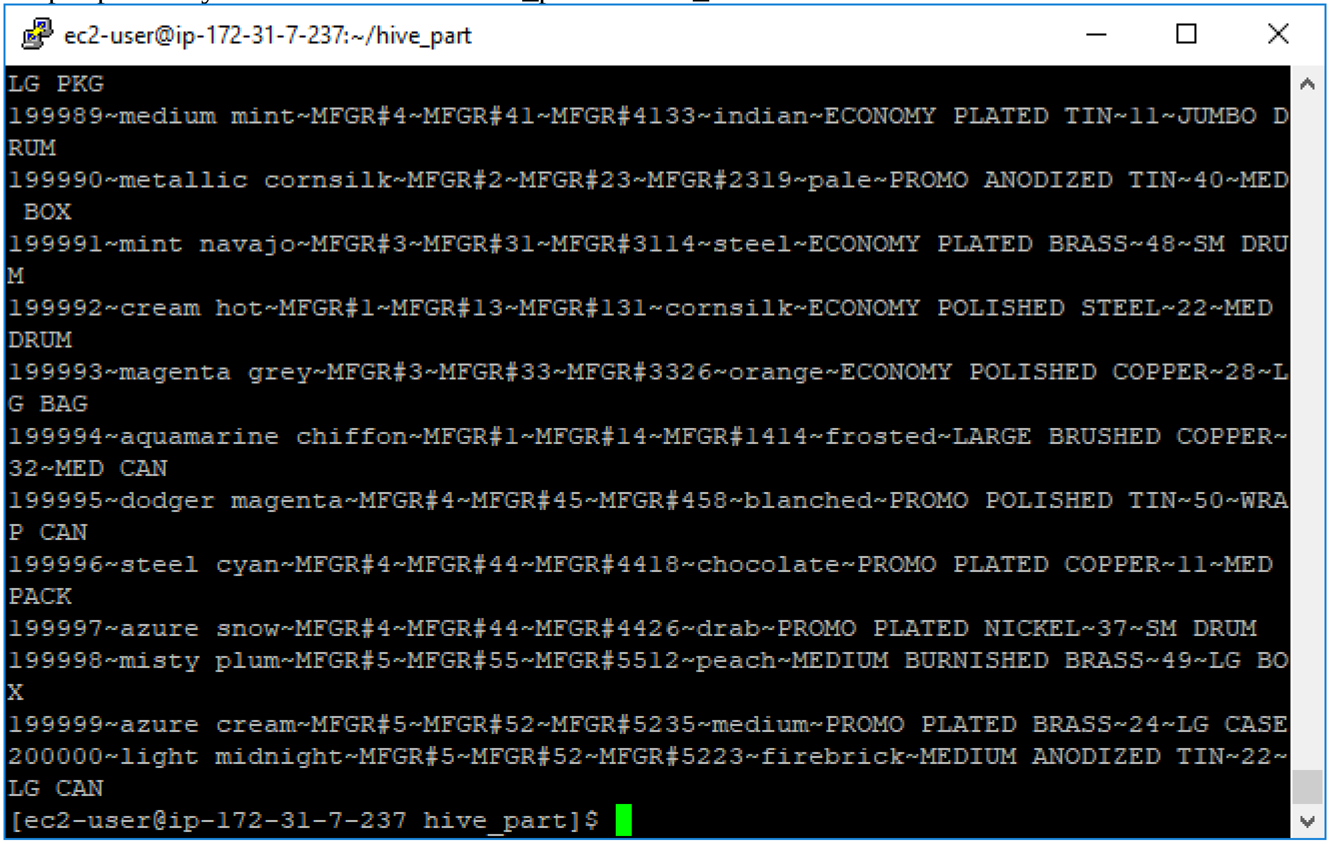
David Beck
CSC555
Project Phase 2
6/12/19

Note: I used the 4-node cluster for this assignment. Almost all commands are listed in the code files attached to this project.

Part 1:

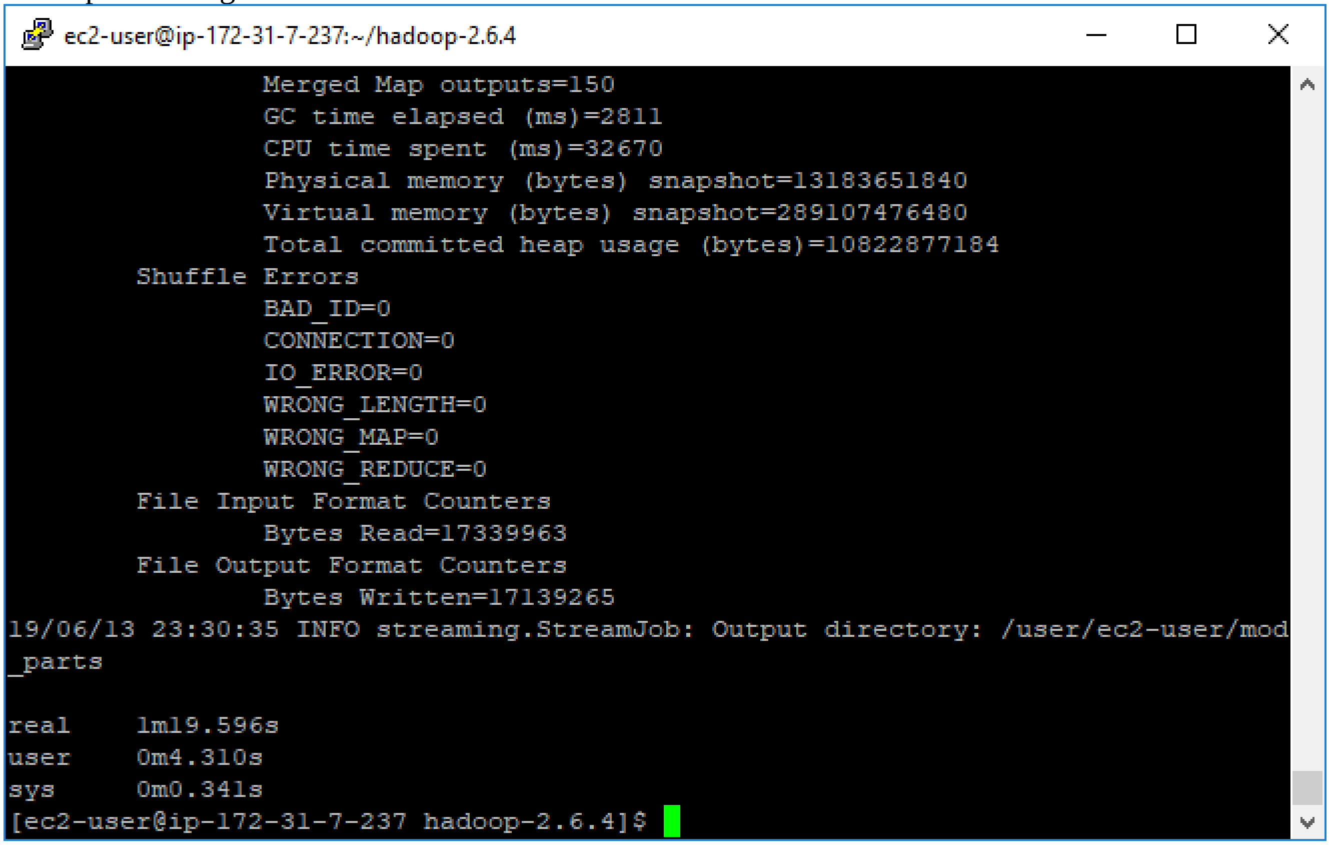
Hive:

