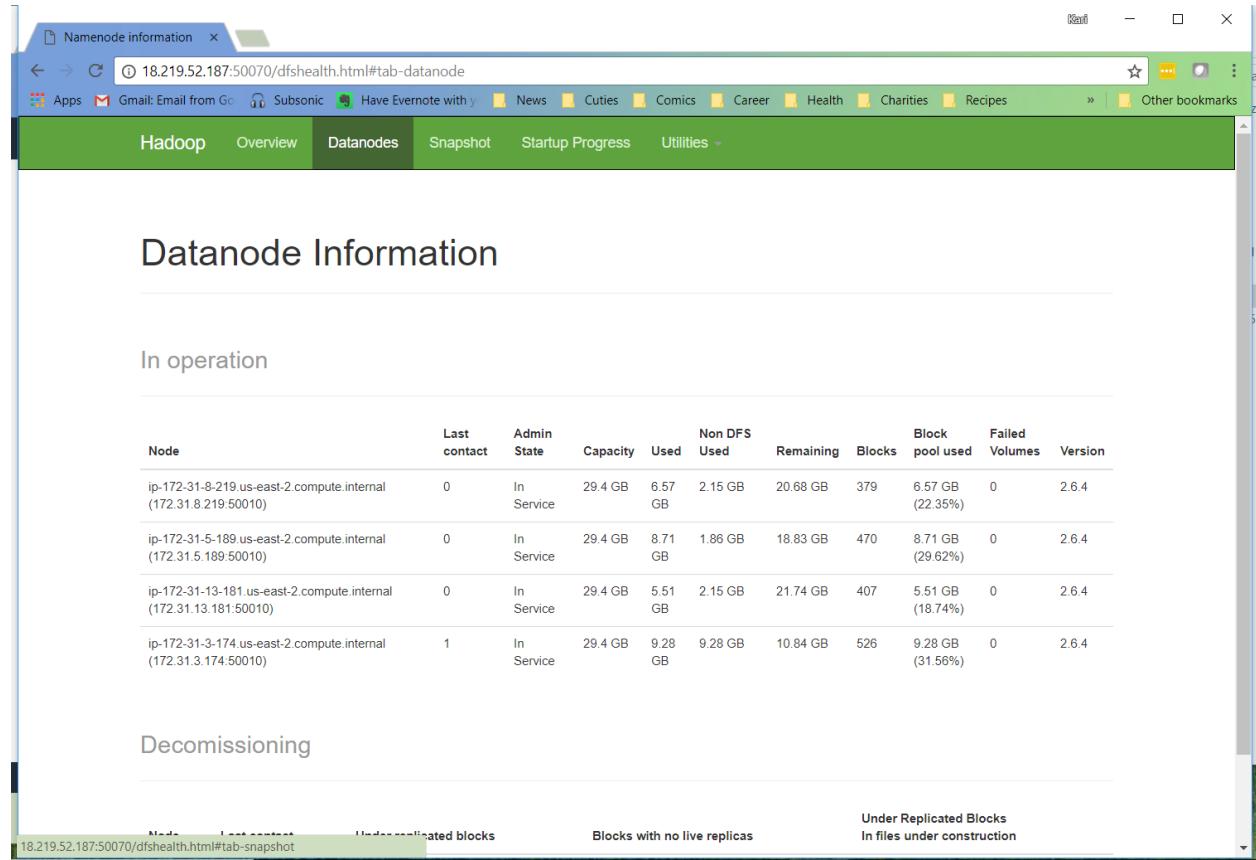


Kari Palmier
CSC 555 Winter 2018
Project Phase 2

I used my 4-node cluster from assignment 5 to perform all of the steps for parts 1 through 3. I used wget to get all of the scale 4 tables from the website given in the project instructions. Before doing this, I deleted any existing tables from the assignments or the project phase 1. I copied the scale 4 tables into HDFS using the hadoop fs -put filename /user/ec2-user command. All instances in the cluster were t2 medium instances with 30 GB of hard drive space each. The HDFS replication factor was set to 2 in the cluster.

4-node cluster screenshot:



The screenshot shows a web browser window titled "Namenode information" with the URL "18.219.52.187:50070/dfshealth.html#tab-datanode". The browser interface includes a top navigation bar with links like "Home", "New tab", "History", "Downloads", "Bookmarks", "Search", and "Help". Below the address bar is a toolbar with icons for "Apps", "Gmail: Email from Google", "Subsonic", "Have Evernote with you", "News", "Cutes", "Comics", "Career", "Health", "Charities", "Recipes", and "Other bookmarks". The main content area has a green header bar with tabs: "Hadoop", "Overview", "Datanodes" (which is selected), "Snapshot", "Startup Progress", and "Utilities". The main content is titled "Datanode Information" and displays two sections: "In operation" and "Decommissioning".

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
ip-172-31-8-219.us-east-2.compute.internal (172.31.8.219:50010)	0	In Service	29.4 GB	6.57 GB	2.15 GB	20.68 GB	379	6.57 GB (22.35%)	0	2.6.4
ip-172-31-5-189.us-east-2.compute.internal (172.31.5.189:50010)	0	In Service	29.4 GB	8.71 GB	1.86 GB	18.83 GB	470	8.71 GB (29.62%)	0	2.6.4
ip-172-31-13-181.us-east-2.compute.internal (172.31.13.181:50010)	0	In Service	29.4 GB	5.51 GB	2.15 GB	21.74 GB	407	5.51 GB (18.74%)	0	2.6.4
ip-172-31-3-174.us-east-2.compute.internal (172.31.3.174:50010)	1	In Service	29.4 GB	9.28 GB	9.28 GB	10.84 GB	526	9.28 GB (31.56%)	0	2.6.4

Decommissioning

Nodes	Last contact	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks In files under construction
18.219.52.187:50070/dfshealth.html#tab-snapshot				

Part 1

A)

Hive lineorder.tbl Transformation From | Separated to CSV

Hive Table Creation Code:

```
create table lineorder(
    lo_orderkey      int,
    lo_linenumber    int,
    lo_custkey       int,
    lo_partkey       int,
    lo_suppkey       int,
    lo_orderdate     int,
    lo_orderpriority varchar(15),
    lo_shippriority  varchar(1),
    lo_quantity      int,
    lo_extendedprice int,
    lo_ordertotalprice int,
    lo_discount      int,
    lo_revenue       int,
    lo_supplycost    int,
    lo_tax           int,
    lo_commitdate    int,
    lo_shipmode      varchar(10))
row format delimited fields
terminated by '|' stored as textfile;

load data local inpath '/home/ec2-user/lineorder.tbl'
overwrite into table lineorder;
```

Hive Table Creation Execution:

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
OK
Time taken: 0.449 seconds
hive> create table lineorder(
    >     lo_orderkey          int,
    >     lo_linenumber        int,
    >     lo_custkey           int,
    >     lo_partkey           int,
    >     lo_suppkey           int,
    >     lo_orderdate         int,
    >     lo_orderpriority     varchar(15),
    >     lo_shippriority      varchar(1),
    >     lo_quantity          int,
    >     lo_extendedprice     int,
    >     lo_ordertotalprice   int,
    >     lo_discount          int,
    >     lo_revenue            int,
    >     lo_supplycost         int,
    >     lo_tax                int,
    >     lo_commitdate         int,
    >     lo_shipmode           varchar(10))
    > row format delimited fields
    > terminated by '|' stored as textfile;
OK
Time taken: 0.06 seconds
hive> load data local inpath '/home/ec2-user/lineorder.tbl'
    > overwrite into table lineorder;
Loading data to table default.lineorder
OK
Time taken: 42.818 seconds
hive>
```

Hive Transformation Code:

```
insert overwrite directory 'lo_hive_csv'
row format delimited
fields terminated by ','
stored as textfile
select * from lineorder;
```

Hive Transformation Execution:

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
hive> insert overwrite directory 'lo_hive_csv'
  > row format delimited
  > fields terminated by ','
  > stored as textfile
  > select * from lineorder;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20180311011106_8eb0730d-d6bd-4095-bfb9-471bba63ffc2
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0004, Tracking URL = http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0004/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0004
Hadoop job information for Stage-1: number of mappers: 10; number of reducers: 0
2018-03-11 01:11:14,820 Stage-1 map = 0%, reduce = 0%
2018-03-11 01:11:23,687 Stage-1 map = 10%, reduce = 0%, Cumulative CPU 2.38 sec
2018-03-11 01:11:29,199 Stage-1 map = 15%, reduce = 0%, Cumulative CPU 49.72 sec
2018-03-11 01:11:32,508 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 64.96 sec
2018-03-11 01:11:37,908 Stage-1 map = 25%, reduce = 0%, Cumulative CPU 87.08 sec
2018-03-11 01:11:38,974 Stage-1 map = 30%, reduce = 0%, Cumulative CPU 91.6 sec
2018-03-11 01:11:41,117 Stage-1 map = 45%, reduce = 0%, Cumulative CPU 103.91 sec
2018-03-11 01:11:43,241 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 108.06 sec
2018-03-11 01:11:37,908 Stage-1 map = 25%, reduce = 0%, Cumulative CPU 87.08 sec
2018-03-11 01:11:38,974 Stage-1 map = 30%, reduce = 0%, Cumulative CPU 91.6 sec
2018-03-11 01:11:41,117 Stage-1 map = 45%, reduce = 0%, Cumulative CPU 103.91 sec
2018-03-11 01:11:43,241 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 108.06 sec
2018-03-11 01:11:44,331 Stage-1 map = 65%, reduce = 0%, Cumulative CPU 117.21 sec
2018-03-11 01:11:51,908 Stage-1 map = 70%, reduce = 0%, Cumulative CPU 137.16 sec
2018-03-11 01:11:52,990 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 144.59 sec
2018-03-11 01:11:54,028 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 151.09 sec
MapReduce Total cumulative CPU time: 2 minutes 31 seconds 90 msec
Ended Job = job_1520720845975_0004
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://172.31.3.174/user/ec2-user/lo_hive_csv/.hive-staging_hive_2018-03-11_01-11-06_127_5642151675634568329-1/-ext-10000
Moving data to: lo_hive_csv
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Cumulative CPU: 151.09 sec   HDFS Read: 2417902701 HDFS Write: 2417756563 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 31 seconds 90 msec
OK
Time taken: 50.157 seconds
hive>
```

The Hive transformation code took 50.157 secs to execute on the 4-node cluster.

Hive CSV File Output Size and File Contents:

