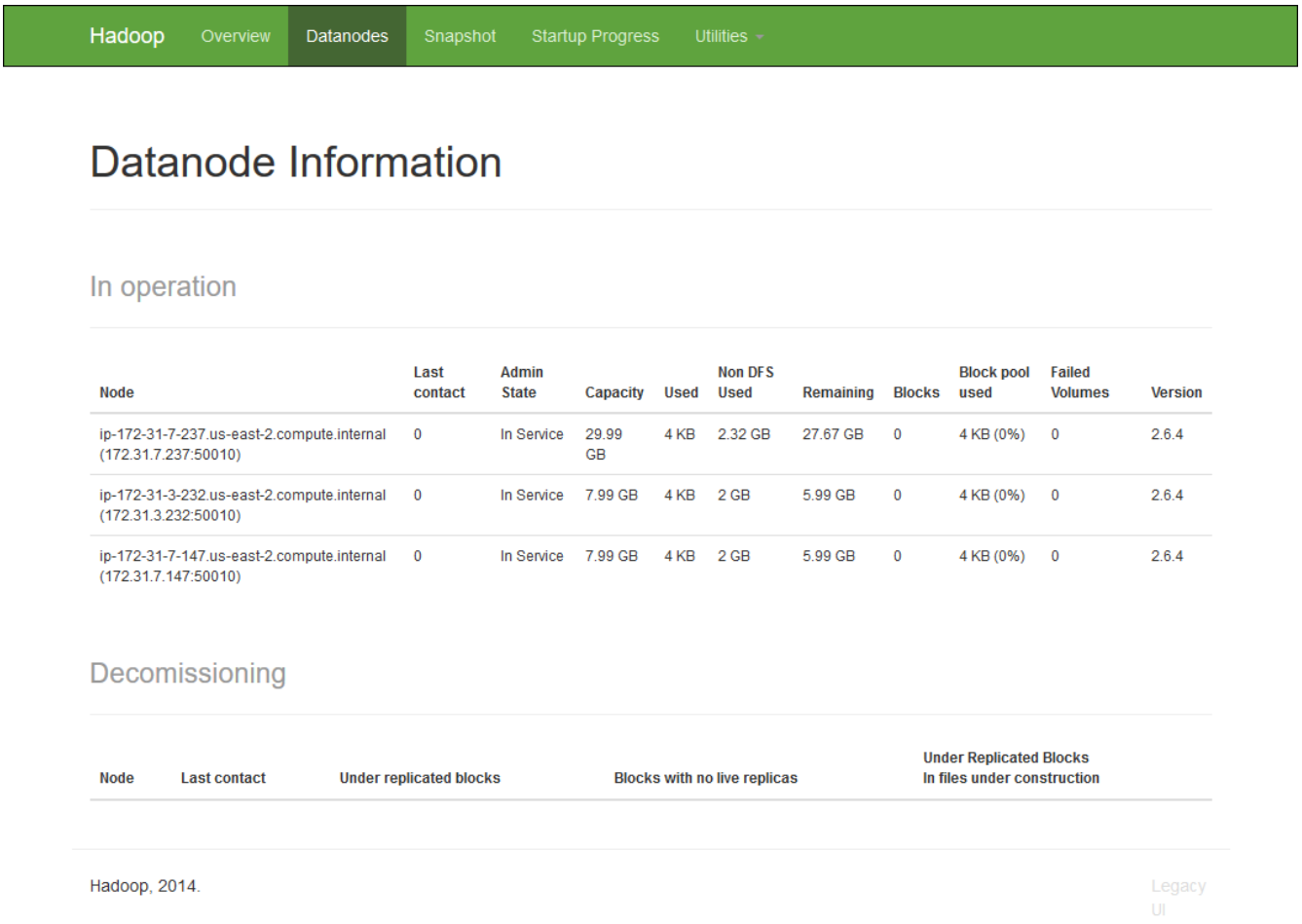
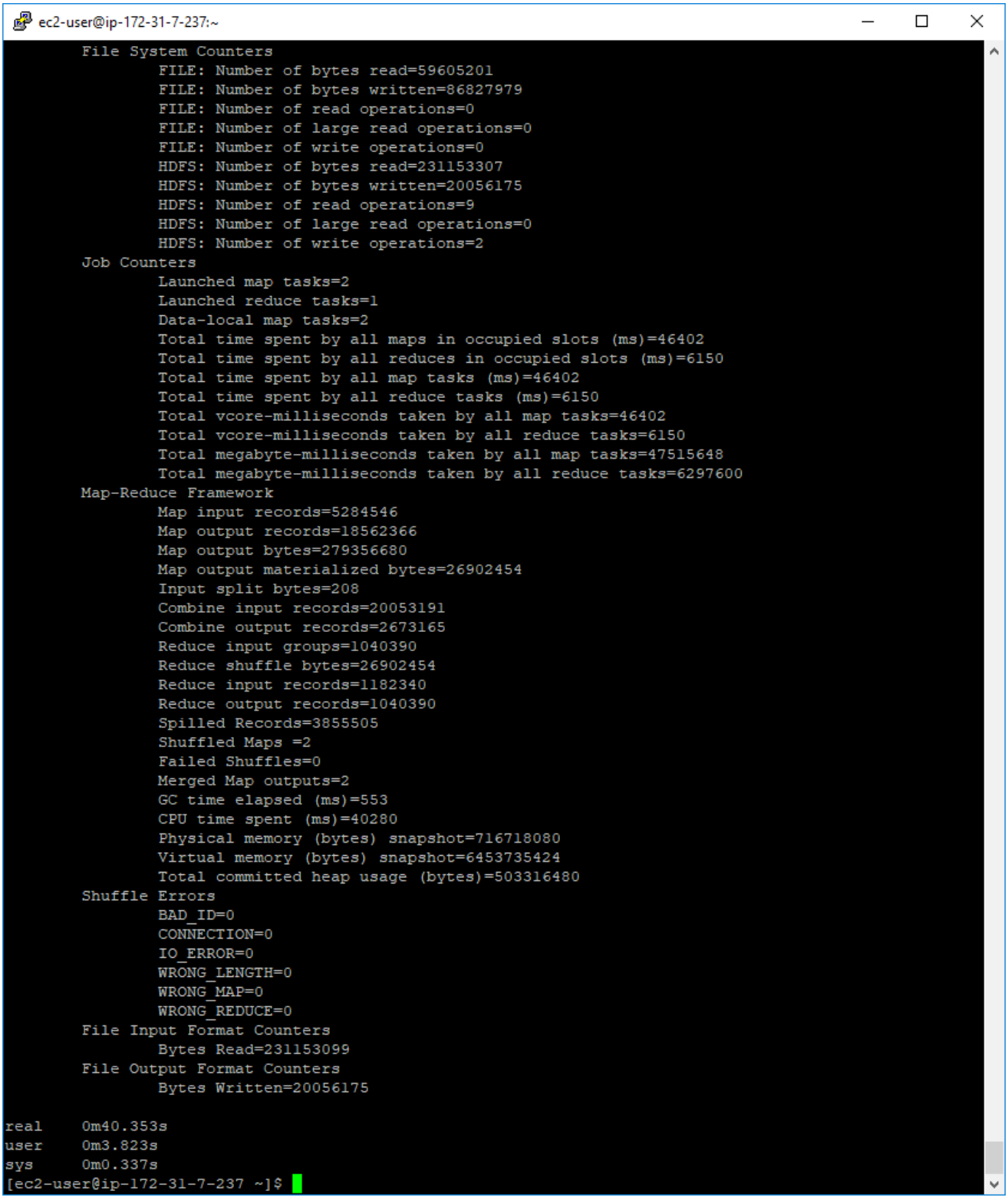
David Beck
CSC555
Project Phase 1
5/19/19

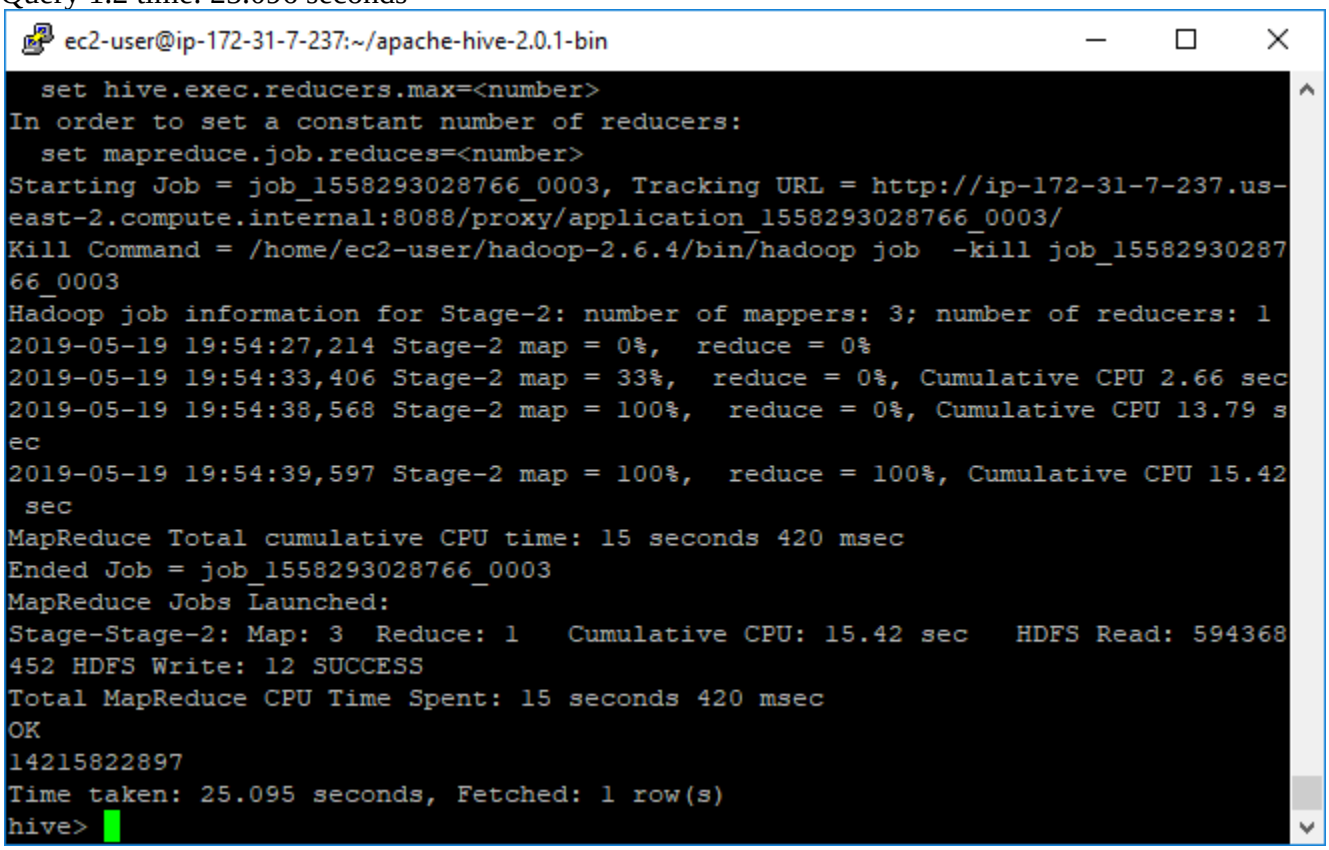
