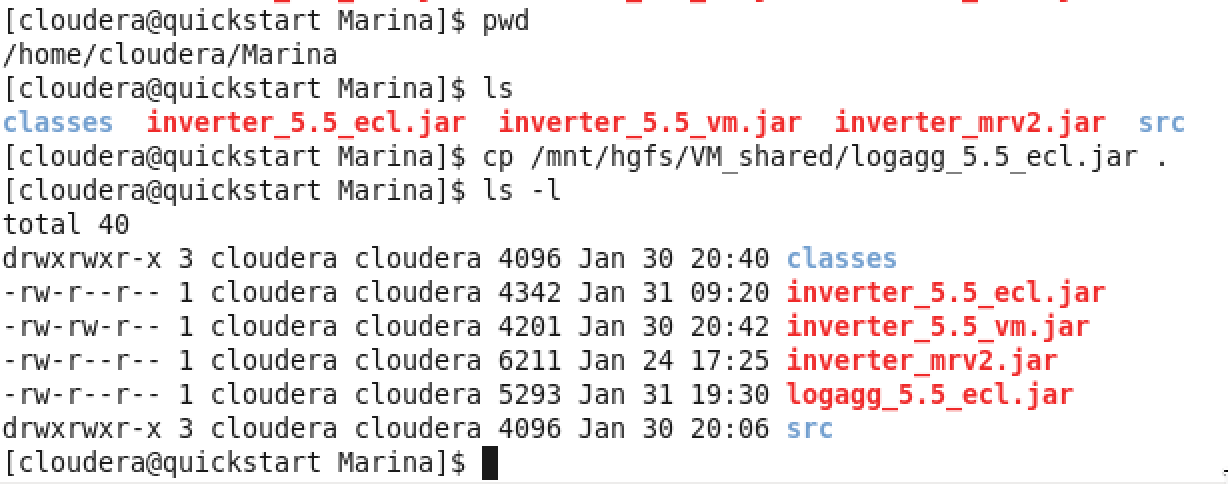


verify we have our created JAR and input data in the shared folder:



[cloudera@quickstart Marina]$ **hadoop fs -copyFromLocal /mnt/hgfs/VM\_shared/NASA\_access\_log\_Jul95 Marina/input\_logs**





checkout what is Hadoop’s classpath:



If you already have an output dir created - delete it:

**hadoop fs -rm -r Marina/agg\_output**

Run the job:

First attempt: use class **cscie63.section1.agg.LogAggregatorFirst** which does not have correct error handling:

[cloudera@quickstart Marina]$ **hadoop jar logagg\_5.5\_ecl.jar cscie63.section1.agg.LogAggregatorFirst Marina/input\_logs Marina/agg\_output**

*16/01/31 19:34:36 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032*

*16/01/31 19:34:37 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.*

*16/01/31 19:34:37 INFO input.FileInputFormat: Total input paths to process : 1*

*16/01/31 19:34:37 INFO mapreduce.JobSubmitter: number of splits:2*

*16/01/31 19:34:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1454295585434\_0001*

*16/01/31 19:34:38 INFO impl.YarnClientImpl: Submitted application application\_1454295585434\_0001*

*16/01/31 19:34:38 INFO mapreduce.Job: The url to track the job:*[*http://quickstart.cloudera:8088/proxy/application\_1454295585434\_0001/*](http://quickstart.cloudera:8088/proxy/application_1454295585434_0001/)

*16/01/31 19:34:38 INFO mapreduce.Job: Running job: job\_1454295585434\_0001*

*16/01/31 19:34:48 INFO mapreduce.Job: Job job\_1454295585434\_0001 running in uber mode : false*

*16/01/31 19:34:48 INFO mapreduce.Job:  map 0% reduce 0%*

*16/01/31 19:35:03 INFO mapreduce.Job:  map 8% reduce 0%*

*16/01/31 19:35:05 INFO mapreduce.Job:  map 23% reduce 0%*

*16/01/31 19:35:06 INFO mapreduce.Job:  map 30% reduce 0%*

*16/01/31 19:35:08 INFO mapreduce.Job:  map 43% reduce 0%*

*16/01/31 19:35:09 INFO mapreduce.Job:  map 51% reduce 0%*

*16/01/31 19:35:09 INFO mapreduce.Job: Task Id : attempt\_1454295585434\_0001\_m\_000001\_0, Status : FAILED*