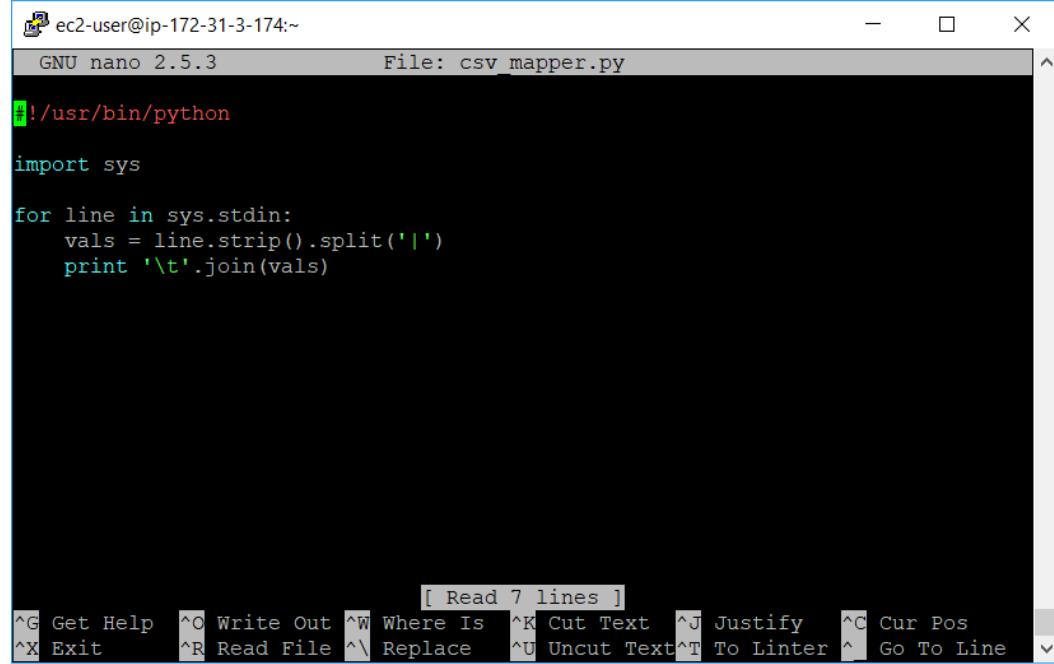
```
ec2-user@ip-172-31-3-174:~  
vielens  
-rw-r--r-- 2 ec2-user supergroup 51039483 2018-03-11 00:12 /user/ec2-user/pa  
rt.tbl  
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:56 /user/ec2-user/re  
commendations  
-rw-r--r-- 2 ec2-user supergroup 3344696 2018-03-11 00:12 /user/ec2-user/su  
pplier.tbl  
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -ls /user/ec2-user/lo_hive_csv/  
Found 10 items  
-rwxr-xr-x 2 ec2-user supergroup 268435460 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000000_0  
-rwxr-xr-x 2 ec2-user supergroup 268435470 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000001_0  
-rwxr-xr-x 2 ec2-user supergroup 268435464 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000002_0  
-rwxr-xr-x 2 ec2-user supergroup 268435475 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000003_0  
-rwxr-xr-x 2 ec2-user supergroup 268435389 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000004_0  
-rwxr-xr-x 2 ec2-user supergroup 268435448 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000005_0  
-rwxr-xr-x 2 ec2-user supergroup 268435435 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000006_0  
-rwxr-xr-x 2 ec2-user supergroup 268435526 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000007_0  
-rwxr-xr-x 2 ec2-user supergroup 268435500 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000008_0  
-rwxr-xr-x 2 ec2-user supergroup 1837396 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv/000009_0  
[ec2-user@ip-172-31-3-174 ~]$  
  
ec2-user@ip-172-31-3-174:~  
_sample.tbl  
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:37 /user/ec2-user/ml  
_dataset  
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:36 /user/ec2-user/mo  
vielens  
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:35 /user/ec2-user/ou  
tput  
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 23:10 /user/ec2-user/ou  
tput11  
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:44 /user/ec2-user/ou  
tput11_orig  
-rw-r--r-- 2 ec2-user supergroup 89519809 2018-03-13 03:23 /user/ec2-user/pa  
rt-00000  
-rw-r--r-- 2 ec2-user supergroup 51039483 2018-03-11 00:12 /user/ec2-user/pa  
rt.tbl  
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 04:45 /user/ec2-user/pi  
g_prejoin  
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:56 /user/ec2-user/re  
commendations  
-rw-r--r-- 2 ec2-user supergroup 3344696 2018-03-11 00:12 /user/ec2-user/su  
pplier.tbl  
-rw-r--r-- 2 ec2-user supergroup 288374 2018-03-13 03:22 /user/ec2-user/sy  
nthetic_control.data  
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:43 /user/ec2-user/te  
stdata  
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -dus /user/ec2-user/lo_hive_csv  
dus: DEPRECATED: Please use 'du -s' instead.  
2417756563 /user/ec2-user/lo_hive_csv  
[ec2-user@ip-172-31-3-174 ~]$
```

```
ec2-user@ip-172-31-3-174:~  
5499653,4,113339,318839,14603,19950331,2-HIGH,0,5,928910,19406684,1,919620,11146 ^  
9,2,19950525,SHIP  
5499653,5,113339,310731,32305,19950331,2-HIGH,0,40,6966880,19406684,7,6479198,10  
4503,0,19950501,RAIL  
5499653,6,113339,48458,14243,19950331,2-HIGH,0,9,1265805,19406684,9,1151882,8438  
7,3,19950523,RAIL  
5499654,1,2989,53874,23755,19950525,3-MEDIUM,0,31,5666397,27269958,3,5496405,109  
672,4,19950705,FOB  
5499654,2,2989,247239,9241,19950525,3-MEDIUM,0,28,3321416,27269958,5,3155345,711  
73,5,19950702,MAIL  
5499654,3,2989,138875,17795,19950525,3-MEDIUM,0,2,382774,27269958,3,371290,11483  
2,6,19950801,REG AIR  
5499654,4,2989,66338,39860,19950525,3-MEDIUM,0,43,5608619,27269958,9,5103843,782  
59,7,19950711,RAIL  
5499654,5,2989,135690,39354,19950525,3-MEDIUM,0,40,6902760,27269958,3,6695677,10  
3541,4,19950722,FOB  
5499654,6,2989,135024,20007,19950525,3-MEDIUM,0,5,529510,27269958,3,513624,63541  
,1,19950705,FOB  
5499654,7,2989,143470,12138,19950525,3-MEDIUM,0,36,5448492,27269958,10,4903642,9  
0808,0,19950803,SHIP  
5499655,1,57451,84565,26583,19930723,1-URGENT,0,13,2014428,12909928,6,1893562,92  
973,0,19930913,REG AIR  
5499655,2,57451,83691,8617,19930723,1-URGENT,0,23,3851787,12909928,5,3659197,100  
481,6,19930831,FOB  
5499655,3,57451,184583,24956,19930723,1-URGENT,0,27,4502466,12909928,0,4502466,1  
00054,1,19931017,FOB  
5499655,4,57451,277321,22255,19930723,1-URGENT,0,19,2466789,12909928,0,2466789,7  
7898,5,19930915,AIR  
5499680,1,43138,267850,6833,19920521,2-HIGH,0,50,9089200,2cat: Filesystem closed  
[ec2-user@ip-172-31-3-174 ~]$
```

The size of the Hive CSV output directory is 2417756563 bytes or approximately 2.418 GB.

Hadoop Streaming lineorder.tbl Transformation From | Separated to CSV

Hadoop Streaming Mapper Code:



The screenshot shows a terminal window titled "ec2-user@ip-172-31-3-174:~". Inside the terminal, the file "File: csv_mapper.py" is being edited in the "GNU nano 2.5.3" editor. The code in the file is as follows:

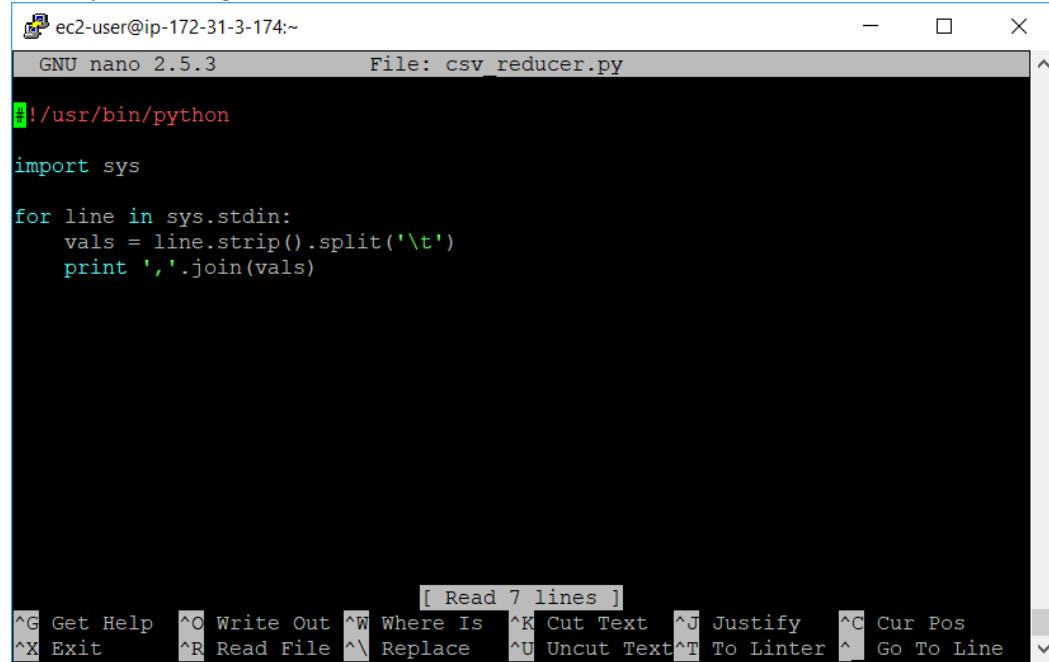
```
#!/usr/bin/python

import sys

for line in sys.stdin:
    vals = line.strip().split('|')
    print '\t'.join(vals)
```

At the bottom of the terminal window, there is a menu bar with various keyboard shortcuts for navigating and editing the text.

Hadoop Streaming Reducer Code:



The screenshot shows a terminal window titled "File: csv_reducer.py" with the following content:

```
GNU nano 2.5.3          File: csv_reducer.py
#!/usr/bin/python

import sys

for line in sys.stdin:
    vals = line.strip().split('\t')
    print ','.join(vals)
```

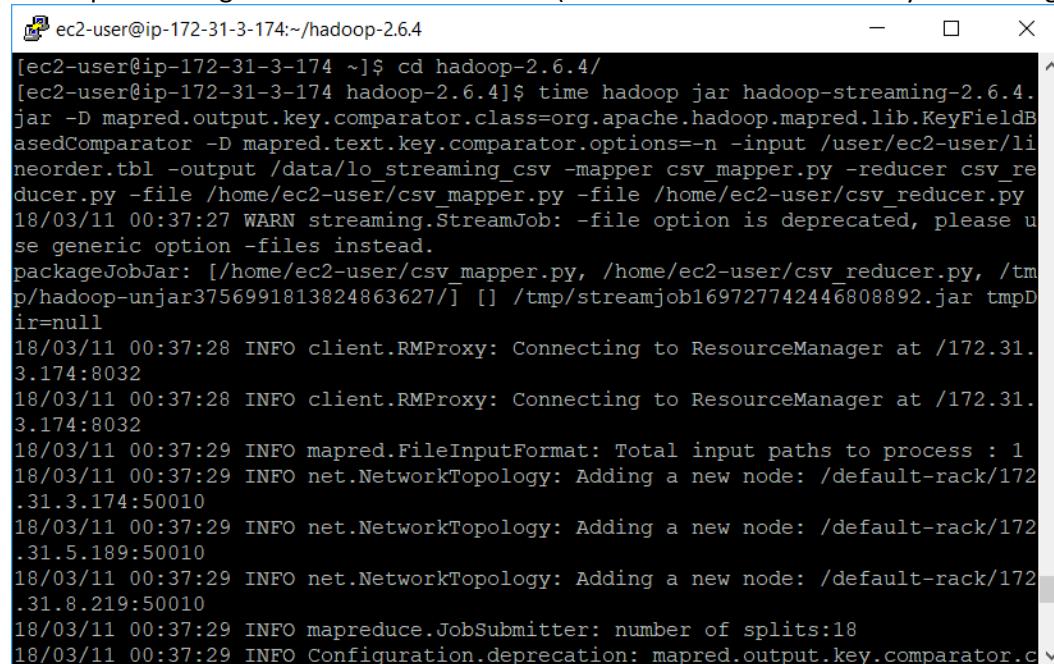
The bottom of the terminal shows the nano editor's command bar with various keyboard shortcuts.