Output pulled by this command: cat hive_part/000000_0



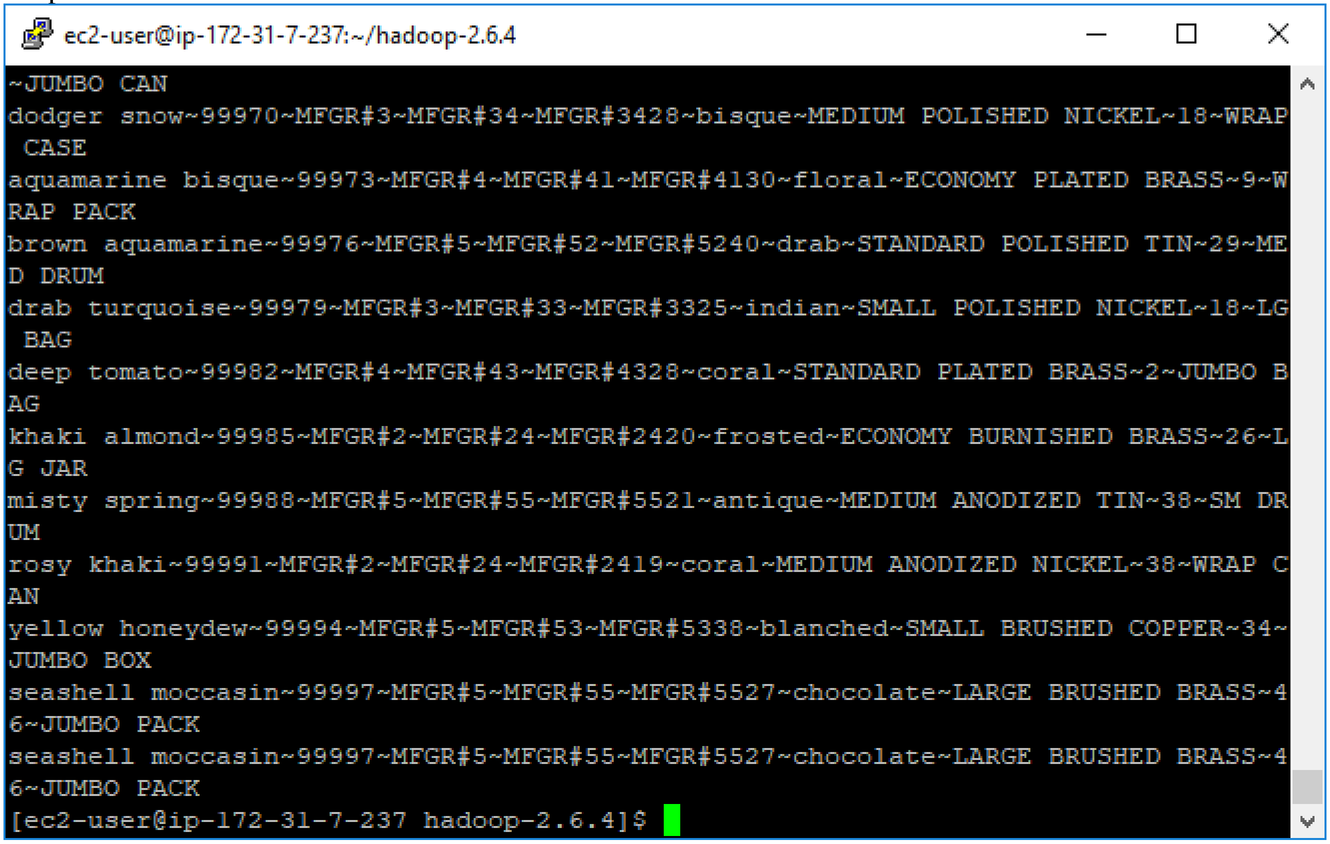
```
LG PKG
199989~medium mint~MFGR#4~MFGR#41~MFGR#4133~indian~ECONOMY PLATED TIN~11~JUMBO D
RUM
199990~metallic cornsilk~MFGR#2~MFGR#23~MFGR#2319~pale~PROMO ANODIZED TIN~40~MED
 BOX
199991~mint navajo~MFGR#3~MFGR#31~MFGR#3114~steel~ECONOMY PLATED BRASS~48~SM DRU
M
199992~cream hot~MFGR#1~MFGR#13~MFGR#131~cornsilk~ECONOMY POLISHED STEEL~22~MED
DRUM
199993~magenta grey~MFGR#3~MFGR#33~MFGR#3326~orange~ECONOMY POLISHED COPPER~28~L
G BAG
199994~aquamarine chiffon~MFGR#1~MFGR#14~MFGR#1414~frosted~LARGE BRUSHED COPPER~
32~MED CAN
199995~dodger magenta~MFGR#4~MFGR#45~MFGR#458~blanched~PROMO POLISHED TIN~50~WRA
P CAN
199996~steel cyan~MFGR#4~MFGR#44~MFGR#4418~chocolate~PROMO PLATED COPPER~11~MED
PACK
199997~azure snow~MFGR#4~MFGR#44~MFGR#4426~drab~PROMO PLATED NICKEL~37~SM DRUM
199998~misty plum~MFGR#5~MFGR#55~MFGR#5512~peach~MEDIUM BURNISHED BRASS~49~LG BO
X
199999~azure cream~MFGR#5~MFGR#52~MFGR#5235~medium~PROMO PLATED BRASS~24~LG CASE
200000~light midnight~MFGR#5~MFGR#52~MFGR#5223~firebrick~MEDIUM ANODIZED TIN~22~
LG CAN
[ec2-user@ip-172-31-7-237 hive_part]$
```

Hadoop streaming:



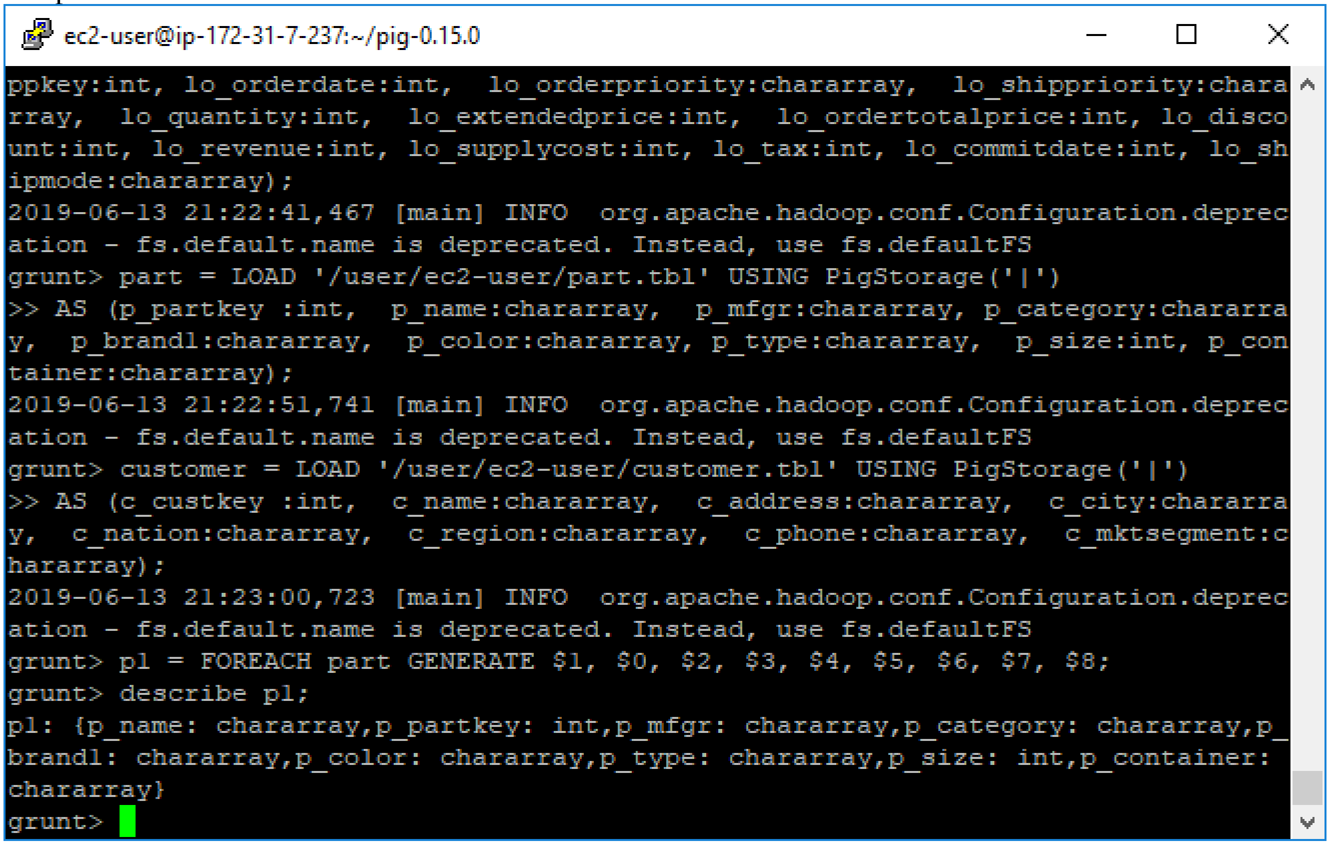
```
                Merged Map outputs=150
                GC time elapsed (ms)=2811
                CPU time spent (ms)=32670
                Physical memory (bytes) snapshot=13183651840
                Virtual memory (bytes) snapshot=289107476480
                Total committed heap usage (bytes)=10822877184
        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=17339963
        File Output Format Counters
                Bytes Written=17139265
19/06/13 23:30:35 INFO streaming.StreamJob: Output directory: /user/ec2-user/mod
_parts

real    1m19.596s
user    0m4.310s
sys     0m0.341s
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```

Output file:



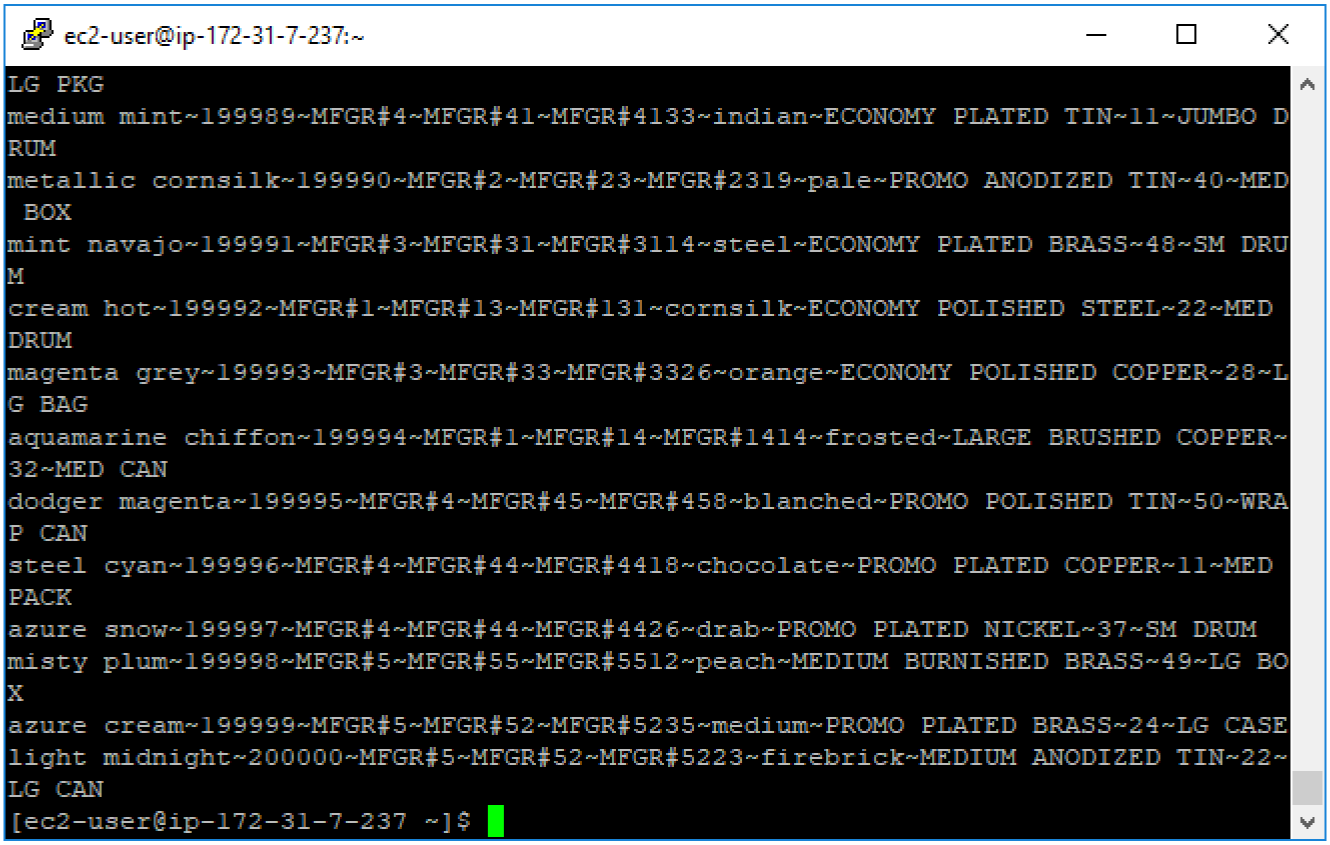
```
~JUMBO CAN
dodger snow~99970~MFGR#3~MFGR#34~MFGR#3428~bisque~MEDIUM POLISHED NICKEL~18~WRAP
 CASE
aquamarine bisque~99973~MFGR#4~MFGR#41~MFGR#4130~floral~ECONOMY PLATED BRASS~9~W
RAP PACK
brown aquamarine~99976~MFGR#5~MFGR#52~MFGR#5240~drab~STANDARD POLISHED TIN~29~ME
D DRUM
drab turquoise~99979~MFGR#3~MFGR#33~MFGR#3325~indian~SMALL POLISHED NICKEL~18~LG
 BAG
deep tomato~99982~MFGR#4~MFGR#43~MFGR#4328~coral~STANDARD PLATED BRASS~2~JUMBO B
AG
khaki almond~99985~MFGR#2~MFGR#24~MFGR#2420~frosted~ECONOMY BURNISHED BRASS~26~L
G JAR
misty spring~99988~MFGR#5~MFGR#55~MFGR#5521~antique~MEDIUM ANODIZED TIN~38~SM DR
UM
rosy khaki~99991~MFGR#2~MFGR#24~MFGR#2419~coral~MEDIUM ANODIZED NICKEL~38~WRAP C
AN
yellow honeydew~99994~MFGR#5~MFGR#53~MFGR#5338~blanched~SMALL BRUSHED COPPER~34~
JUMBO BOX
seashell moccasin~99997~MFGR#5~MFGR#55~MFGR#5527~chocolate~LARGE BRUSHED BRASS~4
6~JUMBO PACK
seashell moccasin~99997~MFGR#5~MFGR#55~MFGR#5527~chocolate~LARGE BRUSHED BRASS~4
6~JUMBO PACK
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```

Pig:
Swap columns:



```
ppkey:int, lo_orderdate:int,  lo_orderpriority:chararray,  lo_shippriority:chara
rray,  lo_quantity:int,  lo_extendedprice:int,  lo_ordertotalprice:int, lo_disco
unt:int, lo_revenue:int, lo_supplycost:int, lo_tax:int, lo_commitdate:int, lo_sh
ipmode:chararray);
2019-06-13 21:22:41,467 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> part = LOAD '/user/ec2-user/part.tbl' USING PigStorage('|')
>> AS (p_partkey :int,  p_name:chararray,  p_mfgr:chararray, p_category:chararra
y,  p_brandl:chararray,  p_color:chararray, p_type:chararray,  p_size:int, p_con
tainer:chararray);
2019-06-13 21:22:51,741 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> customer = LOAD '/user/ec2-user/customer.tbl' USING PigStorage('|')
>> AS (c_custkey :int,  c_name:chararray,  c_address:chararray,  c_city:chararra
y,  c_nation:chararray,  c_region:chararray,  c_phone:chararray,  c_mktsegment:c
hararray);
2019-06-13 21:23:00,723 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> pl = FOREACH part GENERATE $1, $0, $2, $3, $4, $5, $6, $7, $8;
grunt> describe pl;
pl: {p_name: chararray,p_partkey: int,p_mfgr: chararray,p_category: chararray,p_
brandl: chararray,p_color: chararray,p_type: chararray,p_size: int,p_container:
chararray}
grunt>
```