*Error: java.lang.NullPointerException*

*at org.apache.hadoop.io.Text.encode(Text.java:450)*

*at org.apache.hadoop.io.Text.set(Text.java:198)*

*at org.apache.hadoop.io.Text.<init>(Text.java:88)*

*at cscie63.section1.agg.LogAggregatorByHour$LogAggMapper.map(LogAggregatorByHour.java:31)*

*at cscie63.section1.agg.LogAggregatorByHour$LogAggMapper.map(LogAggregatorByHour.java:1)*

*at org.apache.hadoop.mapreduce.Mapper.run(Mapper.java:145)*

*at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:787)*

*at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)*

*at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)*

*at java.security.AccessController.doPrivileged(Native Method)*

*at javax.security.auth.Subject.doAs(Subject.java:415)*

*at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1671)*

*at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)*

*16/01/31 19:35:10 INFO mapreduce.Job:  map 22% reduce 0%*

*16/01/31 19:35:12 INFO mapreduce.Job:  map 32% reduce 0%*

*16/01/31 19:35:15 INFO mapreduce.Job:  map 33% reduce 0%*

*16/01/31 19:35:16 INFO mapreduce.Job:  map 50% reduce 0%*

*16/01/31 19:35:24 INFO mapreduce.Job:  map 63% reduce 0%*

*16/01/31 19:35:27 INFO mapreduce.Job:  map 71% reduce 0%*

*16/01/31 19:35:30 INFO mapreduce.Job:  map 81% reduce 0%*

*16/01/31 19:35:32 INFO mapreduce.Job:  map 81% reduce 17%*

*16/01/31 19:35:32 INFO mapreduce.Job: Task Id : attempt\_1454295585434\_0001\_m\_000001\_1, Status : FAILED*

*Error: java.lang.NullPointerException*

*at org.apache.hadoop.io.Text.encode(Text.java:450)*

*at org.apache.hadoop.io.Text.set(Text.java:198)*

*at org.apache.hadoop.io.Text.<init>(Text.java:88)*

*at cscie63.section1.agg.LogAggregatorByHour$LogAggMapper.map(LogAggregatorByHour.java:31)*

*at cscie63.section1.agg.LogAggregatorByHour$LogAggMapper.map(LogAggregatorByHour.java:1)*

*at org.apache.hadoop.mapreduce.Mapper.run(Mapper.java:145)*

*at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:787)*

*at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)*

*at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)*

*at java.security.AccessController.doPrivileged(Native Method)*

*at javax.security.auth.Subject.doAs(Subject.java:415)*

*at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1671)*

*at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)*

*16/01/31 19:35:33 INFO mapreduce.Job:  map 50% reduce 17%*

*16/01/31 19:35:38 INFO mapreduce.Job: Task Id : attempt\_1454295585434\_0001\_m\_000001\_2, Status : FAILED*

*Error: java.lang.NullPointerException*

*at org.apache.hadoop.io.Text.encode(Text.java:450)*

*at org.apache.hadoop.io.Text.set(Text.java:198)*

*at org.apache.hadoop.io.Text.<init>(Text.java:88)*

*at cscie63.section1.agg.LogAggregatorByHour$LogAggMapper.map(LogAggregatorByHour.java:31)*

*at cscie63.section1.agg.LogAggregatorByHour$LogAggMapper.map(LogAggregatorByHour.java:1)*

*at org.apache.hadoop.mapreduce.Mapper.run(Mapper.java:145)*

*at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:787)*

*at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)*

*at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)*

*at java.security.AccessController.doPrivileged(Native Method)*

*at javax.security.auth.Subject.doAs(Subject.java:415)*

*at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1671)*

*at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)*

*16/01/31 19:35:52 INFO mapreduce.Job:  map 100% reduce 100%*

*16/01/31 19:35:52 INFO mapreduce.Job: Job job\_1454295585434\_0001 failed with state FAILED due to: Task failed task\_1454295585434\_0001\_m\_000001*