Hadoop Streaming Command Used:

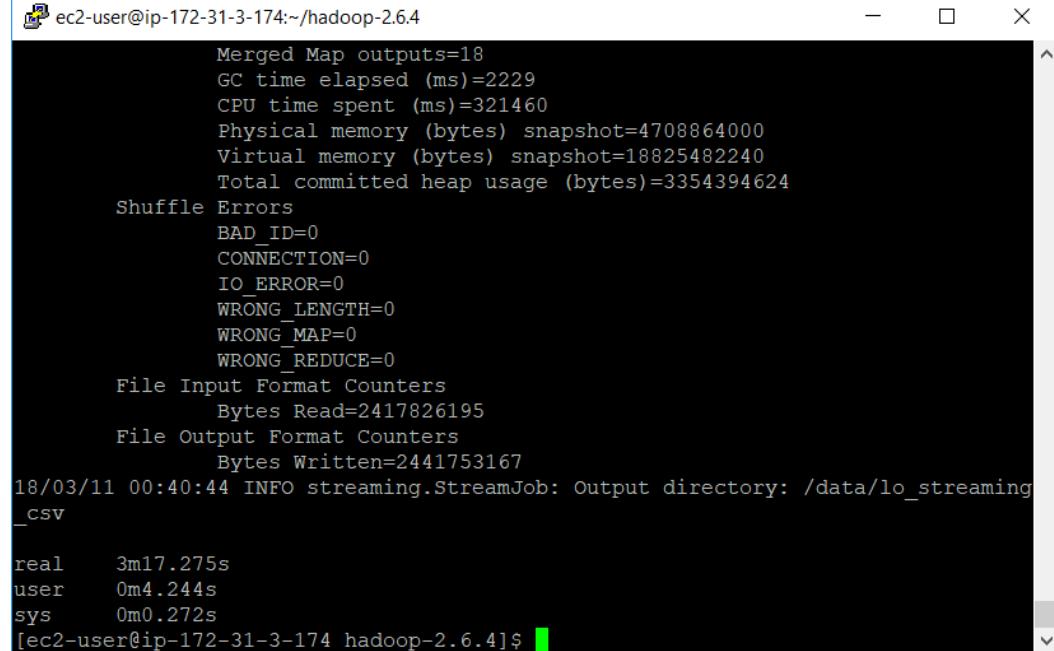
```
time hadoop jar hadoop-streaming-2.6.4.jar -D  
mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D  
mapred.text.key.comparator.options=-n -input /user/ec2-user/lineorder.tbl -output  
/data/lo_streaming_csv -mapper csv_mapper.py -reducer csv_reducer.py -file /home/ec2-  
user/csv_mapper.py -file /home/ec2-user/csv_reducer.py
```

The mapper and reducer keys are set to the `lo_orderkey` value of the `lineorder` table, which is an integer. Note that I used the `KeyFileBasedComparator` and `comparator.options` fields so the key from the mapper would be interpreted as a number instead of a string.

Hadoop Streaming Transformation Execution (first and last screenshots only due to length):



[ec2-user@ip-172-31-3-174 ~]\$ cd hadoop-2.6.4
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]\$ time hadoop jar hadoop-streaming-2.6.4.jar -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D mapred.text.key.comparator.options=-n -input /user/ec2-user/lineorder.tbl -output /data/lo_streaming_csv -mapper csv_mapper.py -reducer csv_reducer.py -file /home/ec2-user/csv_mapper.py -file /home/ec2-user/csv_reducer.py
18/03/11 00:37:27 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/ec2-user/csv_mapper.py, /home/ec2-user/csv_reducer.py, /tmp/hadoop-unjar3756991813824863627/] [] /tmp/streamjob169727742446808892.jar tmpDir=null
18/03/11 00:37:28 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/11 00:37:28 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/11 00:37:29 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/11 00:37:29 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.3.174:50010
18/03/11 00:37:29 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.5.189:50010
18/03/11 00:37:29 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.8.219:50010
18/03/11 00:37:29 INFO mapreduce.JobSubmitter: number of splits:18
18/03/11 00:37:29 INFO Configuration.deprecation: mapred.output.key.comparator.c



[ec2-user@ip-172-31-3-174 ~]\$
Merged Map outputs=18
GC time elapsed (ms)=2229
CPU time spent (ms)=321460
Physical memory (bytes) snapshot=4708864000
Virtual memory (bytes) snapshot=18825482240
Total committed heap usage (bytes)=3354394624
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=2417826195
File Output Format Counters
Bytes Written=2441753167
18/03/11 00:40:44 INFO streaming.StreamJob: Output directory: /data/lo_streaming_csv
real 3m17.275s
user 0m4.244s
sys 0m0.272s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]\$

The Hadoop Streaming transformation code took 3 min and 17.27 secs (or 197.27 secs) to execute on the 4-node cluster.

Hadoop Streaming CSV File Output Size and File Contents:

```
ec2-user@ip-172-31-3-174:~/hadoop-2.6.4
  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=2417826195
  File Output Format Counters
    Bytes Written=2441753167
18/03/11 00:40:44 INFO streaming.StreamJob: Output directory: /data/lo_streaming_csv
real    3m17.275s
user    0m4.244s
sys     0m0.272s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/lo_streaming_csv
Found 2 items
-rw-r--r--  2 ec2-user supergroup          0 2018-03-11 00:40 /data/lo_streaming_csv/_SUCCESS
-rw-r--r--  2 ec2-user supergroup 2441753167 2018-03-11 00:40 /data/lo_streaming_csv/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ 

ec2-user@ip-172-31-3-174:~/hadoop-2.6.4
1812260,2,51985,53814,904,19940930,2-HIGH,0,32,5656992,14352254,1,5600422,106068
,8,19941219,TRUCK
1812260,3,51985,48607,30301,19940930,2-HIGH,0,23,3577880,14352254,2,3506322,9333
6,3,19941220,SHIP
1812260,4,51985,127569,21953,19940930,2-HIGH,0,25,3991400,14352254,3,3871658,957
93,8,19941227,FOB
1812261,3,83008,32002,5593,19950804,2-HIGH,0,50,4670000,18358858,3,4529900,56040
,4,19951020,SHIP
1812261,1,83008,101629,16181,19950804,2-HIGH,0,33,5381046,18358858,4,5165804,978
37,1,19951023,RAIL
1812261,2,83008,66808,19664,19950804,2-HIGH,0,50,8874000,18358858,5,8430300,1064
88,0,19951006,SHIP
1812262,2,109436,121862,13229,19931015,5-LOW,0,40,7535440,17567877,7,7007959,113
031,6,19931207,SHIP
1812262,3,109436,127542,13906,19931015,5-LOW,0,8,1255632,17567877,5,1192850,9417
2,2,19940105,MAIL
1812262,1,109436,165367,29865,19931015,5-LOW,0,4,572944,17567877,2,561485,85941,
3,19931207,TRUCK
1812262,4,109436,395316,26595,19931015,5-LOW,0,9,1270170,17567877,2,1244766,8467
8,7,19931230,SHIP
1812262,5,109436,73400,17114,19931015,5-LOW,0,21,2884140,17567877,0,2884140,8240
4,8,19931122,FOB
1812262,6,109436,26706,28462,19931015,5-LOW,0,17,2775590,17567877,3,2692322,9796
2,3,19931127,[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ 
```

The size of the Hadoop Streaming CSV output file is 2441753167 bytes or approximately 2.442 GB.

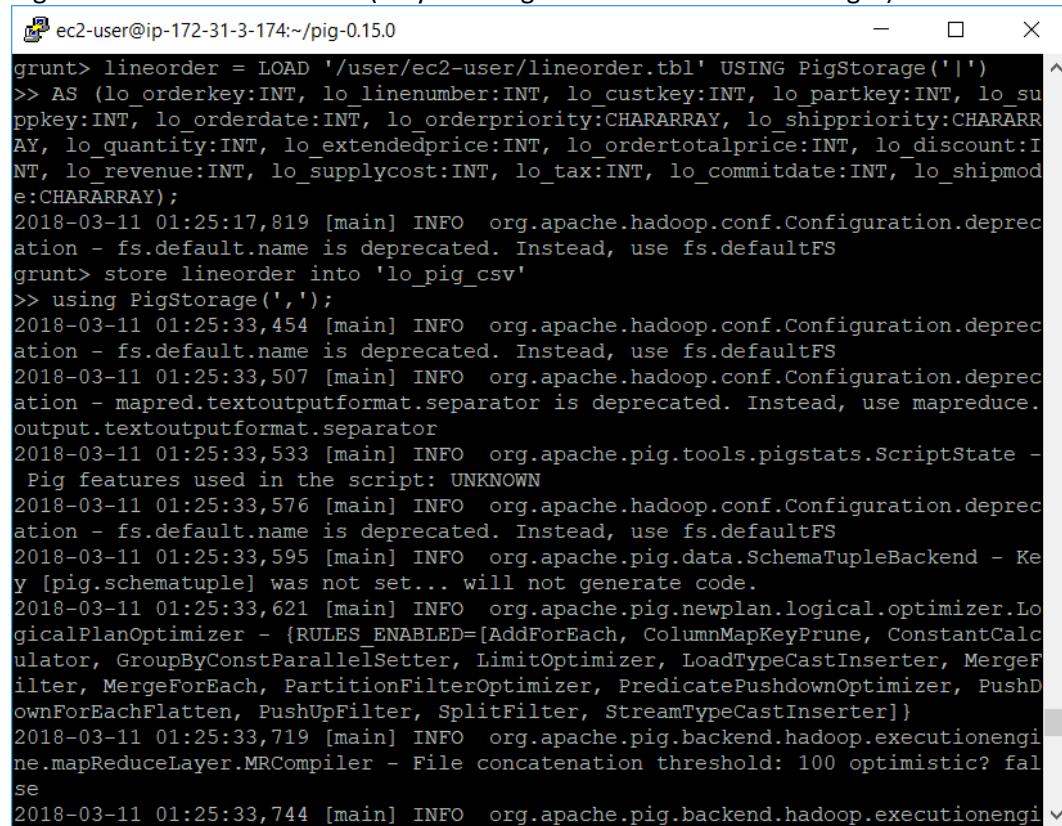
Pig lineorder.tbl Transformation From | Separated to CSV

Pig Transformation Code:

```
lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_suppkey:INT,
lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARRAY, lo_quantity:INT,
lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:INT, lo_revenue:INT, lo_supplycost:INT,
lo_tax:INT, lo_commitdate:INT, lo_shipmode:CHARARRAY);

store lineorder into 'lo_pig_csv'
using PigStorage(',');
```

Pig Transformation Execution (only showing start and finish due to length):



```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
grunt> lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
>> AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_suppkey:INT,
lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARRAY, lo_quantity:INT,
lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:INT, lo_revenue:INT, lo_supplycost:INT,
lo_tax:INT, lo_commitdate:INT, lo_shipmode:CHARARRAY);
2018-03-11 01:25:17,819 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> store lineorder into 'lo_pig_csv'
>> using PigStorage(',');
2018-03-11 01:25:33,454 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-11 01:25:33,507 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2018-03-11 01:25:33,533 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2018-03-11 01:25:33,576 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-11 01:25:33,595 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-03-11 01:25:33,621 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-11 01:25:33,719 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-11 01:25:33,744 [main] INFO  org.apache.pig.backend.hadoop.executionengi
```

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
ne.mapReduceLayer.MapReduceLauncher - 46% complete
2018-03-11 01:26:39,208 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1520720845975_0007]
2018-03-11 01:26:45,218 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /172.31.3.174:8032
2018-03-11 01:26:45,227 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-11 01:26:46,076 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /172.31.3.174:8032
2018-03-11 01:26:46,080 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-11 01:26:46,137 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2018-03-11 01:26:46,138 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /172.31.3.174:8032
2018-03-11 01:26:46,143 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-11 01:26:46,189 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-11 01:26:46,190 [main] INFO org.apache.pig.tools.pigstats.mapreduce.Sim
plePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.4 0.15.0 ec2-user 2018-03-11 01:25:33 2018-03-11 01:26:46 U
KNOWN

Success!
```

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMa
pTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime A
lias Feature Outputs
job_1520720845975_0007 18 0 61 54 58 58 0 0
0 0 lineorder MAP_ONLY hdfs://172.31.3.174/user/ec2-use
r/lo_pig_csv,

Input(s):
Successfully read 23996604 records (2417832927 bytes) from: "/user/ec2-user/line
order.tbl"

Output(s):
Successfully stored 23996604 records (2417756563 bytes) in: "hdfs://172.31.3.174
/user/ec2-user/lo_pig_csv"

Counters:
Total records written : 23996604
Total bytes written : 2417756563
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520720845975_0007
```

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
Counters:
Total records written : 23996604
Total bytes written : 2417756563
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520720845975_0007

2018-03-11 01:26:46,191 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 01:26:46,195 [main] INFO org.apache.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 01:26:46,285 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 01:26:46,289 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 01:26:46,337 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 01:26:46,342 [main] INFO org.apache.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 01:26:46,394 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

The Pig transformation code started at 1:25:33 and ended at 1:26:46, so it took 1 min 13 secs (or 73 secs) to execute on the 4-node cluster.