Part 1:



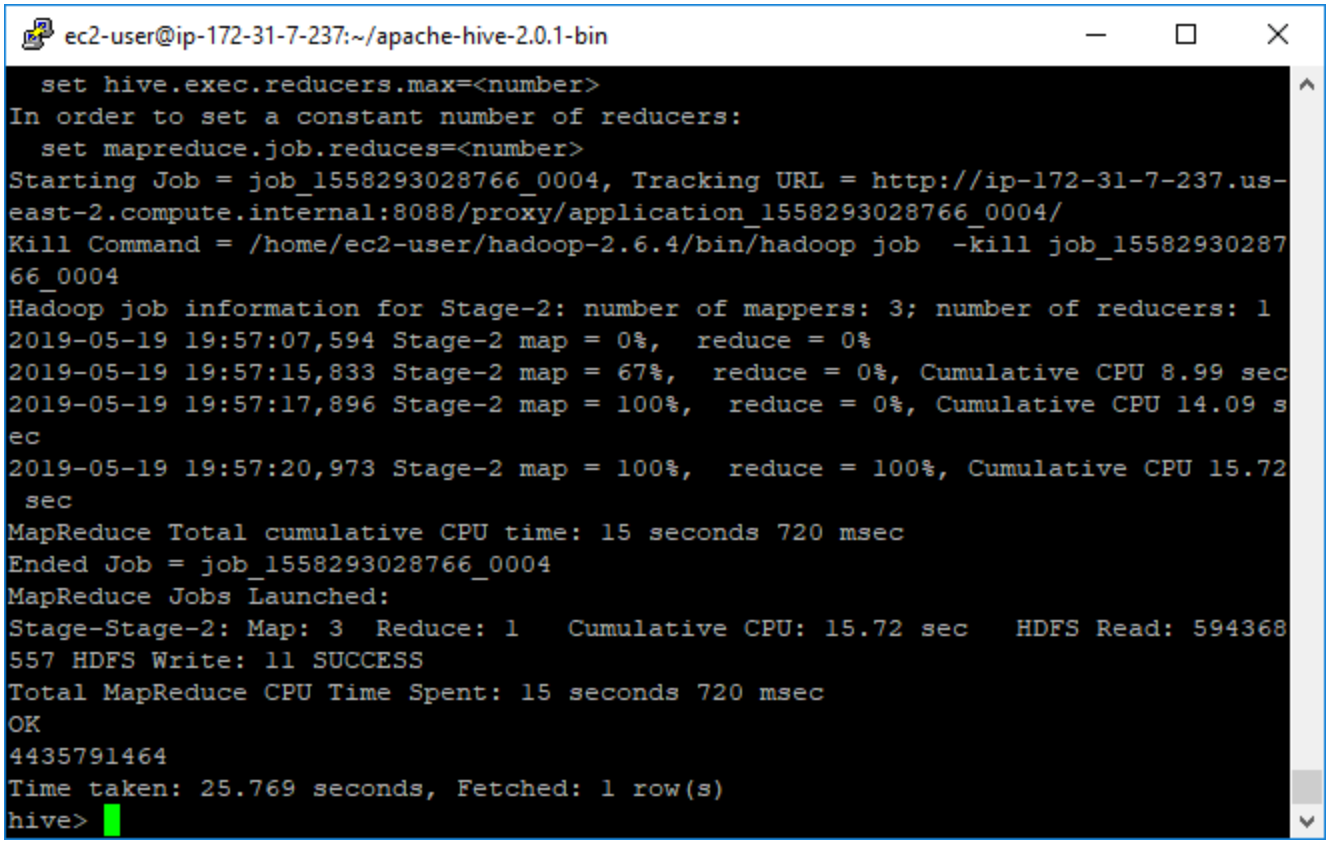


Discussion: The 3-node cluster was measurably faster than the single node at mapping the bioproject.xml data (40 seconds compared to 1 minute 10 seconds). That's a time savings of 30 seconds. The reason the 3-node cluster was able to process this data faster was because the master node could better delegate and divide up the work to the worker nodes and thereby speed up the mapreduce job overall.