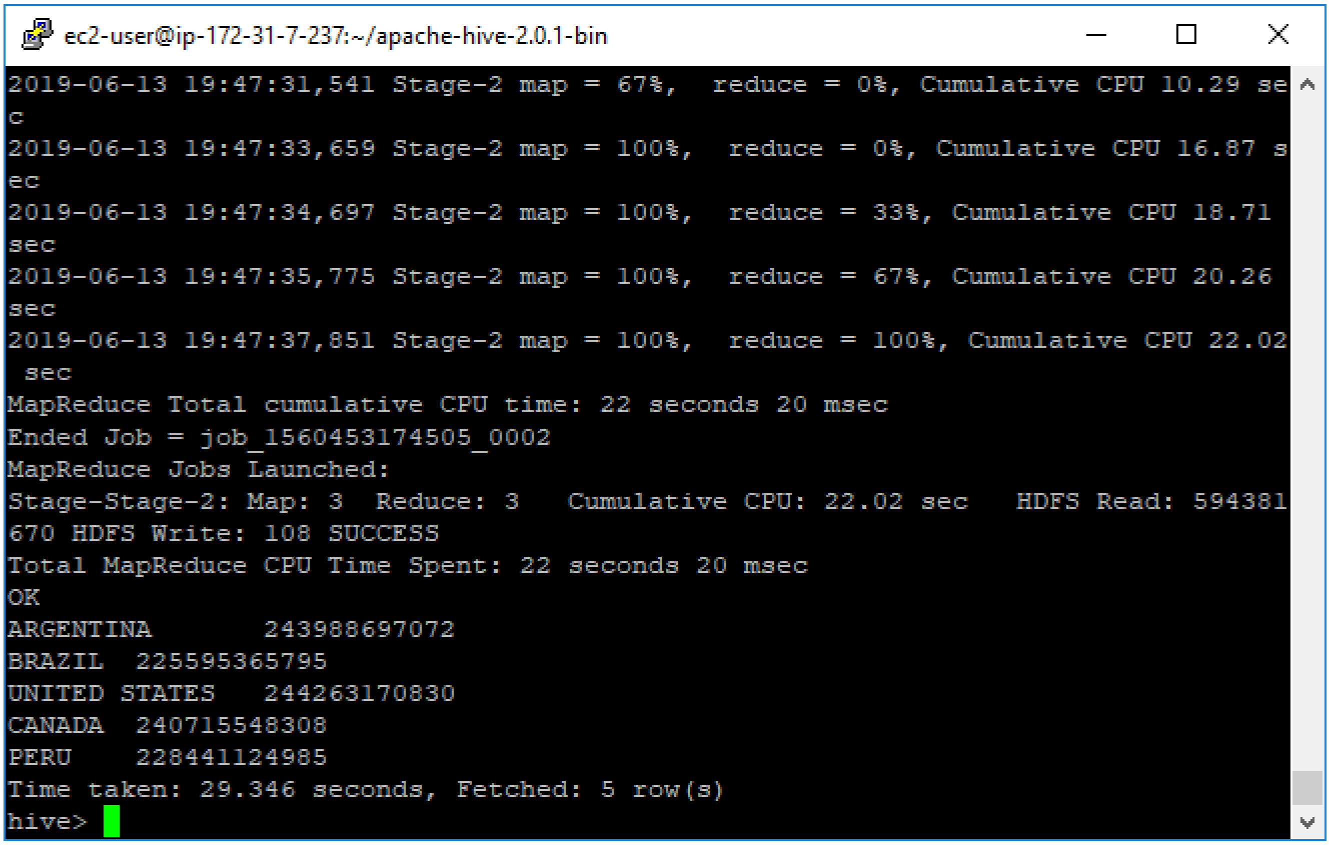
Output: hdfs dfs -cat /user/ec2-user/pig_part.txt/part-m-00000



```
LG PKG
medium mint~199989~MFGR#4~MFGR#41~MFGR#4133~indian~ECONOMY PLATED TIN~11~JUMBO D
RUM
metallic cornsilk~199990~MFGR#2~MFGR#23~MFGR#2319~pale~PROMO ANODIZED TIN~40~MED
 BOX
mint navajo~199991~MFGR#3~MFGR#31~MFGR#3114~steel~ECONOMY PLATED BRASS~48~SM DRU
M
cream hot~199992~MFGR#1~MFGR#13~MFGR#131~cornsilk~ECONOMY POLISHED STEEL~22~MED
DRUM
magenta grey~199993~MFGR#3~MFGR#33~MFGR#3326~orange~ECONOMY POLISHED COPPER~28~L
G BAG
aquamarine chiffon~199994~MFGR#1~MFGR#14~MFGR#1414~frosted~LARGE BRUSHED COPPER~
32~MED CAN
dodger magenta~199995~MFGR#4~MFGR#45~MFGR#458~blanched~PROMO POLISHED TIN~50~WRA
P CAN
steel cyan~199996~MFGR#4~MFGR#44~MFGR#4418~chocolate~PROMO PLATED COPPER~11~MED
PACK
azure snow~199997~MFGR#4~MFGR#44~MFGR#4426~drab~PROMO PLATED NICKEL~37~SM DRUM
misty plum~199998~MFGR#5~MFGR#55~MFGR#5512~peach~MEDIUM BURNISHED BRASS~49~LG BO
X
azure cream~199999~MFGR#5~MFGR#52~MFGR#5235~medium~PROMO PLATED BRASS~24~LG CASE
light midnight~200000~MFGR#5~MFGR#52~MFGR#5223~firebrick~MEDIUM ANODIZED TIN~22~
LG CAN
[ec2-user@ip-172-31-7-237 ~]$
```

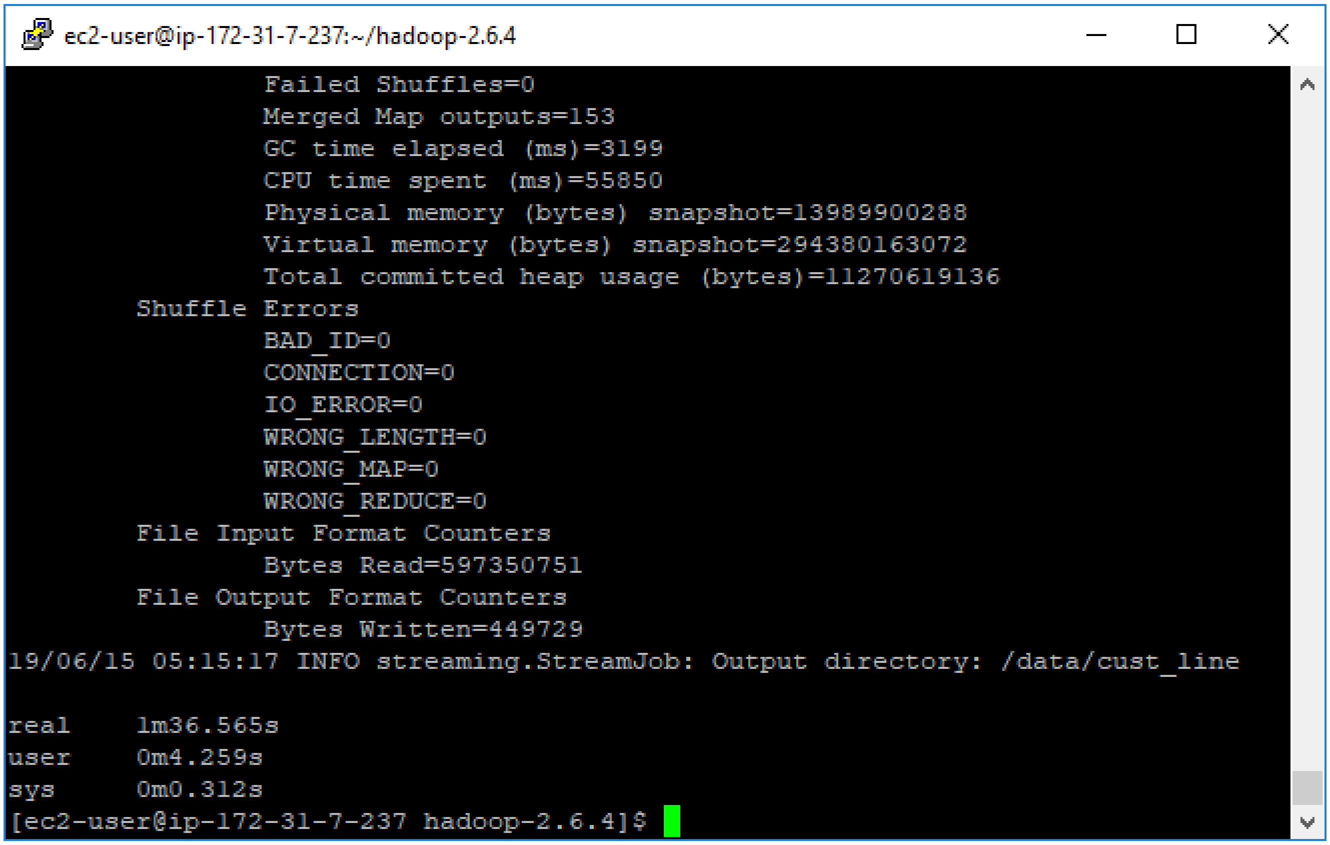
Part 2:

Hive: time 29.346 seconds



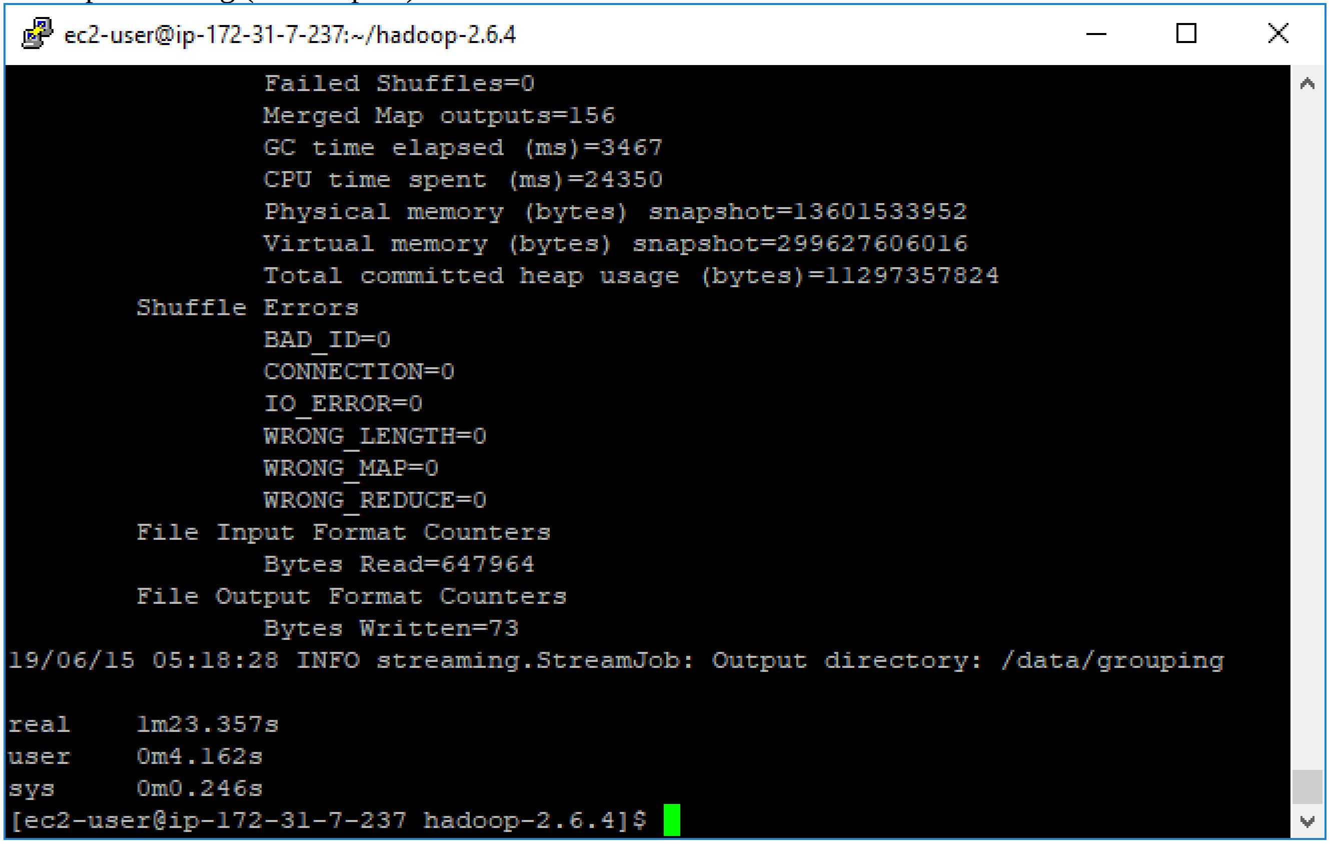
```
2019-06-13 19:47:31,541 Stage-2 map = 67%,  reduce = 0%, Cumulative CPU 10.29 se
c
2019-06-13 19:47:33,659 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 16.87 s
ec
2019-06-13 19:47:34,697 Stage-2 map = 100%,  reduce = 33%, Cumulative CPU 18.71
sec
2019-06-13 19:47:35,775 Stage-2 map = 100%,  reduce = 67%, Cumulative CPU 20.26
sec
2019-06-13 19:47:37,851 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 22.02
 sec
MapReduce Total cumulative CPU time: 22 seconds 20 msec
Ended Job = job_1560453174505_0002
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3  Reduce: 3   Cumulative CPU: 22.02 sec   HDFS Read: 594381
670 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 22 seconds 20 msec
OK
ARGENTINA       243988697072
BRAZIL  225595365795
UNITED STATES   244263170830
CANADA  240715548308
PERU    228441124985
Time taken: 29.346 seconds, Fetched: 5 row(s)
hive>
```

Hadoop streaming (first pass): time 1 minute 36.565 seconds



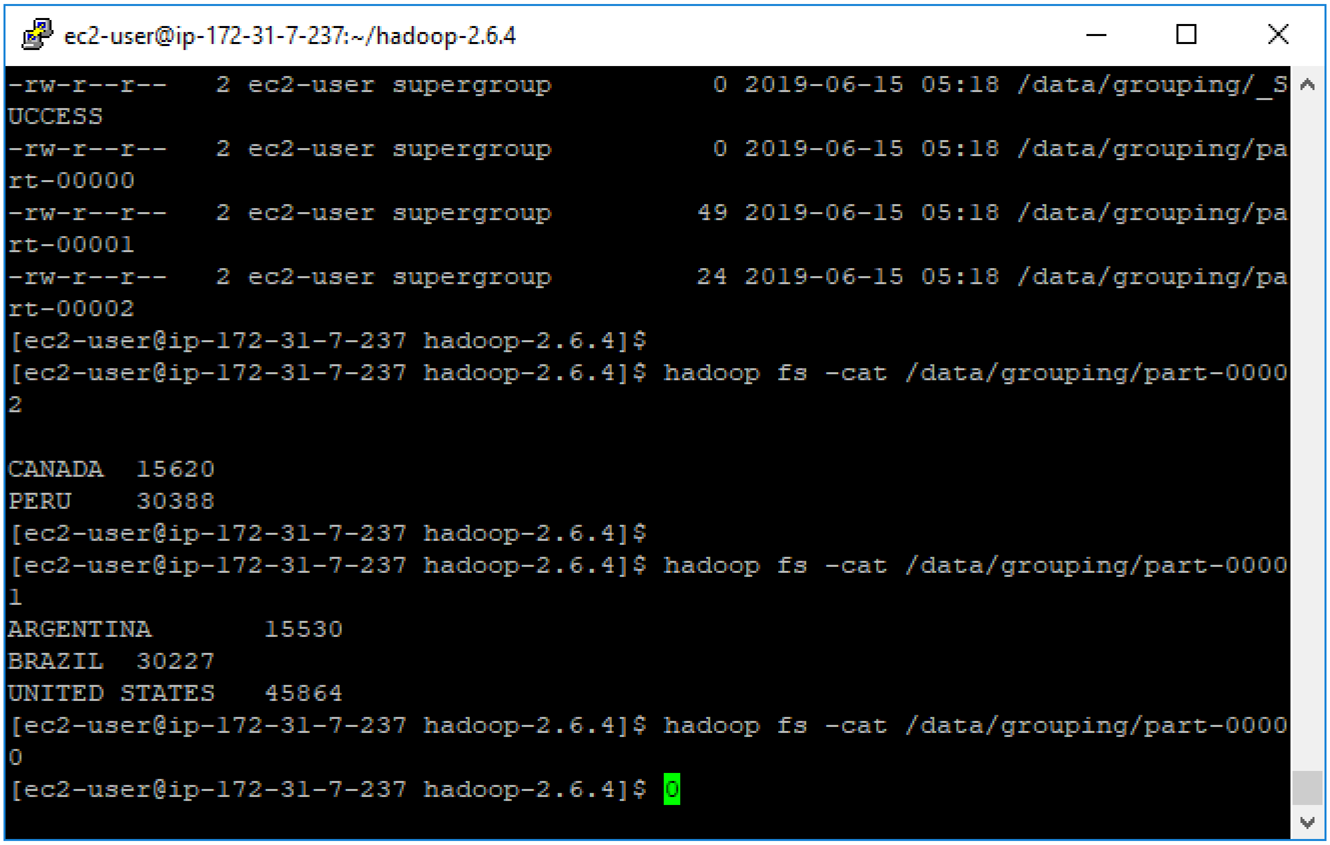
```
                Failed Shuffles=0
                Merged Map outputs=153
                GC time elapsed (ms)=3199
                CPU time spent (ms)=55850
                Physical memory (bytes) snapshot=13989900288
                Virtual memory (bytes) snapshot=294380163072
                Total committed heap usage (bytes)=11270619136
        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=597350751
        File Output Format Counters
                Bytes Written=449729
19/06/15 05:15:17 INFO streaming.StreamJob: Output directory: /data/cust_line

real    1m36.565s
user    0m4.259s
sys     0m0.312s
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```

Hadoop streaming (second pass): time 1 minutes 23.357 seconds



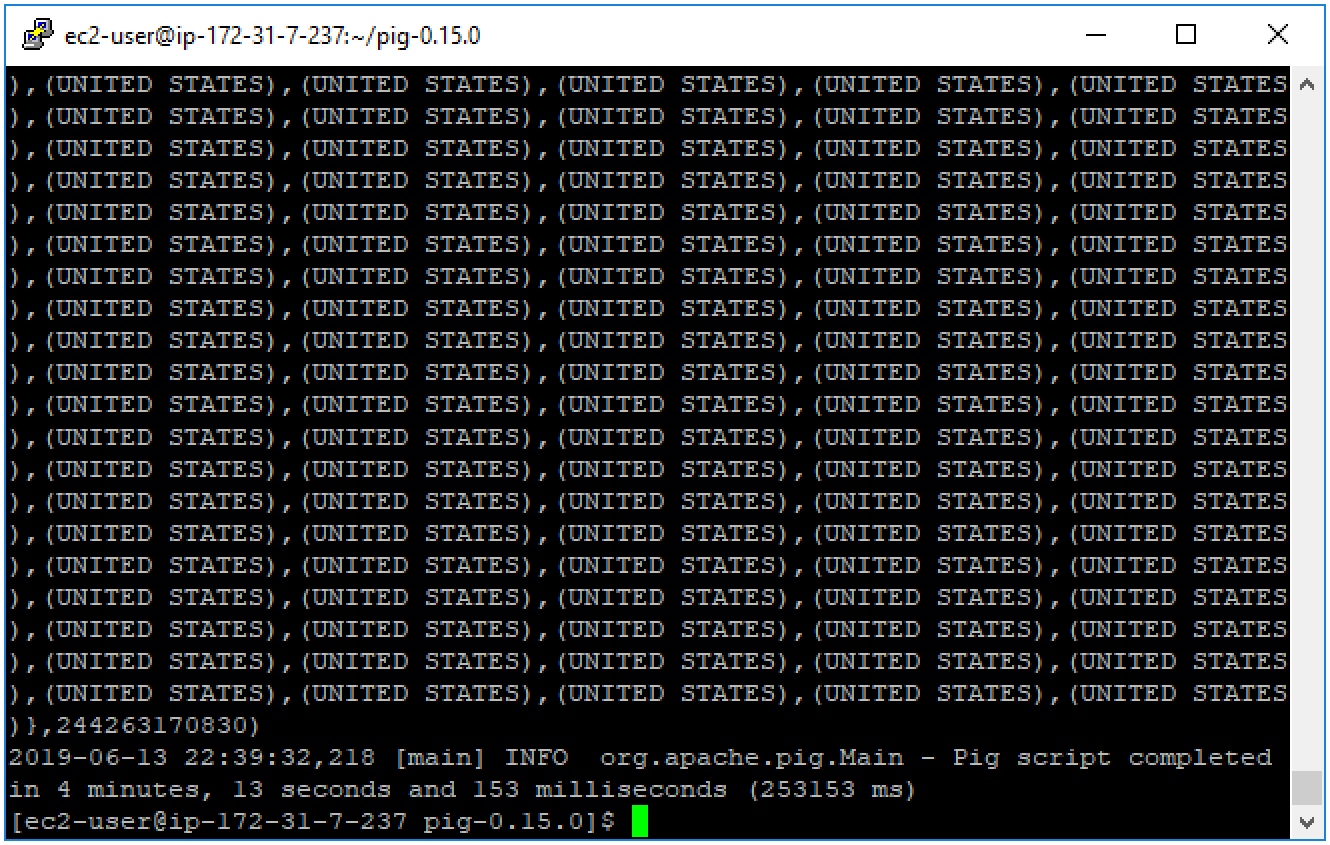
```
                Failed Shuffles=0
                Merged Map outputs=156
                GC time elapsed (ms)=3467
                CPU time spent (ms)=24350
                Physical memory (bytes) snapshot=13601533952
                Virtual memory (bytes) snapshot=299627606016
                Total committed heap usage (bytes)=11297357824
        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=647964
        File Output Format Counters
                Bytes Written=73
19/06/15 05:18:28 INFO streaming.StreamJob: Output directory: /data/grouping