Pig CSV File Output Size and File Contents:

```
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -ls /user/ec2-user/lo_pig_csv/
Found 19 items
-rw-r--r-- 2 ec2-user supergroup 0 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/_SUCCESS
-rw-r--r-- 2 ec2-user supergroup 136055151 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00000
-rw-r--r-- 2 ec2-user supergroup 134217768 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00001
-rw-r--r-- 2 ec2-user supergroup 134217785 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00002
-rw-r--r-- 2 ec2-user supergroup 134217685 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00003
-rw-r--r-- 2 ec2-user supergroup 134217750 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00004
-rw-r--r-- 2 ec2-user supergroup 134217705 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00005
-rw-r--r-- 2 ec2-user supergroup 134217725 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00006
-rw-r--r-- 2 ec2-user supergroup 134217716 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00007
-rw-r--r-- 2 ec2-user supergroup 134217755 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00008
-rw-r--r-- 2 ec2-user supergroup 134217732 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00009
-rw-r--r-- 2 ec2-user supergroup 134217678 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00010
-rw-r--r-- 2 ec2-user supergroup 134217730 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00011
-rw-r--r-- 2 ec2-user supergroup 134217757 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00012

[ec2-user@ip-172-31-3-174 ~]$ ls
_pig_csv/part-m-00003
-rw-r--r-- 2 ec2-user supergroup 134217750 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00004
-rw-r--r-- 2 ec2-user supergroup 134217705 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00005
-rw-r--r-- 2 ec2-user supergroup 134217725 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00006
-rw-r--r-- 2 ec2-user supergroup 134217716 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00007
-rw-r--r-- 2 ec2-user supergroup 134217755 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00008
-rw-r--r-- 2 ec2-user supergroup 134217732 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00009
-rw-r--r-- 2 ec2-user supergroup 134217678 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00010
-rw-r--r-- 2 ec2-user supergroup 134217730 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00011
-rw-r--r-- 2 ec2-user supergroup 134217757 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00012
-rw-r--r-- 2 ec2-user supergroup 134217782 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00013
-rw-r--r-- 2 ec2-user supergroup 134217659 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00014
-rw-r--r-- 2 ec2-user supergroup 134217744 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00015
-rw-r--r-- 2 ec2-user supergroup 134217696 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00016
-rw-r--r-- 2 ec2-user supergroup 134217745 2018-03-11 01:26 /user/ec2-user/lo
_pig_csv/part-m-00017
```

```

ec2-user@ip-172-31-3-174:~ drwxr-xr-x - ec2-user supergroup 0 2018-03-11 03:20 /user/ec2-user/lo_pig_3c
ol drwxr-xr-x - ec2-user supergroup 0 2018-03-11 01:26 /user/ec2-user/lo_pig_cs
v -rw-r--r-- 2 ec2-user supergroup 9500 2018-03-11 00:15 /user/ec2-user/lo_sample
.tbl drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:37 /user/ec2-user/ml_databa
t drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:36 /user/ec2-user/movielens
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:35 /user/ec2-user/output
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 23:10 /user/ec2-user/output11
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:44 /user/ec2-user/output11_
orig -rw-r--r-- 2 ec2-user supergroup 89519809 2018-03-13 03:23 /user/ec2-user/part-0000
0 -rw-r--r-- 2 ec2-user supergroup 51039483 2018-03-11 00:12 /user/ec2-user/part.tbl
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 04:45 /user/ec2-user/pig_prejo
in drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:56 /user/ec2-user/recommend
ations -rw-r--r-- 2 ec2-user supergroup 3344696 2018-03-11 00:12 /user/ec2-user/supplier.
tbl -rw-r--r-- 2 ec2-user supergroup 288374 2018-03-13 03:22 /user/ec2-user/synthetic
_control.data drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:43 /user/ec2-user/testdata
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -dus /user/ec2-user/lo_pig_csv
dus: DEPRECATED: Please use 'du -s' instead.
2417756563 /user/ec2-user/lo_pig_csv
[ec2-user@ip-172-31-3-174 ~]$
```



```

ec2-user@ip-172-31-3-174:~ 0174,1,19931114,MAIL
22721863,2,68347,96283,25879,19930907,2-HIGH,0,22,2814416,10150452,2,2758127,767
56,2,19931107,RAIL
22721888,1,37175,162077,15659,19940730,2-HIGH,0,48,5467536,17645488,7,5084808,68
344,4,19941016,AIR
22721888,2,37175,18984,15708,19940730,2-HIGH,0,46,8753708,17645488,9,7965874,114
178,2,19940923,RAIL
22721888,3,37175,252058,891,19940730,2-HIGH,0,9,909036,17645488,10,818132,60602,
4,19941001,REG AIR
22721888,4,37175,333094,5429,19940730,2-HIGH,0,30,3381240,17645488,0,3381240,676
24,0,19941007,SHIP
22721889,1,76522,213518,23899,19920113,1-URGENT,0,31,4437650,13929437,10,3993885
,85890,7,19920219,SHIP
22721889,2,76522,180446,29464,19920113,1-URGENT,0,46,7021624,13929437,1,6951407,
91586,5,19920317,SHIP
22721889,3,76522,347898,25524,19920113,1-URGENT,0,12,2335056,13929437,2,2288354,
116752,3,19920411,SHIP
22721890,1,45106,273884,3001,19970605,1-URGENT,0,28,5202036,22463423,7,4837893,1
11472,2,19970829,AIR
22721890,2,45106,366648,34961,19970605,1-URGENT,0,49,8401687,22463423,10,7561518
,102877,7,19970902,RAIL
22721890,3,45106,246408,26606,19970605,1-URGENT,0,23,3115097,22463423,4,2990493,
81263,6,19970711,RAIL
22721890,4,45106,165838,2131,19970605,1-URGENT,0,21,3998043,22463423,8,3678199,1
14229,1,19970715,REG AIR
22721890,5,45106,35373,2820,19970605,1-URGENT,0,21,2747577,22463423,8,2527770,78
502,1,19970811,SHIP
22721891,1,37873,199352,31081,19930820,5-LOW,0,36,5224860,10121604,3,50cat: File
system closed
[ec2-user@ip-172-31-3-174 ~]$
```

The size of the Pig CSV output directory is 2417756563 bytes or approximately 2.418 GB.

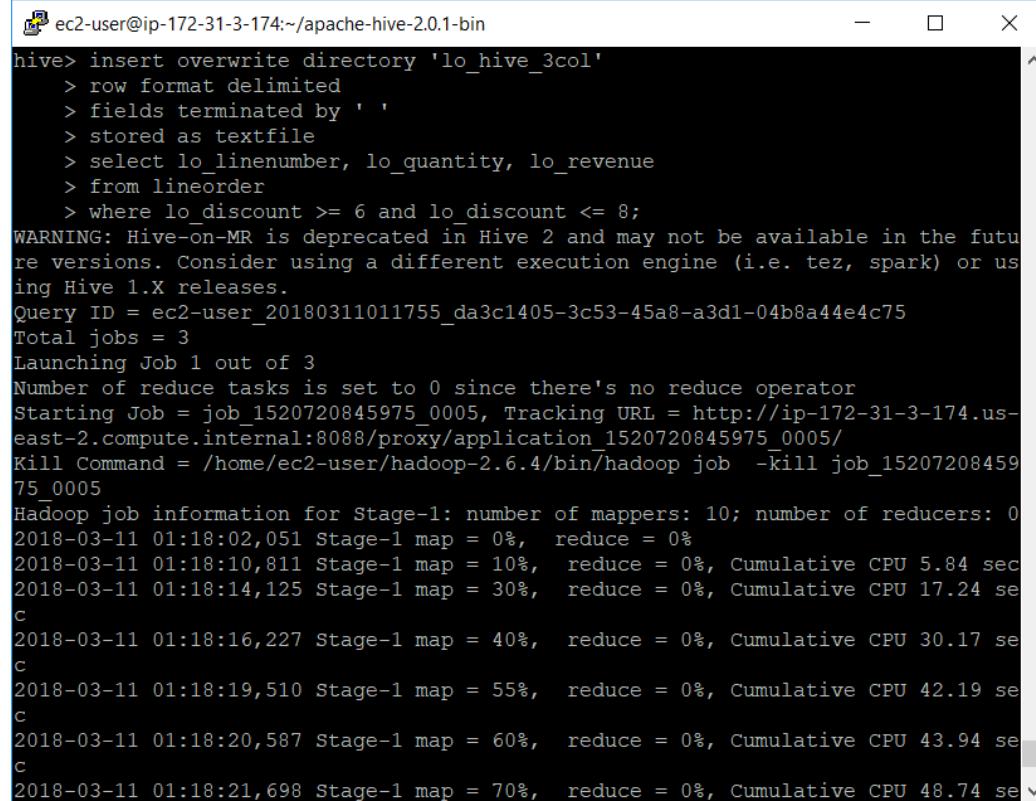
B)

Hive lineorder.tbl 3 Column Extraction

Hive 3 Column Extraction Code:

```
insert overwrite directory 'lo_hive_3col'
row format delimited
fields terminated by ''
stored as textfile
select lo_linenumber, lo_quantity, lo_revenue
from lineorder
where lo_discount >= 6 and lo_discount <= 8;
```

Hive 3 Column Extraction Execution:



The screenshot shows a terminal window titled 'ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin'. The user has run a complex Hive query to extract three columns from the 'lineorder' table, filtering rows where the discount is between 6 and 8. The command includes several nested 'select' statements and a 'where' clause. A warning message is displayed about the deprecation of Hive-on-MR. The terminal also shows the job ID, total jobs (3), and the start of the Hadoop job, detailing the number of mappers and reducers, and the progress of the Stage-1 map tasks over time.

```
hive> insert overwrite directory 'lo_hive_3col'
  > row format delimited
  > fields terminated by ''
  > stored as textfile
  > select lo_linenumber, lo_quantity, lo_revenue
  > from lineorder
  > where lo_discount >= 6 and lo_discount <= 8;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20180311011755_da3c1405-3c53-45a8-a3d1-04b8a44e4c75
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0005, Tracking URL = http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0005/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0005
Hadoop job information for Stage-1: number of mappers: 10; number of reducers: 0
2018-03-11 01:18:02,051 Stage-1 map = 0%,  reduce = 0%
2018-03-11 01:18:10,811 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 5.84 sec
2018-03-11 01:18:14,125 Stage-1 map = 30%,  reduce = 0%, Cumulative CPU 17.24 sec
2018-03-11 01:18:16,227 Stage-1 map = 40%,  reduce = 0%, Cumulative CPU 30.17 sec
2018-03-11 01:18:19,510 Stage-1 map = 55%,  reduce = 0%, Cumulative CPU 42.19 sec
2018-03-11 01:18:20,587 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 43.94 sec
2018-03-11 01:18:21,698 Stage-1 map = 70%,  reduce = 0%, Cumulative CPU 48.74 sec
```

```

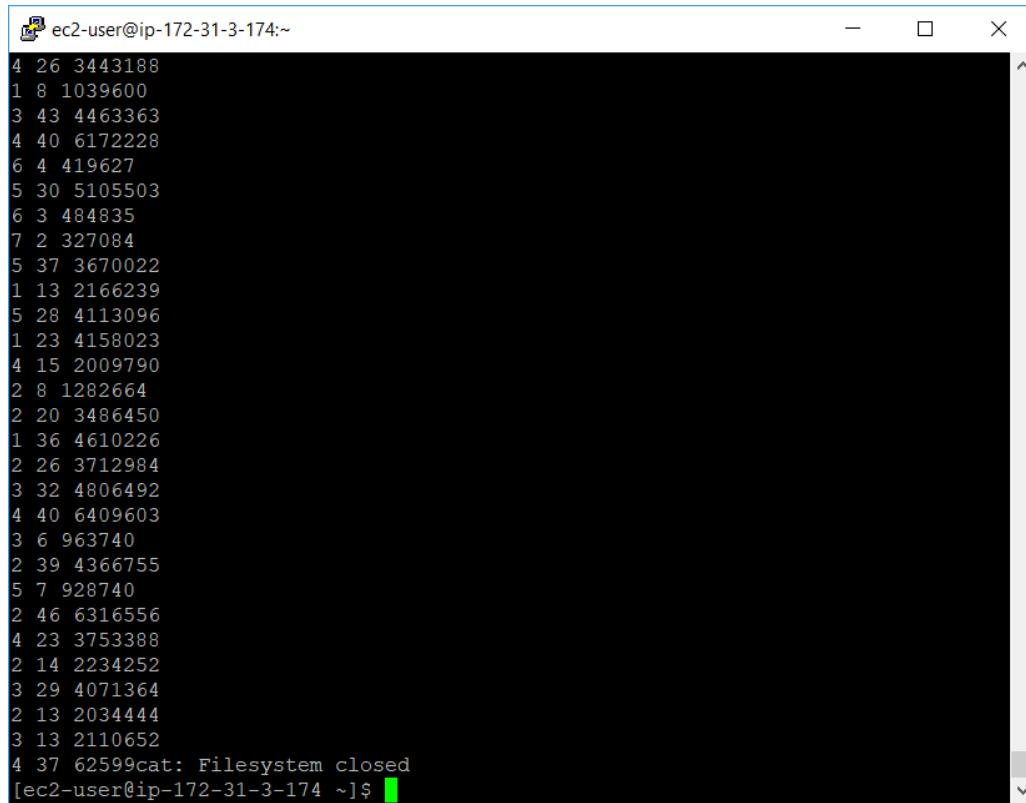
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
2018-03-11 01:18:20,587 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 43.94 sec
2018-03-11 01:18:21,698 Stage-1 map = 70%,  reduce = 0%, Cumulative CPU 48.74 sec
2018-03-11 01:18:22,785 Stage-1 map = 85%,  reduce = 0%, Cumulative CPU 52.49 sec
2018-03-11 01:18:24,895 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 54.49 sec
MapReduce Total cumulative CPU time: 54 seconds 490 msec
Ended Job = job_1520720845975_0005
Stage-3 is filtered out by condition resolver.
Stage-2 is selected by condition resolver.
Stage-4 is filtered out by condition resolver.
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0006, Tracking URL = http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0006/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 0
2018-03-11 01:18:30,644 Stage-2 map = 0%,  reduce = 0%
2018-03-11 01:18:41,994 Stage-2 map = 45%,  reduce = 0%, Cumulative CPU 7.1 sec
2018-03-11 01:18:45,085 Stage-2 map = 78%,  reduce = 0%, Cumulative CPU 10.28 sec
2018-03-11 01:18:46,116 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 11.48 sec
MapReduce Total cumulative CPU time: 11 seconds 480 msec
Ended Job = job_1520720845975_0006
Moving data to: lo_hive_3col
MapReduce Jobs Launched:
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
MapReduce Total cumulative CPU time: 54 seconds 490 msec
Ended Job = job_1520720845975_0005
Stage-3 is filtered out by condition resolver.
Stage-2 is selected by condition resolver.
Stage-4 is filtered out by condition resolver.
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0006, Tracking URL = http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0006/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 0
2018-03-11 01:18:30,644 Stage-2 map = 0%,  reduce = 0%
2018-03-11 01:18:41,994 Stage-2 map = 45%,  reduce = 0%, Cumulative CPU 7.1 sec
2018-03-11 01:18:45,085 Stage-2 map = 78%,  reduce = 0%, Cumulative CPU 10.28 sec
2018-03-11 01:18:46,116 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 11.48 sec
MapReduce Total cumulative CPU time: 11 seconds 480 msec
Ended Job = job_1520720845975_0006
Moving data to: lo_hive_3col
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Cumulative CPU: 54.49 sec  HDFS Read: 2417894121 HDFS Write: 82975501 SUCCESS
Stage-Stage-2: Map: 1  Cumulative CPU: 11.48 sec  HDFS Read: 82978541 HDFS Write: 82975501 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 5 seconds 970 msec
OK
Time taken: 51.232 seconds
hive>

```

The Hive 3 column extraction code took 51.232 secs to execute on the 4-node cluster.

Hive 3 Column File Output Size and File Contents:

```
ec2-user@ip-172-31-3-174:~$ ls -l /user/ec2-user/lo_hive_3col/
total 128
drwxr-xr-x 2 ec2-user supergroup 82975501 2018-03-11 01:18 /user/ec2-user/lo_hive_3col/000000_0
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -ls /user/ec2-user/lo_hive_3col/
Found 1 items
drwxr-xr-x 2 ec2-user supergroup 82975501 2018-03-11 01:18 /user/ec2-user/lo_hive_3col/000000_0
[ec2-user@ip-172-31-3-174 ~]$ ^[[1;32m
ec2-user@ip-172-31-3-174:~$ ls -l /user/ec2-user/lo_pig_3c/
total 128
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-11 03:20 /user/ec2-user/lo_pig_3c/0
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-11 01:26 /user/ec2-user/lo_pig_3c/1
drwxr-xr-x 2 ec2-user supergroup 9500 2018-03-11 00:15 /user/ec2-user/lo_sample
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-05 00:37 /user/ec2-user/ml_dataset
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-05 00:36 /user/ec2-user/movielens
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-13 03:35 /user/ec2-user/output
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-13 23:10 /user/ec2-user/output11
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-13 03:44 /user/ec2-user/output11_orig
drwxr-xr-x 2 ec2-user supergroup 89519809 2018-03-13 03:23 /user/ec2-user/part-00000
drwxr-xr-x 2 ec2-user supergroup 51039483 2018-03-11 00:12 /user/ec2-user/part.tbl
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-11 04:45 /user/ec2-user/pig_prejoin
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-05 00:56 /user/ec2-user/recommendations
drwxr-xr-x 2 ec2-user supergroup 3344696 2018-03-11 00:12 /user/ec2-user/supplier
drwxr-xr-x 2 ec2-user supergroup 288374 2018-03-13 03:22 /user/ec2-user/synthetic
drwxr-xr-x 1 ec2-user supergroup 0 2018-03-13 03:43 /user/ec2-user/testdata
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -dus /user/ec2-user/lo_hive_3col
dus: DEPRECATED: Please use 'du -s' instead.
82975501 /user/ec2-user/lo_hive_3col
[ec2-user@ip-172-31-3-174 ~]$ ^[[1;32m
```



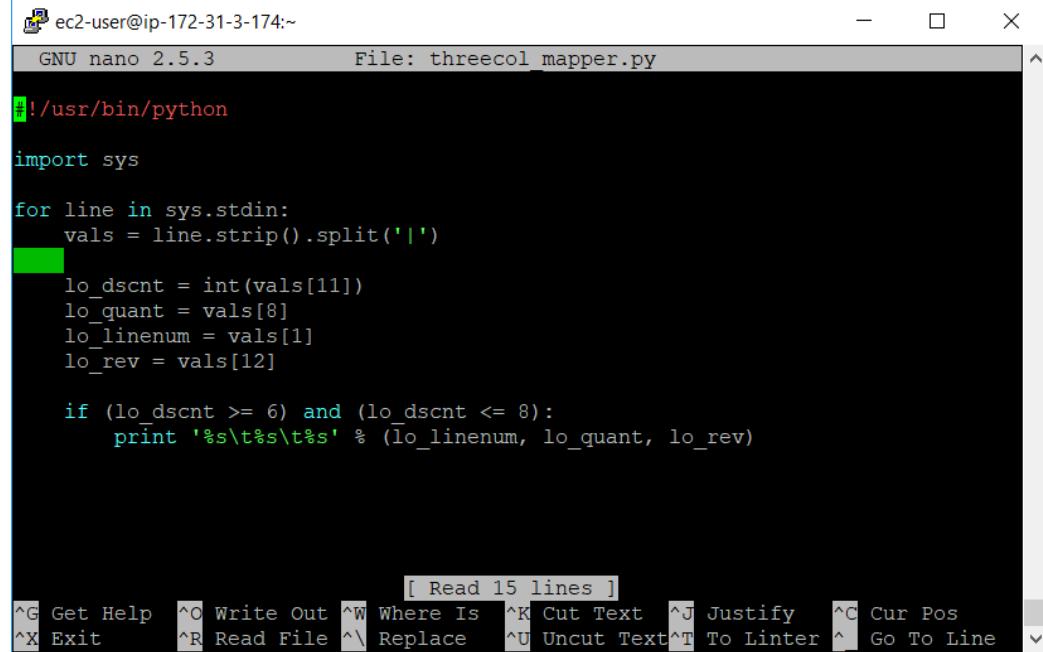
A screenshot of a terminal window titled "ec2-user@ip-172-31-3-174:~". The window contains a large amount of text output from a command, likely a file listing or log dump. The text is mostly numerical values (e.g., 4 26 3443188, 1 8 1039600) and ends with the message "cat: Filesystem closed". The terminal has a standard Windows-style interface with minimize, maximize, and close buttons at the top right.

```
4 26 3443188
1 8 1039600
3 43 4463363
4 40 6172228
6 4 419627
5 30 5105503
6 3 484835
7 2 327084
5 37 3670022
1 13 2166239
5 28 4113096
1 23 4158023
4 15 2009790
2 8 1282664
2 20 3486450
1 36 4610226
2 26 3712984
3 32 4806492
4 40 6409603
3 6 963740
2 39 4366755
5 7 928740
2 46 6316556
4 23 3753388
2 14 2234252
3 29 4071364
2 13 2034444
3 13 2110652
4 37 62599cat: Filesystem closed
[ec2-user@ip-172-31-3-174 ~]$
```

The size of the Hive 3 column output directory is 82975501 bytes or approximately 82.976 MB.

Hadoop Streaming lineorder.tbl 3 Column Extraction

Hadoop Streaming Mapper Code:



The screenshot shows a terminal window titled "GNU nano 2.5.3" with the file "threecol_mapper.py" open. The code is a Python script for a Hadoop Streamer. It reads from standard input, splits each line by '|', and extracts specific fields (lo_dscnt, lo_quant, lo_linenum, lo_rev) into a single output line if lo_dscnt is between 6 and 8. The terminal also displays a menu bar at the bottom with various keyboard shortcuts.

```
ec2-user@ip-172-31-3-174:~$ GNU nano 2.5.3          File: threecol_mapper.py
#!/usr/bin/python

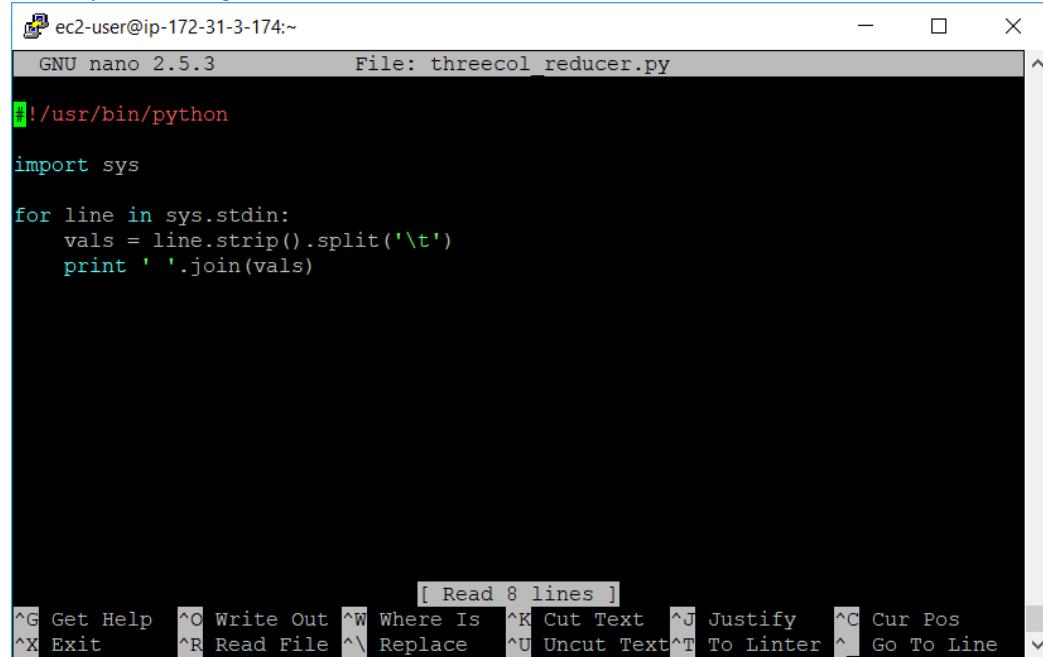
import sys

for line in sys.stdin:
    vals = line.strip().split('|')
    lo_dscnt = int(vals[11])
    lo_quant = vals[8]
    lo_linenum = vals[1]
    lo_rev = vals[12]

    if (lo_dscnt >= 6) and (lo_dscnt <= 8):
        print '%s\t%s\t%s' % (lo_linenum, lo_quant, lo_rev)

[ Read 15 lines ]
^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace   ^U Uncut Text^T To Linter  ^  Go To Line
```