*Job failed as tasks failed. failedMaps:1 failedReduces:0*

*16/01/31 19:35:53 INFO mapreduce.Job: Counters: 39*

*File System Counters*

*FILE: Number of bytes read=0*

*FILE: Number of bytes written=26006887*

*FILE: Number of read operations=0*

*FILE: Number of large read operations=0*

*FILE: Number of write operations=0*

*HDFS: Number of bytes read=134221974*

*HDFS: Number of bytes written=0*

*HDFS: Number of read operations=3*

*HDFS: Number of large read operations=0*

*HDFS: Number of write operations=0*

*Job Counters*

*Failed map tasks=4*

*Killed reduce tasks=1*

*Launched map tasks=5*

*Launched reduce tasks=1*

*Other local map tasks=3*

*Data-local map tasks=2*

*Total time spent by all maps in occupied slots (ms)=91433*

*Total time spent by all reduces in occupied slots (ms)=33278*

*Total time spent by all map tasks (ms)=91433*

*Total time spent by all reduce tasks (ms)=33278*

*Total vcore-seconds taken by all map tasks=91433*

*Total vcore-seconds taken by all reduce tasks=33278*

*Total megabyte-seconds taken by all map tasks=93627392*

*Total megabyte-seconds taken by all reduce tasks=34076672*

*Map-Reduce Framework*

*Map input records=1233097*

*Map output records=1233097*

*Map output bytes=23428843*

*Map output materialized bytes=25895043*

*Input split bytes=150*

*Combine input records=0*

*Spilled Records=1233097*

*Failed Shuffles=0*

*Merged Map outputs=0*

*GC time elapsed (ms)=290*

*CPU time spent (ms)=8710*

*Physical memory (bytes) snapshot=247508992*

*Virtual memory (bytes) snapshot=1501298688*

*Total committed heap usage (bytes)=212996096*

*File Input Format Counters*

*Bytes Read=134221824*

[cloudera@quickstart Marina]$ hadoop fs -ls Marina

Found 4 items

drwxr-xr-x   - cloudera cloudera          0 2016-01-31 19:35 Marina/agg\_output

drwxr-xr-x   - cloudera cloudera          0 2016-01-31 19:28 Marina/input\_logs

drwxr-xr-x   - cloudera cloudera          0 2016-01-24 17:34 Marina/inverter\_input

drwxr-xr-x   - cloudera cloudera          0 2016-01-31 09:23 Marina/inverter\_output

[cloudera@quickstart Marina]$ hadoop fs -ls Marina/agg\_output

[cloudera@quickstart Marina]$

Running in Eclipse - Debugging

The culprit:

2016-02-03 21:02:16,758 INFO  [LocalJobRunner Map Task Executor #0] mapred.MapTask **(MapTask.java:createSortingCollector(403)) - Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer**

**ERROR parsing DATE for this line: alyssa.p**

**, storing as an UNKNOWN date**

After the Fix:

Use the class **cscie63.section1.agg.LogAggregatorByHour** with the correct error handling:

[cloudera@quickstart Marina]$ **hadoop jar logagg\_5.5\_ecl.jar cscie63.section1.agg.LogAggregatorByHour Marina/input\_logs Marina/agg\_output**

*16/02/03 18:05:41 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032*

*16/02/03 18:05:42 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.*

*16/02/03 18:05:42 INFO input.FileInputFormat: Total input paths to process : 1*

*16/02/03 18:05:42 INFO mapreduce.JobSubmitter: number of splits:2*

*16/02/03 18:05:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1454544632686\_0003*

*16/02/03 18:05:43 INFO impl.YarnClientImpl: Submitted application application\_1454544632686\_0003*

*16/02/03 18:05:43 INFO mapreduce.Job: The url to track the job:*[*http://quickstart.cloudera:8088/proxy/application\_1454544632686\_0003/*](http://quickstart.cloudera:8088/proxy/application_1454544632686_0003/)