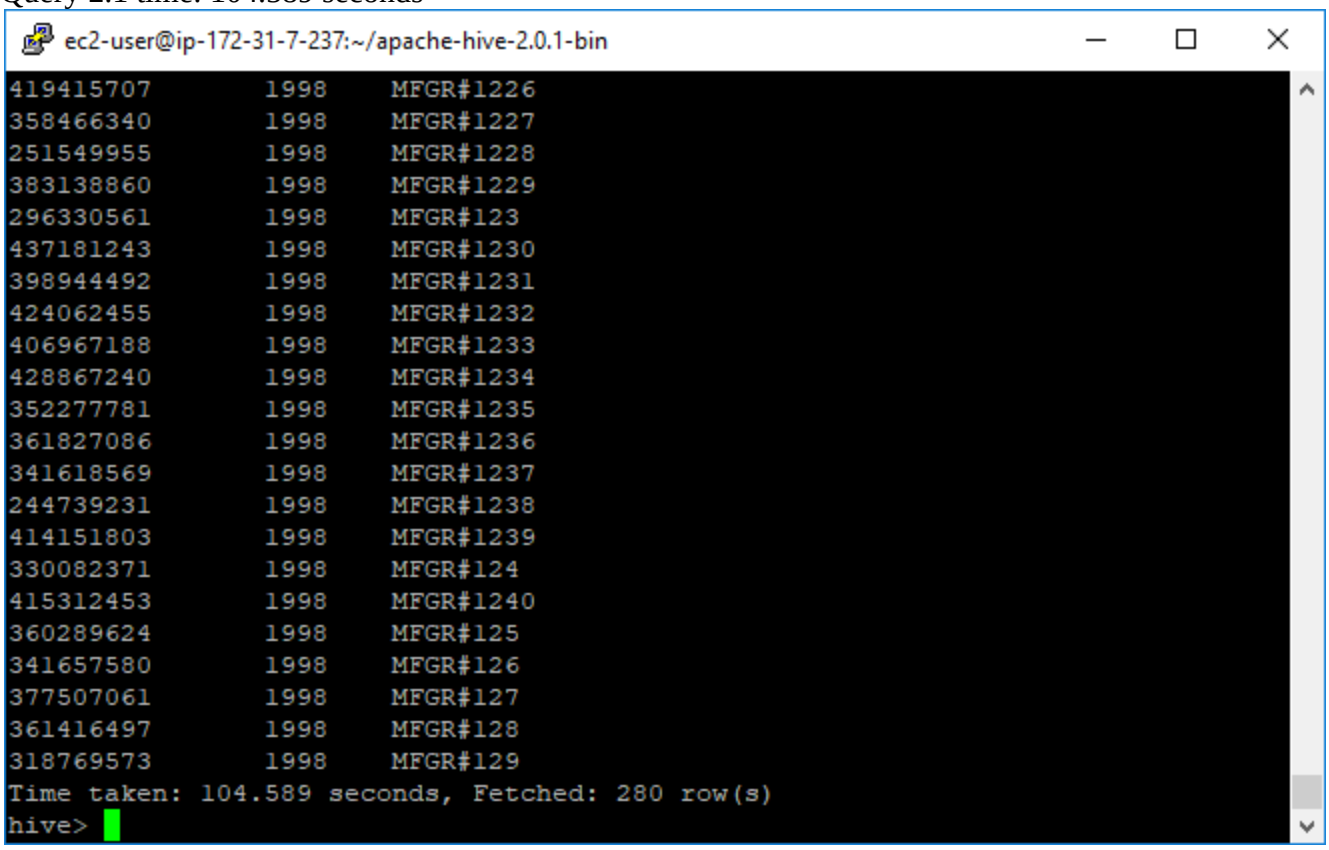
Part 2:
Query 1.2 time: 25.096 seconds



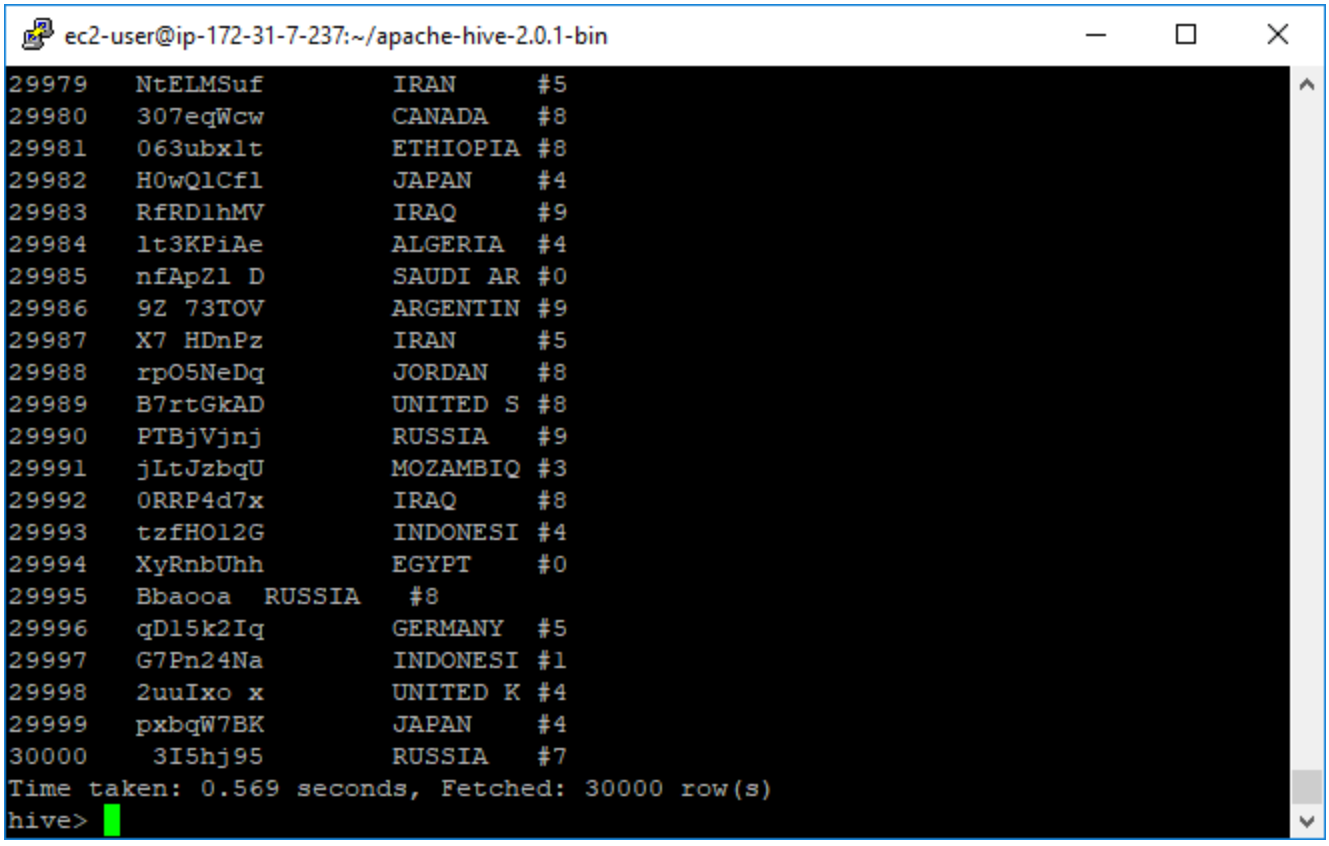
Query 1.3 time: 25.796 seconds


ec2-user@ip-172-31-7-237:~/apache-hive-2.0.1-bin
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1558293028766_0004, Tracking URL = http://ip-172-31-7-237.us-east-2.compute.internal:8088/proxy/application_1558293028766_0004/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1558293028766_0004
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2019-05-19 19:57:07,594 Stage-2 map = 0%, reduce = 0%
2019-05-19 19:57:15,833 Stage-2 map = 67%, reduce = 0%, Cumulative CPU 8.99 sec
2019-05-19 19:57:17,896 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 14.09 sec
2019-05-19 19:57:20,973 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 15.72 sec