Hadoop Streaming Reducer Code:



The screenshot shows a terminal window titled "File: threecol_reducer.py" in the "GNU nano 2.5.3" editor. The code is a Python script that reads from standard input, strips whitespace, splits each line by tabs, and then prints the values joined by spaces. The terminal also displays a command-line interface with various keyboard shortcuts.

```
ec2-user@ip-172-31-3-174:~$ GNU nano 2.5.3          File: threecol_reducer.py
#!/usr/bin/python

import sys

for line in sys.stdin:
    vals = line.strip().split('\t')
    print ' '.join(vals)

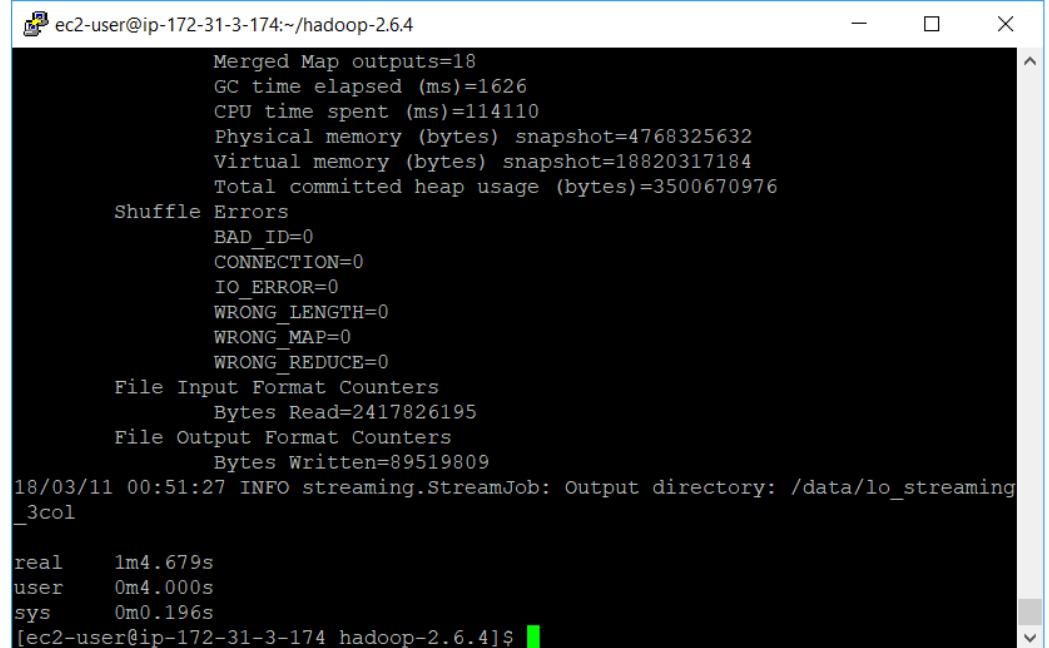
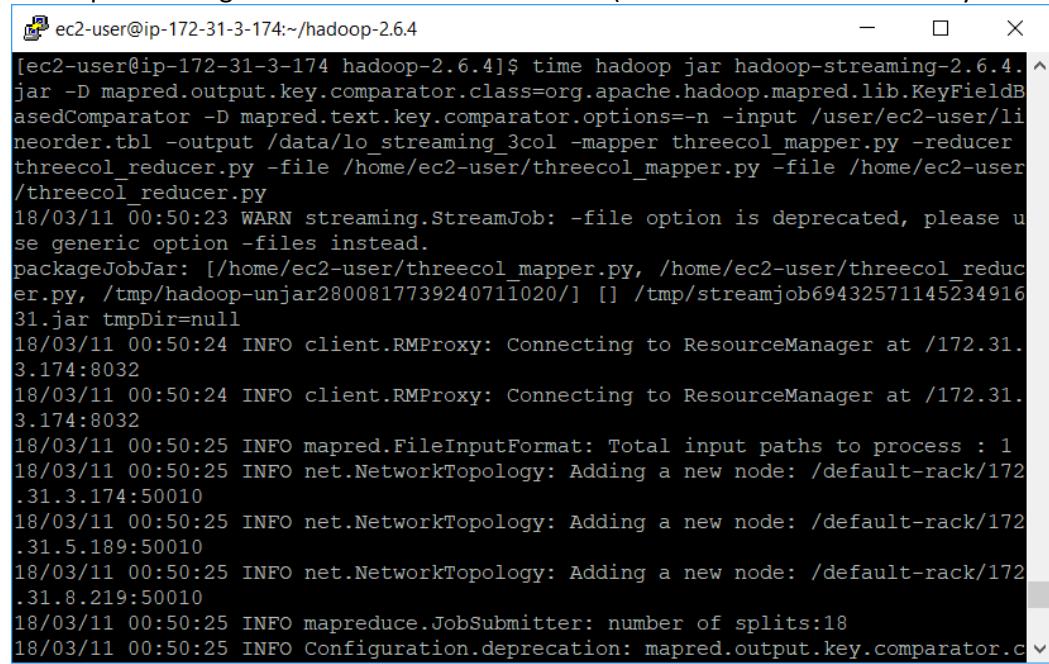
[ Read 8 lines ]
^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace   ^U Uncut Text^T To Linter  ^ Go To Line
```

Hadoop Streaming Command Used:

```
time hadoop jar hadoop-streaming-2.6.4.jar -D
mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D
mapred.text.key.comparator.options=-n -input /user/ec2-user/lineorder.tbl -output
/data/lo_streaming_3col -mapper threecol_mapper.py -reducer threecol_reducer.py -file /home/ec2-
user/threecol_mapper.py -file /home/ec2-user/threecol_reducer.py
```

The mapper key is set to the `lo_linenumber` value of the `lineorder` table, which is an integer. There is no key set in the reducer output because the goal is to have the columns separated by spaces (a key would induce a tab). Note that I used the `KeyFileBasedComparator` and `comparator.options` fields so the key from the mapper would be interpreted as a number instead of a string.

Hadoop Streaming 3 Column Extraction Execution (first and last screenshots only due to length):



```
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ time hadoop jar hadoop-streaming-2.6.4.jar -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D mapred.text.key.comparator.options=-n -input /user/ec2-user/li_neorder.tbl -output /data/lo_streaming_3col -mapper threecol_mapper.py -reducer threecol_reducer.py -file /home/ec2-user/threecol_mapper.py -file /home/ec2-user/threecol_reducer.py
18/03/11 00:50:23 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/ec2-user/threecol_mapper.py, /home/ec2-user/threecol_reducer.py, /tmp/hadoop-unjar2800817739240711020/] [] /tmp/streamjob6943257114523491631.jar tmpDir=null
18/03/11 00:50:24 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/11 00:50:24 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/11 00:50:25 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/11 00:50:25 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.3.174:50010
18/03/11 00:50:25 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.5.189:50010
18/03/11 00:50:25 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.8.219:50010
18/03/11 00:50:25 INFO mapreduce.JobSubmitter: number of splits:18
18/03/11 00:50:25 INFO Configuration.deprecation: mapred.output.key.comparator.c
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ Merged Map outputs=18
GC time elapsed (ms)=1626
CPU time spent (ms)=114110
Physical memory (bytes) snapshot=4768325632
Virtual memory (bytes) snapshot=18820317184
Total committed heap usage (bytes)=3500670976
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=2417826195
File Output Format Counters
Bytes Written=89519809
18/03/11 00:51:27 INFO streaming.StreamJob: Output directory: /data/lo_streaming_3col
real    1m4.679s
user    0m4.000s
sys     0m0.196s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$
```

The Hadoop Streaming transformation code took 1 min and 4.679 secs (or 64.679 secs) to execute on the 4-node cluster

Hadoop Streaming 3 Column File Output Size and File Contents:

```
ec2-user@ip-172-31-3-174:~/hadoop-2.6.4
  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=2417826195
  File Output Format Counters
    Bytes Written=89519809
18/03/11 00:51:27 INFO streaming.StreamJob: Output directory: /data/lo_streaming_3col
real    1m4.679s
user    0m4.000s
sys     0m0.196s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/lo_streaming_3col/
Found 2 items
-rw-r--r--  2 ec2-user supergroup      0 2018-03-11 00:51 /data/lo_streaming_3col/_SUCCESS
-rw-r--r--  2 ec2-user supergroup  89519809 2018-03-11 00:51 /data/lo_streaming_3col/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$
ec2-user@ip-172-31-3-174:~/hadoop-2.6.4
1 39 5162780
1 16 1862389
1 42 5554422
1 26 4012816
1 43 6332656
1 48 8156703
1 44 4485486
1 9 1114289
1 16 1833616
1 44 7270343
1 18 1942868
1 48 5343943
1 7 1150061
1 35 3571255
1 18 3123042
1 5 692991
1 44 6483155
1 12 2021979
1 23 2258762
1 10 1385906
1 35 4769942
1 48 8769911
1 2 229561
1 44 5887856[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$
```

The size of the Hadoop Streaming 3 column output file is 89519809 bytes or approximately 89.52 MB.

Pig lineorder.tbl 3 Column Extraction

Pig 3 Column Extraction Code:

```
lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_suppkey:INT,
lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARRAY, lo_quantity:INT,
lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:INT, lo_revenue:INT, lo_supplycost:INT,
lo_tax:INT, lo_commitdate:INT, lo_shipmode:CHARARRAY);

lo_filter = FILTER lineorder BY lo_discount >= 6 AND lo_discount <= 8;
lo_three = FOREACH lo_filter GENERATE lo_linenumber, lo_quantity, lo_revenue;
store lo_three into 'lo_pig_3col' using PigStorage(' ');
```

Pig 3 Column Extraction Execution (only showing start and finish due to length):

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
-
X
grunt> lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
>> AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_su
ppkey:INT, lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARR
AY, lo_quantity:INT, lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:I
NT, lo_revenue:INT, lo_supplycost:INT, lo_tax:INT, lo_commitdate:INT, lo_shipmod
e:CHARARRAY);
2018-03-11 03:19:12,024 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> lo_filter = FILTER lineorder BY lo_discount >= 6 AND lo_discount <= 8;
grunt> lo_three = FOREACH lo_filter GENERATE lo_linenumber, lo_quantity, lo_reve
nue;
grunt> store lo_three into 'lo_pig_3col' using PigStorage(' ');
2018-03-11 03:19:37,201 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-11 03:19:37,230 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.
output.textoutputformat.separator
2018-03-11 03:19:37,251 [main] INFO org.apache.pig.tools.pigstats.ScriptState -
Pig features used in the script: FILTER
2018-03-11 03:19:37,273 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-11 03:19:37,277 [main] INFO org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2018-03-11 03:19:37,302 [main] INFO org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalc
ulator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeF
ilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushD
ownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-11 03:19:37,323 [main] INFO org.apache.pig.newplan.logical.rules.Column
PruneVisitor - Columns pruned for lineorder: $0, $2, $3, $4, $5, $6, $7, $9, $10
```

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
ne.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1520720845975_0024]
2018-03-11 03:20:20,716 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - 49% complete
2018-03-11 03:20:20,716 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1520720845975_0024]
2018-03-11 03:20:23,723 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /172.31.3.174:8032
2018-03-11 03:20:23,729 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-11 03:20:23,891 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /172.31.3.174:8032
2018-03-11 03:20:23,896 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-11 03:20:23,976 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2018-03-11 03:20:23,977 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /172.31.3.174:8032
2018-03-11 03:20:23,981 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-11 03:20:24,019 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-11 03:20:24,021 [main] INFO org.apache.pig.tools.pigstats.mapreduce.Sim
plePigStats - Script Statistics:
```

HadoopVersion	PigVersion	UserId	StartedAt	FinishedAt	Features
2.6.4	0.15.0	ec2-user	2018-03-11 03:19:37	2018-03-11 03:20:24	FILTER

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
FILTER
Success!