*16/02/03 18:05:43 INFO mapreduce.Job: Running job: job\_1454544632686\_0003*

*16/02/03 18:05:51 INFO mapreduce.Job: Job job\_1454544632686\_0003 running in uber mode : false*

*16/02/03 18:05:51 INFO mapreduce.Job:  map 0% reduce 0%*

*16/02/03 18:06:06 INFO mapreduce.Job:  map 25% reduce 0%*

*16/02/03 18:06:08 INFO mapreduce.Job:  map 58% reduce 0%*

*16/02/03 18:06:09 INFO mapreduce.Job:  map 83% reduce 0%*

*16/02/03 18:06:10 INFO mapreduce.Job:  map 100% reduce 0%*

*16/02/03 18:06:21 INFO mapreduce.Job:  map 100% reduce 100%*

*16/02/03 18:06:21 INFO mapreduce.Job: Job job\_1454544632686\_0003 completed successfully*

*16/02/03 18:06:22 INFO mapreduce.Job: Counters: 49*

*File System Counters*

*FILE: Number of bytes read=39726014*

*FILE: Number of bytes written=79787511*

*FILE: Number of read operations=0*

*FILE: Number of large read operations=0*

*FILE: Number of write operations=0*

*HDFS: Number of bytes read=205246764*

*HDFS: Number of bytes written=13203*

*HDFS: Number of read operations=9*

*HDFS: Number of large read operations=0*