MapReduce Total cumulative CPU time: 15 seconds 720 msec
Ended Job = job_1558293028766_0004
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3 Reduce: 1 Cumulative CPU: 15.72 sec HDFS Read: 594368557 HDFS Write: 11 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 720 msec
OK
4435791464
Time taken: 25.769 seconds, Fetched: 1 row(s)
hive>

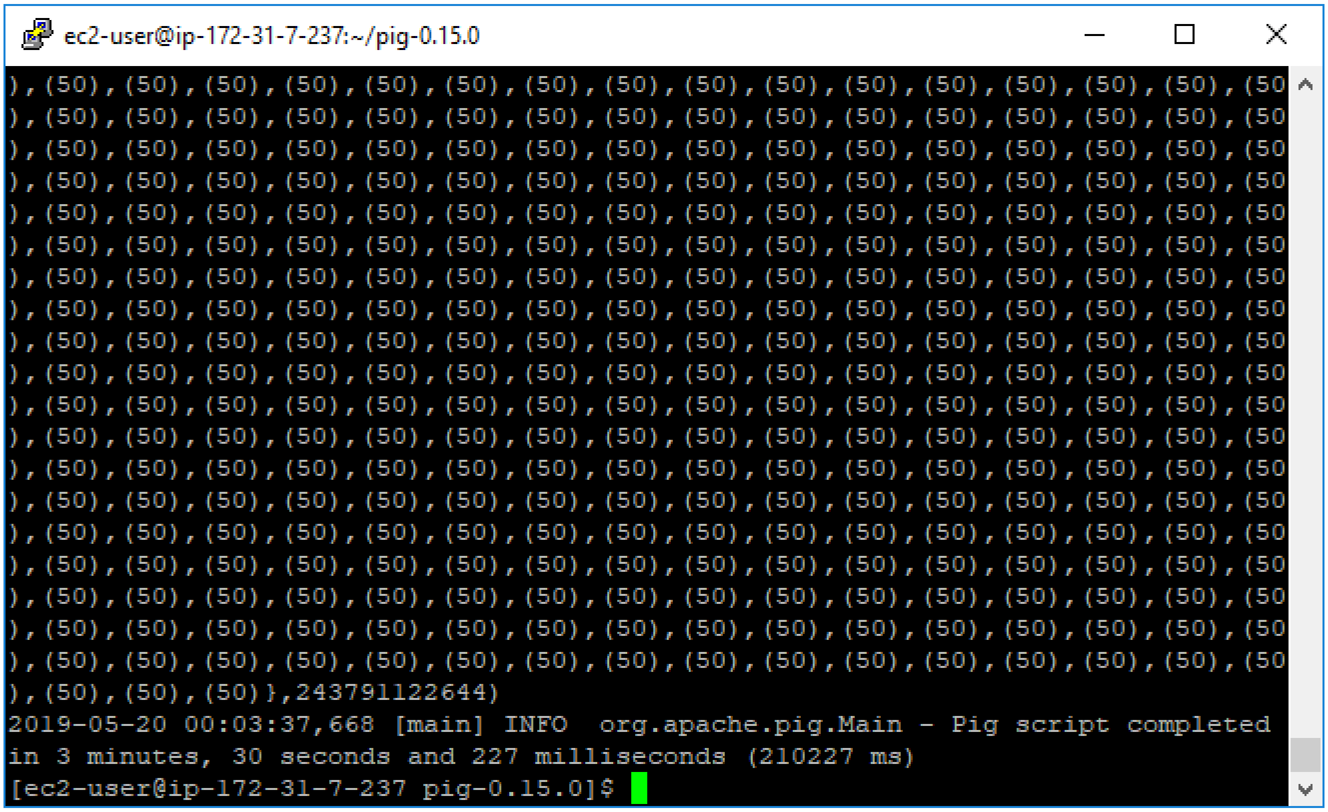
Query 2.1 time: 104.589 seconds



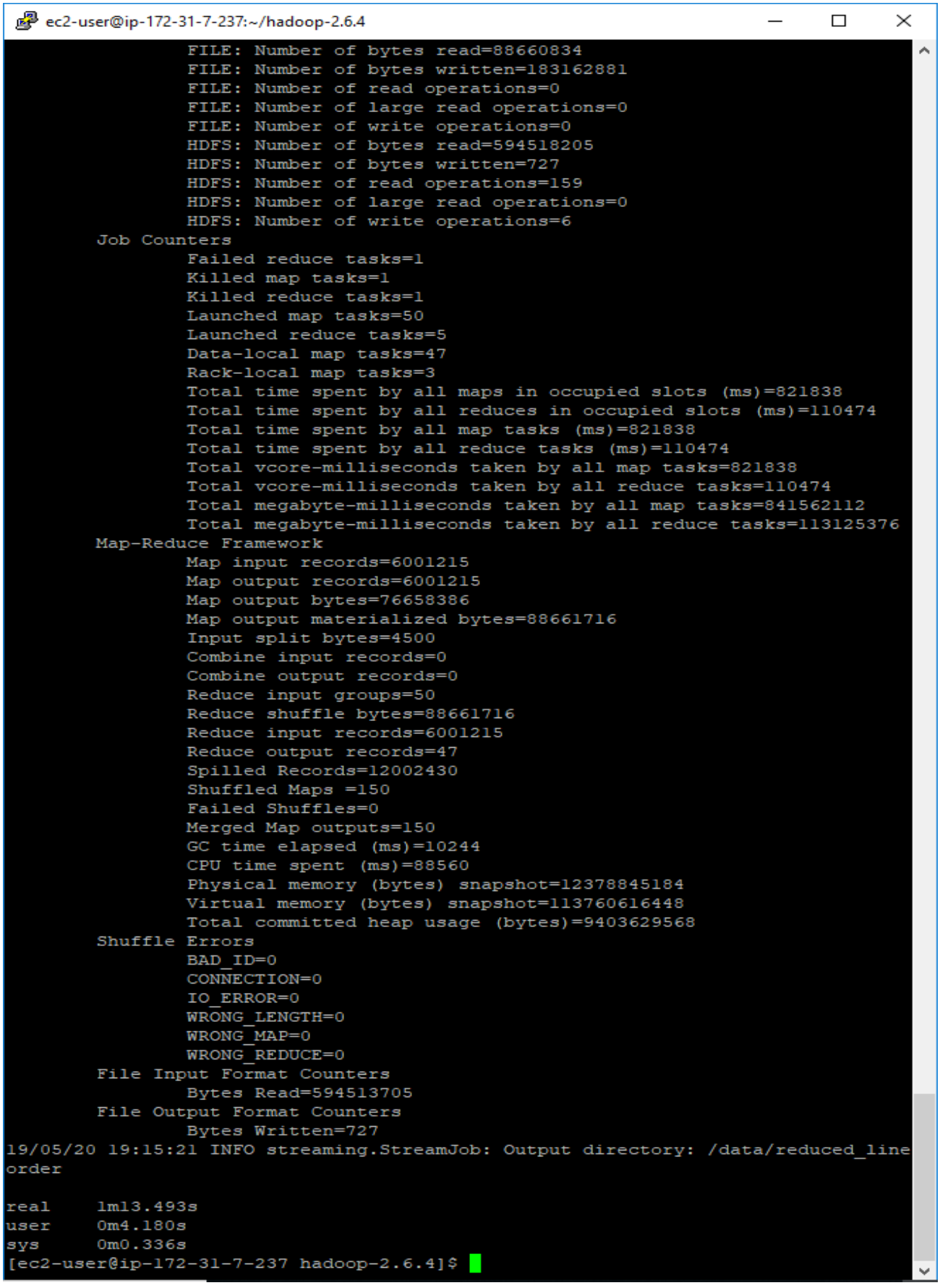
Select transform results:



Part 3:
Total Query 0.1-0.3 time: 3 minutes, 30 seconds and 225 milliseconds



Part 4:


ec2-user@ip-172-31-7-237:~/hadoop-2.6.4
FILE: Number of bytes read=88660834
FILE: Number of bytes written=183162881
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=594518205
HDFS: Number of bytes written=727