real    1m23.357s
user    0m4.162s
sys     0m0.246s
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```

Output:



```
-rw-r--r--   2 ec2-user supergroup          0 2019-06-15 05:18 /data/grouping/_S
UCCESS
-rw-r--r--   2 ec2-user supergroup          0 2019-06-15 05:18 /data/grouping/pa
rt-00000
-rw-r--r--   2 ec2-user supergroup         49 2019-06-15 05:18 /data/grouping/pa
rt-00001
-rw-r--r--   2 ec2-user supergroup         24 2019-06-15 05:18 /data/grouping/pa
rt-00002
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -cat /data/grouping/part-0000
2

CANADA  15620
PERU    30388
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -cat /data/grouping/part-0000
1
ARGENTINA       15530
BRAZIL  30227
UNITED STATES   45864
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -cat /data/grouping/part-0000
0
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ 0
```

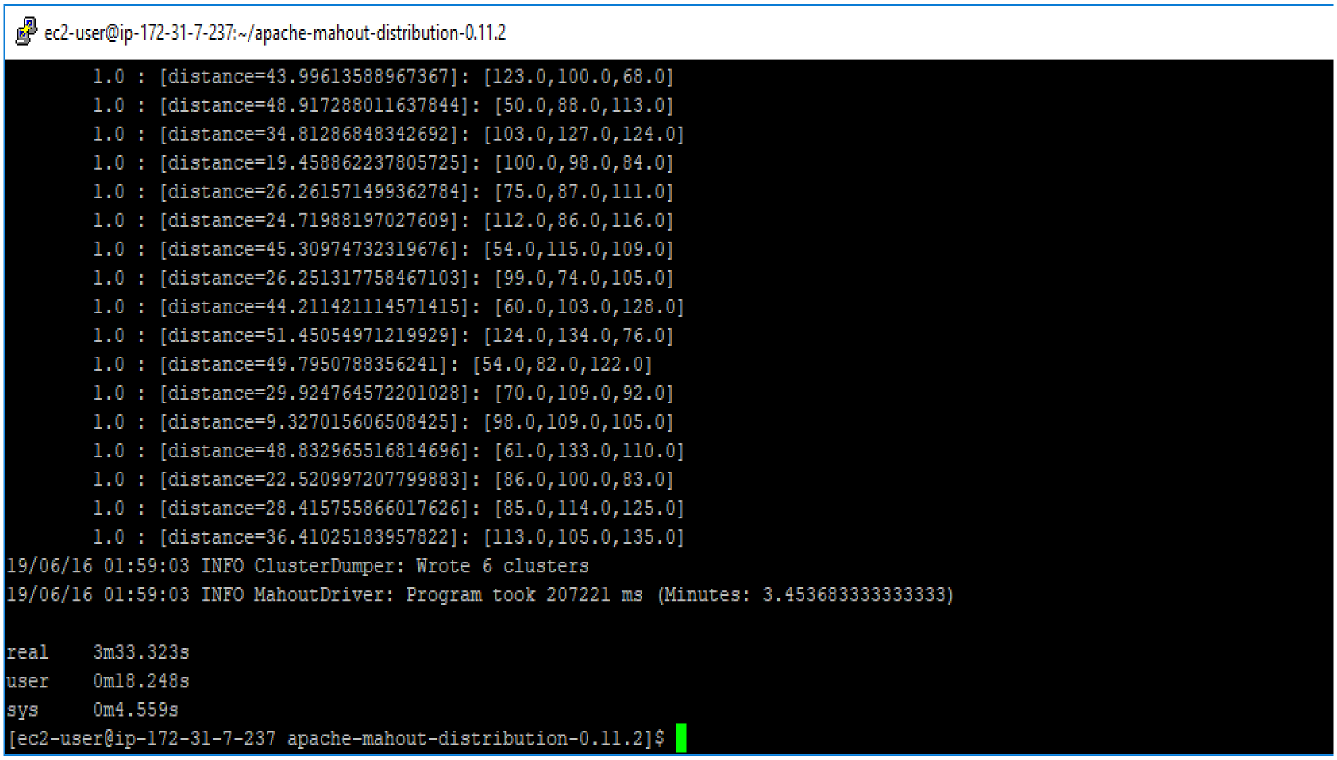
Pig: time 4 minutes, 13.153 seconds



```
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES),(UNITED STATES
)},244263170830)
2019-06-13 22:39:32,218 [main] INFO  org.apache.pig.Main - Pig script completed
in 4 minutes, 13 seconds and 153 milliseconds (253153 ms)
[ec2-user@ip-172-31-7-237 pig-0.15.0]$
```

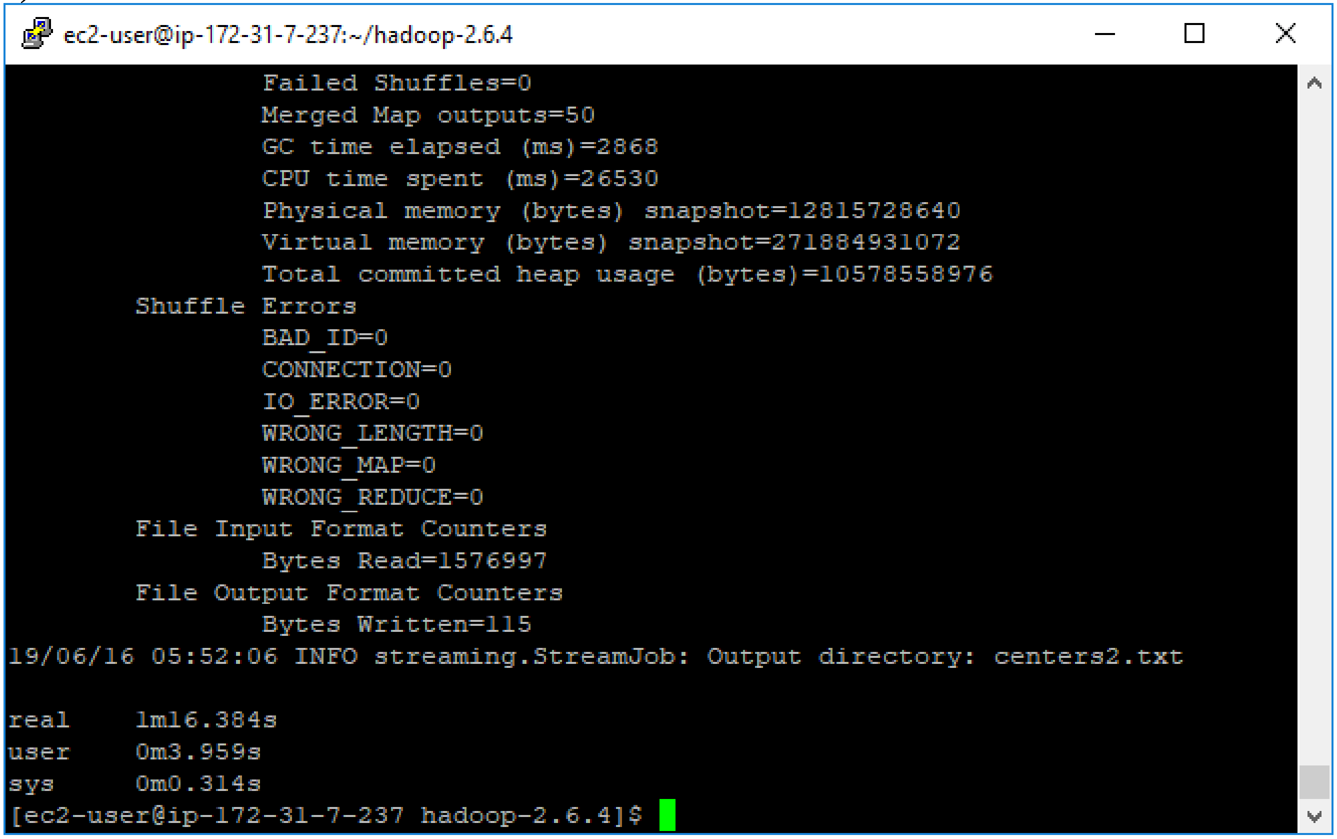
Part 3:

a) Mahout run on sample data created. Time: 3 minutes 33.323 seconds


ec2-user@ip-172-31-7-237:~/apache-mahout-distribution-0.11.2

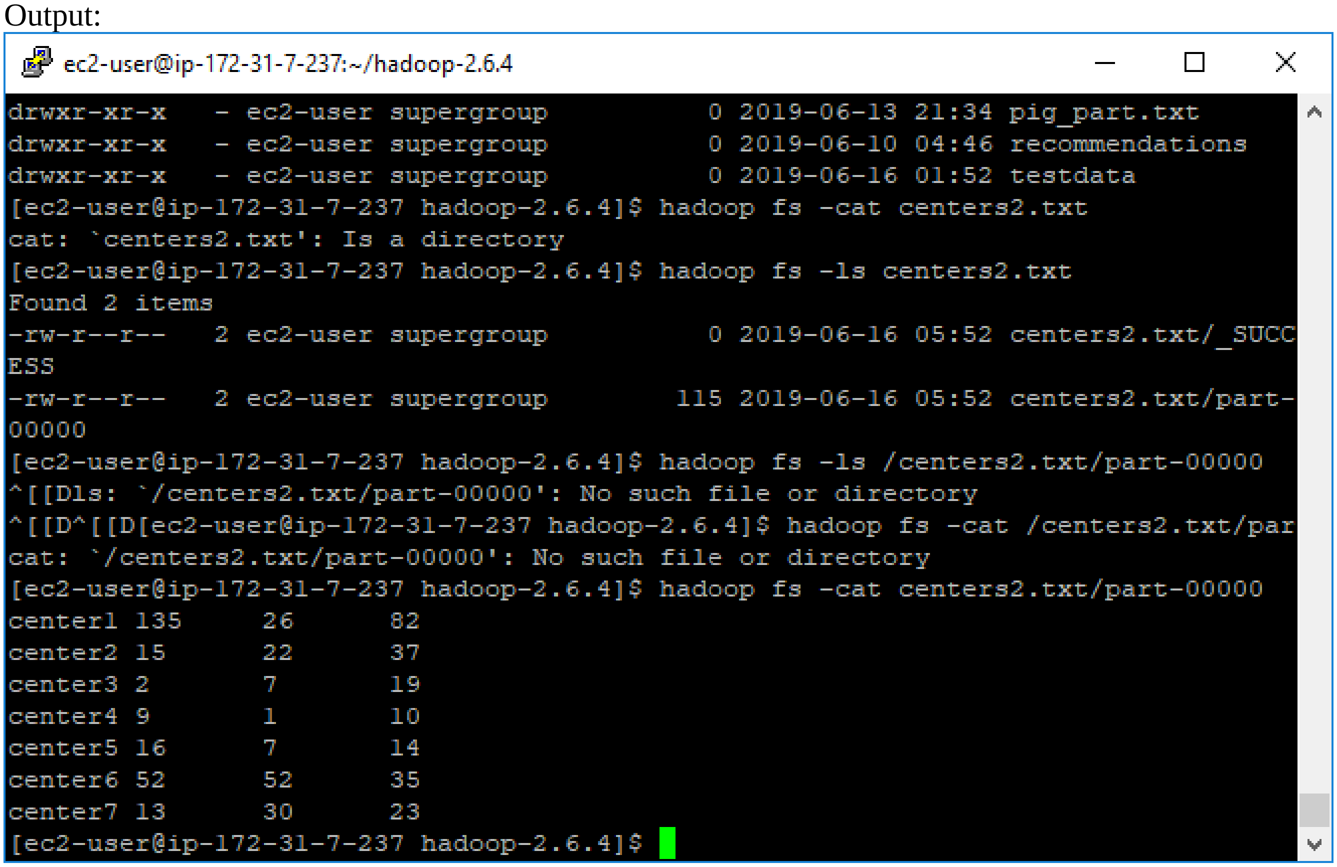
```
        1.0 : [distance=43.99613588967367]: [123.0,100.0,68.0]
        1.0 : [distance=48.917288011637844]: [50.0,88.0,113.0]
        1.0 : [distance=34.81286848342692]: [103.0,127.0,124.0]
        1.0 : [distance=19.458862237805725]: [100.0,98.0,84.0]
        1.0 : [distance=26.261571499362784]: [75.0,87.0,111.0]
        1.0 : [distance=24.71988197027609]: [112.0,86.0,116.0]
        1.0 : [distance=45.30974732319676]: [54.0,115.0,109.0]
        1.0 : [distance=26.251317758467103]: [99.0,74.0,105.0]
        1.0 : [distance=44.211421114571415]: [60.0,103.0,128.0]
        1.0 : [distance=51.45054971219929]: [124.0,134.0,76.0]
        1.0 : [distance=49.7950788356241]: [54.0,82.0,122.0]
        1.0 : [distance=29.924764572201028]: [70.0,109.0,92.0]
        1.0 : [distance=9.327015606508425]: [98.0,109.0,105.0]
        1.0 : [distance=48.832965516814696]: [61.0,133.0,110.0]
        1.0 : [distance=22.520997207799883]: [86.0,100.0,83.0]
        1.0 : [distance=28.415755866017626]: [85.0,114.0,125.0]
        1.0 : [distance=36.41025183957822]: [113.0,105.0,135.0]
19/06/16 01:59:03 INFO ClusterDumper: Wrote 6 clusters
19/06/16 01:59:03 INFO MahoutDriver: Program took 207221 ms (Minutes: 3.4536833333333333)

real    3m33.323s
user    0m18.248s
sys     0m4.559s
[ec2-user@ip-172-31-7-237 apache-mahout-distribution-0.11.2]$
```