*HDFS: Number of write operations=2*

*Job Counters*

*Launched map tasks=2*

*Launched reduce tasks=1*

*Data-local map tasks=2*

*Total time spent by all maps in occupied slots (ms)=32202*

*Total time spent by all reduces in occupied slots (ms)=8897*

*Total time spent by all map tasks (ms)=32202*

*Total time spent by all reduce tasks (ms)=8897*

*Total vcore-seconds taken by all map tasks=32202*

*Total vcore-seconds taken by all reduce tasks=8897*

*Total megabyte-seconds taken by all map tasks=32974848*

*Total megabyte-seconds taken by all reduce tasks=9110528*

*Map-Reduce Framework*

*Map input records=1891715*

*Map output records=1891715*

*Map output bytes=35942578*

*Map output materialized bytes=39726020*

*Input split bytes=300*

*Combine input records=0*

*Combine output records=0*

*Reduce input groups=663*

*Reduce shuffle bytes=39726020*

*Reduce input records=1891715*

*Reduce output records=663*

*Spilled Records=3783430*

*Shuffled Maps =2*

*Failed Shuffles=0*

*Merged Map outputs=2*

*GC time elapsed (ms)=619*

*CPU time spent (ms)=11150*

*Physical memory (bytes) snapshot=665718784*

*Virtual memory (bytes) snapshot=4509782016*

*Total committed heap usage (bytes)=489791488*

*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=205246464*

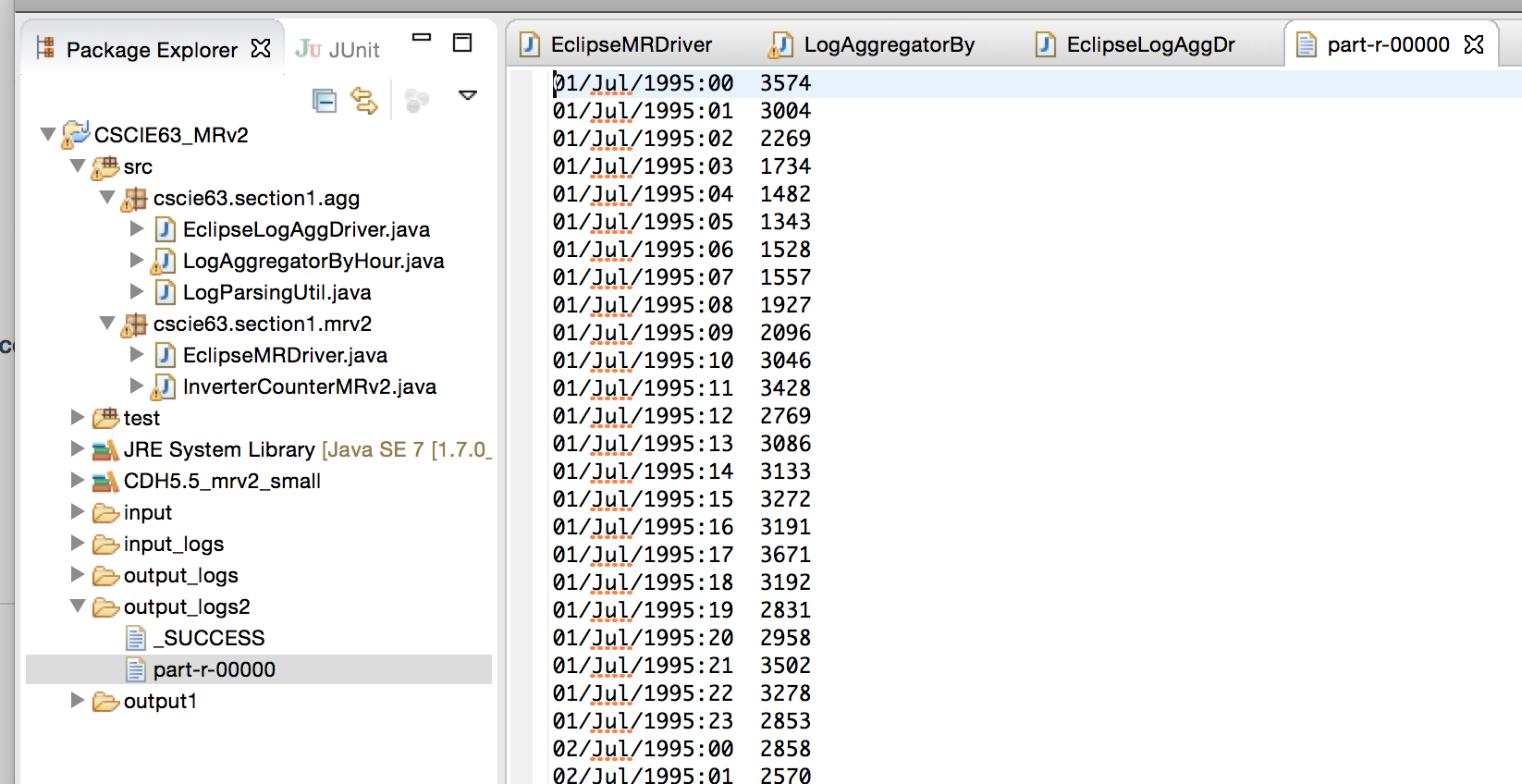
*File Output Format Counters*

*Bytes Written=13203*

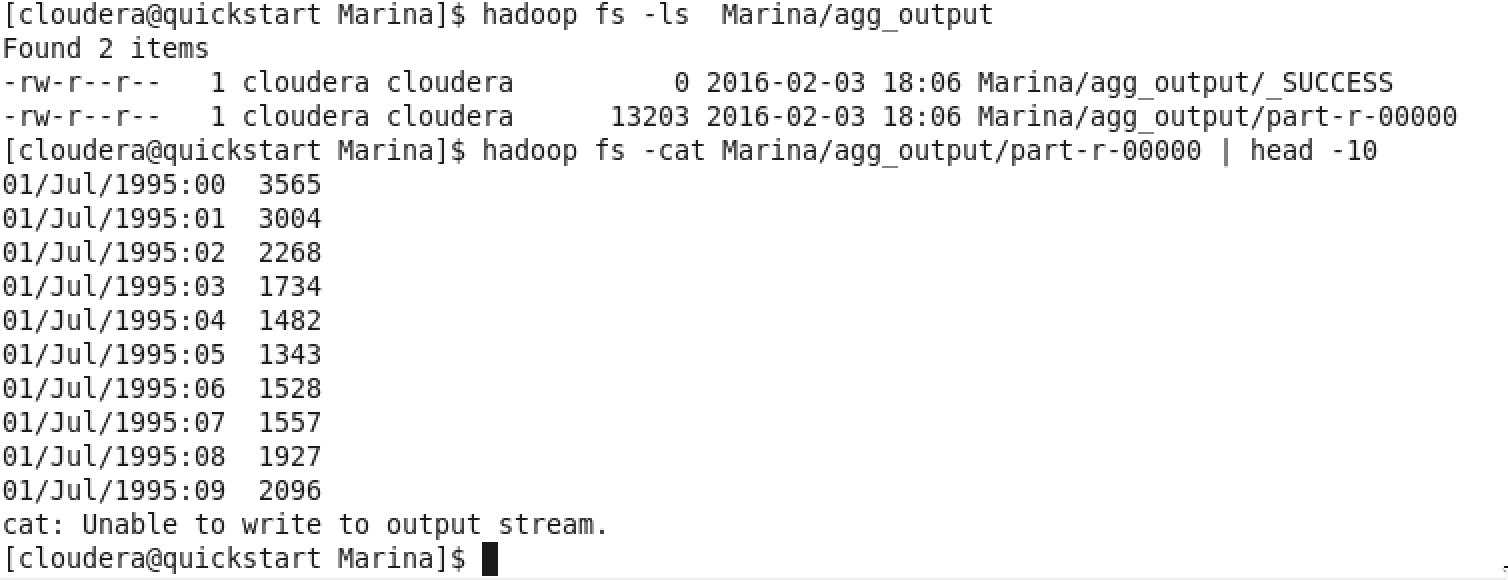
[cloudera@quickstart Marina]$

Inspect output:

In Eclipse:

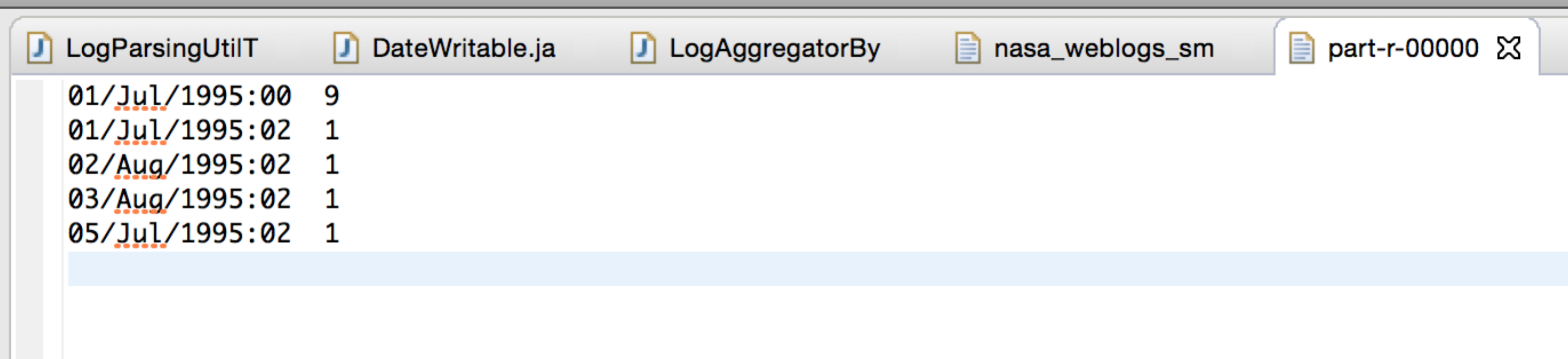


On QuickStart VM:



What is the problem with this version of the log aggregation job?

Using the provided nasa\_weblogs\_small.txt as input – run the job and inspect the results. You will see:



Notice how the entry for July 5th comes after the one for Aug 3rd?

That’s because we are treating the dates as String key types, which are sorted alphabetically, not as real Dates.

To fix that – we need to have keys of type Date. The next version of this aggregation job will do that.

Hadoop 2.6 API reference:

<https://hadoop.apache.org/docs/r2.6.0/api/index.html>