HDFS: Number of read operations=159
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
Job Counters
Failed reduce tasks=1
Killed map tasks=1
Killed reduce tasks=1
Launched map tasks=50
Launched reduce tasks=5
Data-local map tasks=47
Rack-local map tasks=3
Total time spent by all maps in occupied slots (ms)=821838
Total time spent by all reduces in occupied slots (ms)=110474
Total time spent by all map tasks (ms)=821838
Total time spent by all reduce tasks (ms)=110474
Total vcore-milliseconds taken by all map tasks=821838
Total vcore-milliseconds taken by all reduce tasks=110474
Total megabyte-milliseconds taken by all map tasks=841562112
Total megabyte-milliseconds taken by all reduce tasks=113125376
Map-Reduce Framework
Map input records=6001215
Map output records=6001215
Map output bytes=76658386
Map output materialized bytes=88661716
Input split bytes=4500
Combine input records=0
Combine output records=0
Reduce input groups=50
Reduce shuffle bytes=88661716
Reduce input records=6001215
Reduce output records=47
Spilled Records=12002430
Shuffled Maps =150
Failed Shuffles=0
Merged Map outputs=150
GC time elapsed (ms)=10244
CPU time spent (ms)=88560
Physical memory (bytes) snapshot=12378845184
Virtual memory (bytes) snapshot=113760616448
Total committed heap usage (bytes)=9403629568
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=594513705
File Output Format Counters
Bytes Written=727
19/05/20 19:15:21 INFO streaming.StreamJob: Output directory: /data/reduced_line
order
real 1m13.493s
user 0m4.180s
sys 0m0.336s
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$