b) First Pass: time 1 minute 16.384 seconds



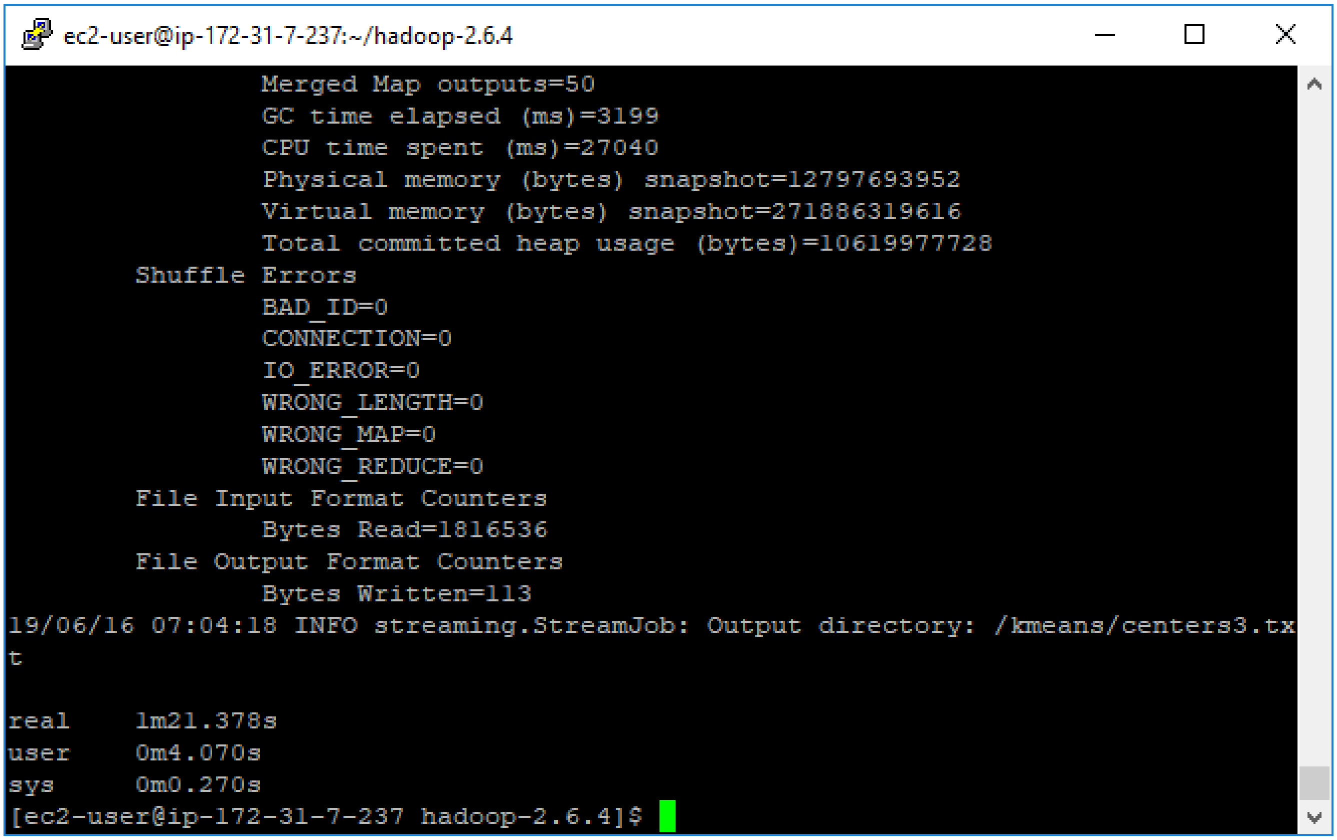
```
                Failed Shuffles=0
                Merged Map outputs=50
                GC time elapsed (ms)=2868
                CPU time spent (ms)=26530
                Physical memory (bytes) snapshot=12815728640
                Virtual memory (bytes) snapshot=271884931072
                Total committed heap usage (bytes)=10578558976
        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=1576997
        File Output Format Counters
                Bytes Written=115
19/06/16 05:52:06 INFO streaming.StreamJob: Output directory: centers2.txt

real    1m16.384s
user    0m3.959s
sys     0m0.314s
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```

Output:



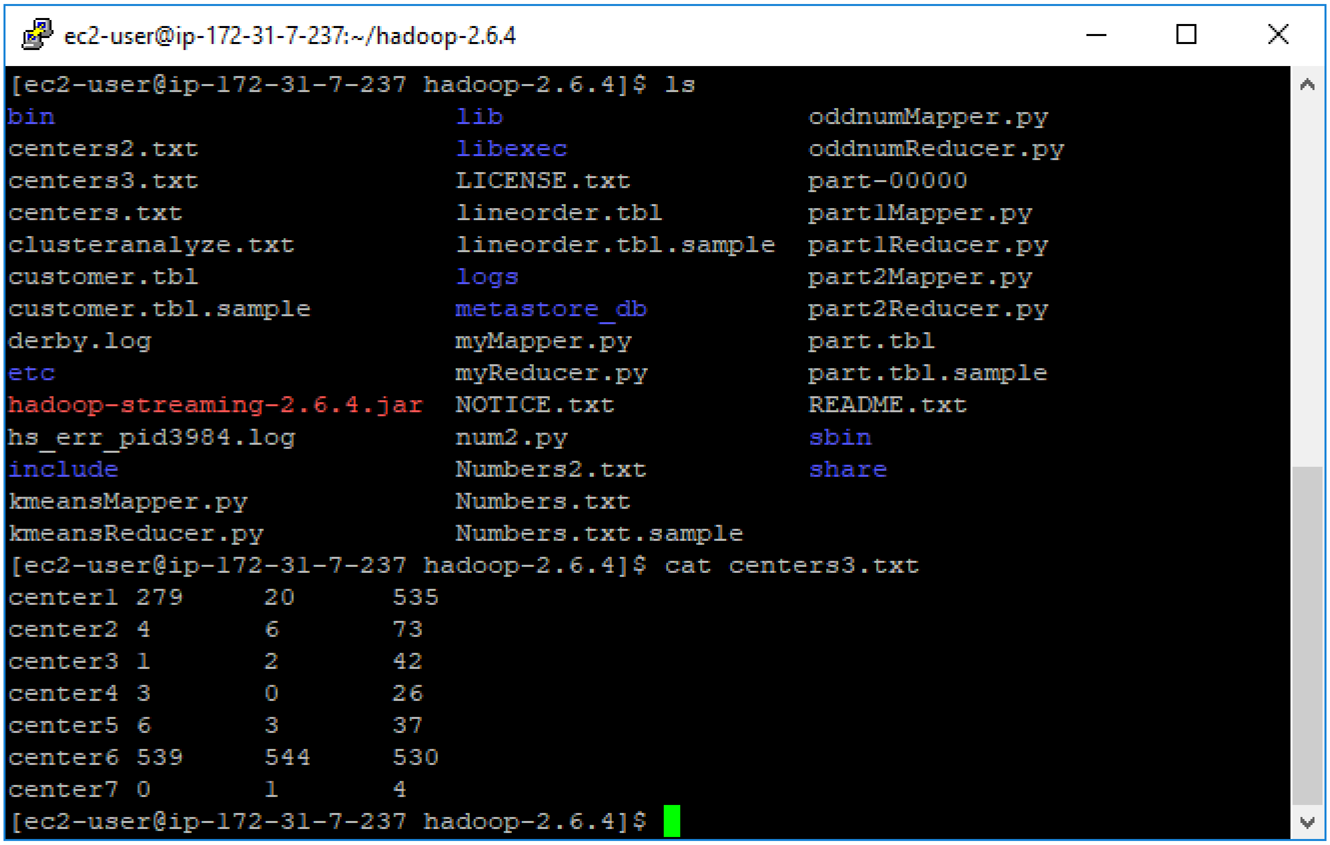
```
drwxr-xr-x   - ec2-user supergroup          0 2019-06-13 21:34 pig_part.txt
drwxr-xr-x   - ec2-user supergroup          0 2019-06-10 04:46 recommendations
drwxr-xr-x   - ec2-user supergroup          0 2019-06-16 01:52 testdata
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -cat centers2.txt
cat: `centers2.txt': Is a directory
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -ls centers2.txt
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2019-06-16 05:52 centers2.txt/_SUCC
ESS
-rw-r--r--   2 ec2-user supergroup        115 2019-06-16 05:52 centers2.txt/part-
00000
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -ls /centers2.txt/part-00000
^[[Dls: `/centers2.txt/part-00000': No such file or directory
^[[D^[[D[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -cat /centers2.txt/par
cat: `/centers2.txt/part-00000': No such file or directory
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ hadoop fs -cat centers2.txt/part-00000
center1 135     26      82
center2 15      22      37
center3 2       7       19
center4 9       1       10
center5 16      7       14
center6 52      52      35
center7 13      30      23
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```

Use command to copy centers2.txt to local:
hadoop fs -copyToLocal centers2.txt/part-00000 /home/ec2-user/hadoop-2.6.4/centers2.txt
Second Pass: time 1 minute 21.378 seconds



```
                Merged Map outputs=50
                GC time elapsed (ms)=3199
                CPU time spent (ms)=27040
                Physical memory (bytes) snapshot=12797693952
                Virtual memory (bytes) snapshot=271886319616
                Total committed heap usage (bytes)=10619977728
        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=1816536
        File Output Format Counters
                Bytes Written=113
19/06/16 07:04:18 INFO streaming.StreamJob: Output directory: /kmeans/centers3.tx
t

real    1m21.378s
user    0m4.070s
sys     0m0.270s
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```

Output:



```
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ ls
bin                         lib                   oddnumMapper.py
centers2.txt                libexec               oddnumReducer.py
centers3.txt                LICENSE.txt           part-00000
centers.txt                 lineorder.tbl         part1Mapper.py
clusteranalyze.txt          lineorder.tbl.sample  part1Reducer.py
customer.tbl                logs                  part2Mapper.py
customer.tbl.sample         metastore_db          part2Reducer.py
derby.log                   myMapper.py           part.tbl
etc                         myReducer.py          part.tbl.sample
hadoop-streaming-2.6.4.jar  NOTICE.txt            README.txt
hs_err_pid3984.log          num2.py               sbin
include                     Numbers2.txt          share
kmeansMapper.py             Numbers.txt
kmeansReducer.py            Numbers.txt.sample
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$ cat centers3.txt
center1 279     20      535
center2 4       6       73
center3 1       2       42
center4 3       0       26
center5 6       3       37
center6 539     544     530
center7 0       1       4
[ec2-user@ip-172-31-7-237 hadoop-2.6.4]$
```