Job Stats (time in seconds):
JobId    Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMa
pTime    MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReducetime   A
lias    Feature Outputs
job_1520720845975_0024  18        0          36            23            30            28            0            0
0          lineorder,lo_filter,lo_three    MAP_ONLY      hdfs://172.31.3.
174/user/ec2-user/lo_pig_3col

Input(s):
Successfully read 23996604 records (2417832927 bytes) from: "/user/ec2-user/line
order.tbl"

Output(s):
Successfully stored 6544308 records (82975501 bytes) in: "hdfs://172.31.3.174/us
er/ec2-user/lo_pig_3col"

Counters:
Total records written : 6544308
Total bytes written : 82975501
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520720845975_0024
```

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
Counters:
Total records written : 6544308
Total bytes written : 82975501
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520720845975_0024

2018-03-11 03:20:24,022 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 03:20:24,026 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegator - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 03:20:24,097 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 03:20:24,102 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegator - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 03:20:24,137 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 03:20:24,142 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegator - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 03:20:24,180 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

The Pig transformation code started at 3:19:37 and ended at 3:20:24, so it took 53 secs to execute on the 4-node cluster.

Pig 3 Column File Output Size and File Contents:

```
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -ls /user/ec2-user/lo_pig_3col
Found 19 items
-rw-r--r-- 2 ec2-user supergroup          0 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/_SUCCESS
-rw-r--r-- 2 ec2-user supergroup 4652851 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00000
-rw-r--r-- 2 ec2-user supergroup 4657516 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00001
-rw-r--r-- 2 ec2-user supergroup 4638562 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00002
-rw-r--r-- 2 ec2-user supergroup 4629637 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00003
-rw-r--r-- 2 ec2-user supergroup 4624740 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00004
-rw-r--r-- 2 ec2-user supergroup 4639966 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00005
-rw-r--r-- 2 ec2-user supergroup 4628915 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00006
-rw-r--r-- 2 ec2-user supergroup 4626694 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00007
-rw-r--r-- 2 ec2-user supergroup 4599797 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00008
-rw-r--r-- 2 ec2-user supergroup 4587677 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00009
-rw-r--r-- 2 ec2-user supergroup 4588608 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00010
-rw-r--r-- 2 ec2-user supergroup 4589677 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00011
-rw-r--r-- 2 ec2-user supergroup 4591547 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00012
[ec2-user@ip-172-31-3-174 ~]$ ^

[ec2-user@ip-172-31-3-174 ~]$ ^

_pig_3col/part-m-00003
-rw-r--r-- 2 ec2-user supergroup 4624740 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00004
-rw-r--r-- 2 ec2-user supergroup 4639966 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00005
-rw-r--r-- 2 ec2-user supergroup 4628915 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00006
-rw-r--r-- 2 ec2-user supergroup 4626694 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00007
-rw-r--r-- 2 ec2-user supergroup 4599797 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00008
-rw-r--r-- 2 ec2-user supergroup 4587677 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00009
-rw-r--r-- 2 ec2-user supergroup 4588608 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00010
-rw-r--r-- 2 ec2-user supergroup 4589677 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00011
-rw-r--r-- 2 ec2-user supergroup 4591547 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00012
-rw-r--r-- 2 ec2-user supergroup 4587482 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00013
-rw-r--r-- 2 ec2-user supergroup 4576497 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00014
-rw-r--r-- 2 ec2-user supergroup 4592352 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00015
-rw-r--r-- 2 ec2-user supergroup 4579748 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00016
-rw-r--r-- 2 ec2-user supergroup 4583235 2018-03-11 03:20 /user/ec2-user/lo
_pig_3col/part-m-00017
[ec2-user@ip-172-31-3-174 ~]$ ^
```

```

ec2-user@ip-172-31-3-174:~ - ec2-user supergroup 0 2018-03-11 03:20 /user/ec2-user/lo_pig_3c
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 01:26 /user/ec2-user/lo_pig_cs
v -rw-r--r-- 2 ec2-user supergroup 9500 2018-03-11 00:15 /user/ec2-user/lo_sample
tbl drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:37 /user/ec2-user/ml_database
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:36 /user/ec2-user/movielens
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:35 /user/ec2-user/output
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 23:10 /user/ec2-user/output11
drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:44 /user/ec2-user/output11_
orig -rw-r--r-- 2 ec2-user supergroup 89519809 2018-03-13 03:23 /user/ec2-user/part-0000
0 -rw-r--r-- 2 ec2-user supergroup 51039483 2018-03-11 00:12 /user/ec2-user/part.tbl
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 04:45 /user/ec2-user/pig_prejoin
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:56 /user/ec2-user/recommendations
-rw-r--r-- 2 ec2-user supergroup 3344696 2018-03-11 00:12 /user/ec2-user/supplier.
tbl -rw-r--r-- 2 ec2-user supergroup 288374 2018-03-13 03:22 /user/ec2-user/synthetic
control.data drwxr-xr-x - ec2-user supergroup 0 2018-03-13 03:43 /user/ec2-user/testdata
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -dus /user/ec2-user/lo_pig_3col
dus: DEPRECATED: Please use 'du -s' instead.
82975501 /user/ec2-user/lo_pig_3col
[ec2-user@ip-172-31-3-174 ~]$
```



```

ec2-user@ip-172-31-3-174:~
1 15 2742294
2 5 802181
4 24 3436230
5 18 1952017
6 25 2641893
2 48 8977059
3 7 795678
3 2 226393
1 36 6206656
3 10 1145991
4 39 6183128
5 26 4821250
5 7 924384
6 37 3683196
5 30 3530509
1 18 1925233
2 48 5994278
1 48 5353005
3 30 3344773
5 50 7103386
2 42 4299490
2 36 5341745
2 23 3384817
4 46 5032652
6 22 2869122
1 19 2113756
2 18 2494634
4 50 5904187
2 38 4931073
1 16 [ec2-user@ip-172-31-3-174 ~]$
```

The size of the Pig 3 column output directory is 82975501 bytes or approximately 82.976 MB.

Part 2

A)

Hive Query 2.1

Hive Remaining Table Creation Code:

```
create table dwdate(
  d_datekey      int,
  d_date        varchar(19),
  d_dayofweek    varchar(10),
  d_month       varchar(10),
  d_year        int,
  d_yeарmonthnum   int,
  d_yeарmonth    varchar(8),
  d_daynuminweek  int,
  d_daynuminmonth  int,
  d_daynuminyear   int,
  d_monthnuminyear  int,
  d_weeknuminyear   int,
  d_sellingseason   varchar(13),
  d_lastdayinweekfl  varchar(1),
  d_lastdayinmonthfl  varchar(1),
  d_holidayfl     varchar(1),
  d_weekdayfl      varchar(1))
row format delimited fields
terminated by '|' stored as textfile;

load data local inpath '/home/ec2-user/dwdate.tbl'
overwrite into table dwdate;
```

```
create table part(
  p_partkey    int,
  p_name       varchar(22),
  p_mfgr       varchar(6),
  p_category    varchar(7),
  p_brand1      varchar(9),
  p_color       varchar(11),
  p_type        varchar(25),
  p_size        int,
  p_container    varchar(10))
row format delimited fields
terminated by '|' stored as textfile;
```

```
load data local inpath '/home/ec2-user/part.tbl'
overwrite into table part;
```

```
create table supplier(
    s_suppkey  int,
    s_name     varchar(25),
    s_address  varchar(25),
    s_city     varchar(10),
    s_nation   varchar(15),
    s_region   varchar(12),
    s_phone    varchar(15))
row format delimited fields
terminated by '|' stored as textfile;
```

```
load data local inpath '/home/ec2-user/supplier.tbl'
overwrite into table supplier;
```

```
create table customer (
    c_custkey  int,
    c_name     varchar(25),
    c_address  varchar(25),
    c_city     varchar(10),
    c_nation   varchar(15),
    c_region   varchar(12),
    c_phone    varchar(15),
    c_mktsegment varchar(10))
row format delimited fields
terminated by '|' stored as textfile;
```

```
load data local inpath '/home/ec2-user/customer.tbl'
overwrite into table customer;
```

Hive Remaining Table Creation Execution:

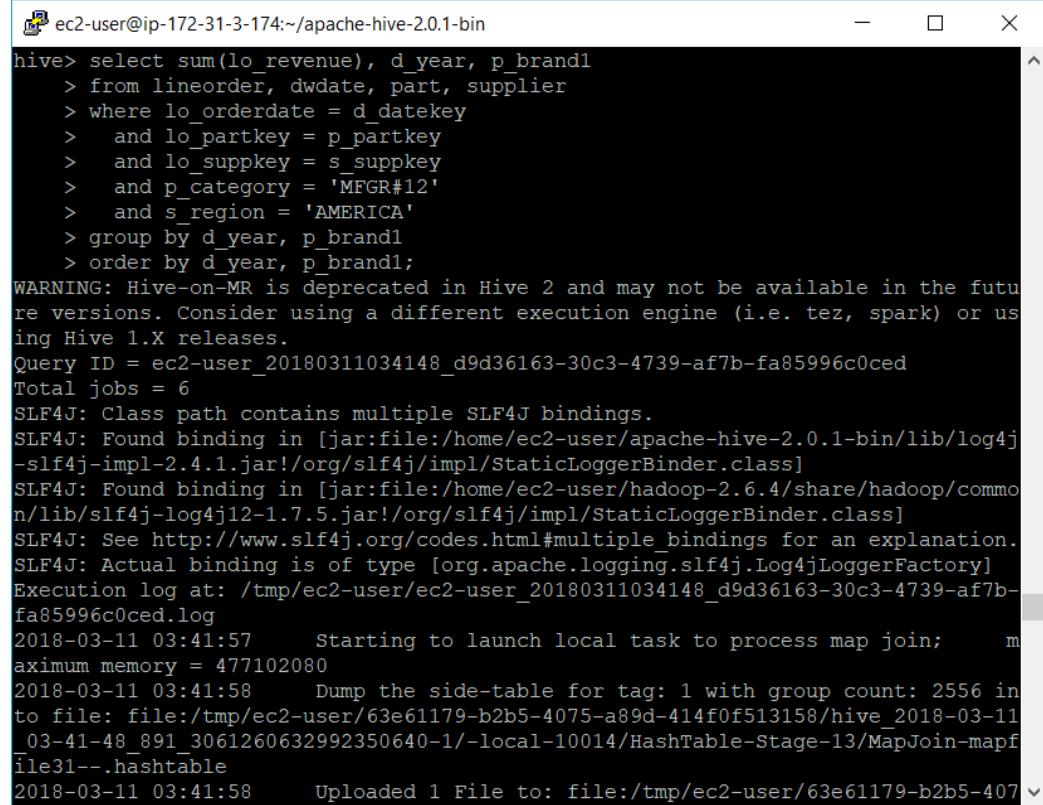
```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
OK
Time taken: 0.104 seconds
hive> create table dwdate(
    >     d_datekey          int,
    >     d_date              varchar(19),
    >     d_dayofweek         varchar(10),
    >     d_month             varchar(10),
    >     d_year              int,
    >     d_yeарmonthnum      int,
    >     d_yeарmonth         varchar(8),
    >     d_daynuminweek      int,
    >     d_daynuminmonth     int,
    >     d_daynuminyear      int,
    >     d_monthnuminyear    int,
    >     d_weeknuminyear     int,
    >     d_sellingseason     varchar(13),
    >     d_lastdayinweekfl   varchar(1),
    >     d_lastdayinmonthfl  varchar(1),
    >     d_holidayfl         varchar(1),
    >     d_weekdayfl         varchar(1)
    > row format delimited fields
    > terminated by '|' stored as textfile;
OK
Time taken: 0.339 seconds
hive> load data local inpath '/home/ec2-user/dwdate.tbl'
    > overwrite into table dwdate;
Loading data to table default.dwdate
OK
Time taken: 0.449 seconds
hive>
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
>     lo_shipmode          varchar(10))
> row format delimited fields
> terminated by '|' stored as textfile;
OK
Time taken: 0.06 seconds
hive> load data local inpath '/home/ec2-user/lineorder.tbl'
    > overwrite into table lineorder;
Loading data to table default.lineorder
OK
Time taken: 42.818 seconds
hive> create table part(
    >     p_partkey        int,
    >     p_name           varchar(22),
    >     p_mfgr            varchar(6),
    >     p_category        varchar(7),
    >     p_brand1          varchar(9),
    >     p_color           varchar(11),
    >     p_type            varchar(25),
    >     p_size             int,
    >     p_container        varchar(10))
    > row format delimited fields
    > terminated by '|' stored as textfile;
OK
Time taken: 0.05 seconds
hive> load data local inpath '/home/ec2-user/part.tbl'
    > overwrite into table part;
Loading data to table default.part
OK
Time taken: 0.642 seconds
hive>
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
> p_type      varchar(25),
> p_size      int,
> p_container varchar(10))
> row format delimited fields
> terminated by '|' stored as textfile;
OK
Time taken: 0.05 seconds
hive> load data local inpath '/home/ec2-user/part.tbl'
    > overwrite into table part;
Loading data to table default.part
OK
Time taken: 0.642 seconds
hive> create table supplier(
    > s_suppkey      int,
    > s_name      varchar(25),
    > s_address      varchar(25),
    > s_city      varchar(10),
    > s_nation      varchar(15),
    > s_region      varchar(12),
    > s_phone      varchar(15))
    > row format delimited fields
    > terminated by '|' stored as textfile;
OK
Time taken: 0.058 seconds
hive> load data local inpath '/home/ec2-user/supplier.tbl'
    > overwrite into table supplier;
Loading data to table default.supplier
OK
Time taken: 0.2 seconds
hive>
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
> s_region      varchar(12),
> s_phone      varchar(15))
> row format delimited fields
> terminated by '|' stored as textfile;
OK
Time taken: 0.058 seconds
hive> load data local inpath '/home/ec2-user/supplier.tbl'
    > overwrite into table supplier;
Loading data to table default.supplier
OK
Time taken: 0.2 seconds
hive> create table customer (
    > c_custkey      int,
    > c_name      varchar(25),
    > c_address      varchar(25),
    > c_city      varchar(10),
    > c_nation      varchar(15),
    > c_region      varchar(12),
    > c_phone      varchar(15),
    > c_mktsegment  varchar(10))
    > row format delimited fields
    > terminated by '|' stored as textfile;
OK
Time taken: 0.047 seconds
hive> load data local inpath '/home/ec2-user/customer.tbl'
    > overwrite into table customer;
Loading data to table default.customer
OK
Time taken: 0.261 seconds
hive>
```

Hive Query 2.1 Code:

```
select sum(lo_revenue), d_year, p_brand1
from lineorder, dwdate, part, supplier
where lo_orderdate = d_datekey
and lo_partkey = p_partkey
and lo_suppkey = s_suppkey
and p_category = 'MFGR#12'
and s_region = 'AMERICA'
group by d_year, p_brand1
order by d_year, p_brand1;
```

Hive Query 2.1 Execution (only beginning, start of each map/reduce, and end are shown):



```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
hive> select sum(lo_revenue), d_year, p_brand1
> from lineorder, dwdate, part, supplier
> where lo_orderdate = d_datekey
>   and lo_partkey = p_partkey
>   and lo_suppkey = s_suppkey
>   and p_category = 'MFGR#12'
>   and s_region = 'AMERICA'
> group by d_year, p_brand1
> order by d_year, p_brand1;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20180311034148_d9d36163-30c3-4739-af7b-fa85996c0ced
Total jobs = 6
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar!/_org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/_org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20180311034148_d9d36163-30c3-4739-af7b-fa85996c0ced.log
2018-03-11 03:41:57      Starting to launch local task to process map join;      m
maximum memory = 477102080
2018-03-11 03:41:58      Dump the side-table for tag: 1 with group count: 2556 in
to file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11
_03-41-48_891_3061260632992350640-1/-local-10014/HashTable-Stage-13/MapJoin-mapf
ile31--.hashtable
2018-03-11 03:41:58      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-407
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
2018-03-11 03:41:58      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11_03-41-48_891_3061260632992350640-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile31--.hashtable (67039 bytes)
2018-03-11 03:41:58      End of local task; Time Taken: 1.242 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 6
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0025, Tracking URL = http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0025/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0025
Hadoop job information for Stage-13: number of mappers: 10; number of reducers: 0
2018-03-11 03:42:04,698 Stage-13 map = 0%,  reduce = 0%
2018-03-11 03:42:19,187 Stage-13 map = 5%,  reduce = 0%, Cumulative CPU 33.54 sec
2018-03-11 03:42:20,339 Stage-13 map = 20%,  reduce = 0%, Cumulative CPU 41.27 sec
2018-03-11 03:42:21,432 Stage-13 map = 30%,  reduce = 0%, Cumulative CPU 52.9 sec
2018-03-11 03:42:22,548 Stage-13 map = 35%,  reduce = 0%, Cumulative CPU 60.55 sec
2018-03-11 03:42:23,674 Stage-13 map = 40%,  reduce = 0%, Cumulative CPU 63.44 sec
2018-03-11 03:42:27,889 Stage-13 map = 45%,  reduce = 0%, Cumulative CPU 79.43 sec
2018-03-11 03:42:28,942 Stage-13 map = 55%,  reduce = 0%, Cumulative CPU 82.06 sec
2018-03-11 03:42:31,060 Stage-13 map = 70%,  reduce = 0%, Cumulative CPU 88.57 sec
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
2018-03-11 03:42:31,060 Stage-13 map = 70%,  reduce = 0%, Cumulative CPU 88.57 sec
2018-03-11 03:42:35,413 Stage-13 map = 75%,  reduce = 0%, Cumulative CPU 99.81 sec
2018-03-11 03:42:36,448 Stage-13 map = 95%,  reduce = 0%, Cumulative CPU 105.46 sec
2018-03-11 03:42:38,515 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 108.04 sec
MapReduce Total cumulative CPU time: 1 minutes 48 seconds 40 msec
Ended Job = job_1520720845975_0025
Stage-15 is filtered out by condition resolver.
Stage-16 is filtered out by condition resolver.
Stage-2 is selected by condition resolver.
Launching Job 2 out of 6
Number of reduce tasks not specified. Estimated from input data size: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1520720845975_0026, Tracking URL = http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0026/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0026
Hadoop job information for Stage-2: number of mappers: 5; number of reducers: 4
2018-03-11 03:42:46,921 Stage-2 map = 0%,  reduce = 0%
2018-03-11 03:42:59,294 Stage-2 map = 20%,  reduce = 0%, Cumulative CPU 17.01 sec
2018-03-11 03:43:00,327 Stage-2 map = 40%,  reduce = 0%, Cumulative CPU 30.01 sec
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
east-2.compute.internal:8088/proxy/application_1520720845975_0027/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0027
Hadoop job information for Stage-4: number of mappers: 3; number of reducers: 1
2018-03-11 03:44:26,988 Stage-4 map = 0%, reduce = 0%
2018-03-11 03:44:34,267 Stage-4 map = 67%, reduce = 0%, Cumulative CPU 7.2 sec
2018-03-11 03:44:35,298 Stage-4 map = 100%, reduce = 0%, Cumulative CPU 11.35 s
ec
2018-03-11 03:44:40,448 Stage-4 map = 100%, reduce = 100%, Cumulative CPU 12.61
sec
MapReduce Total cumulative CPU time: 12 seconds 610 msec
Ended Job = job_1520720845975_0027
Launching Job 4 out of 6
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1520720845975_0028, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0028/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0028
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2018-03-11 03:44:46,228 Stage-5 map = 0%, reduce = 0%
2018-03-11 03:44:51,405 Stage-5 map = 100%, reduce = 0%, Cumulative CPU 0.89 se
c
2018-03-11 03:44:57,590 Stage-5 map = 100%, reduce = 100%, Cumulative CPU 2.1 s
ec
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
1457307800    1998    MFGR#1220
1530901152    1998    MFGR#1221
1358086103    1998    MFGR#1222
1409115372    1998    MFGR#1223
1568652498    1998    MFGR#1224
1451906941    1998    MFGR#1225
1640583696    1998    MFGR#1226
1565657860    1998    MFGR#1227
1607890751    1998    MFGR#1228
1350601347    1998    MFGR#1229
1470503353    1998    MFGR#123
1441898473    1998    MFGR#1230
1445039464    1998    MFGR#1231
1710140678    1998    MFGR#1232
1538979218    1998    MFGR#1233
1532309319    1998    MFGR#1234
1598713364    1998    MFGR#1235
1577658136    1998    MFGR#1236
1532687418    1998    MFGR#1237
1285428693    1998    MFGR#1238
1459545128    1998    MFGR#1239
1525737275    1998    MFGR#124
1587370161    1998    MFGR#1240
1477715730    1998    MFGR#125
1466946762    1998    MFGR#126
1686460729    1998    MFGR#127
1538644707    1998    MFGR#128
1207004714    1998    MFGR#129
Time taken: 190.815 seconds, Fetched: 280 row(s)
hive>
```

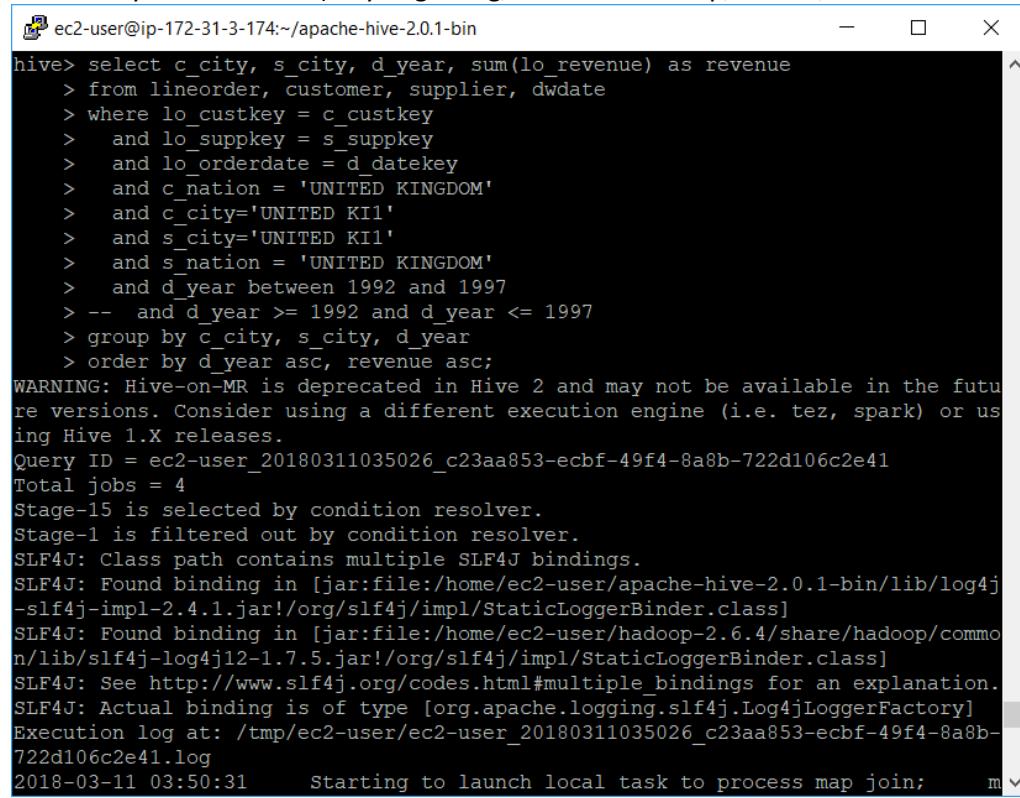
The Hive Query 2.1 took 190.815 secs to run on a 4-node cluster. Note that 4 Map/Reduce passes were executed.

Hive Query 3.3

Hive Query 3.3 Code:

```
select c_city, s_city, d_year, sum(lo_revenue) as revenue
from lineorder, customer, supplier, dwdate
where lo_custkey = c_custkey
and lo_suppkey = s_suppkey
and lo_orderdate = d_datekey
and c_nation = 'UNITED KINGDOM'
and c_city='UNITED KI1'
and s_city='UNITED KI1'
and s_nation = 'UNITED KINGDOM'
and d_year between 1992 and 1997
-- and d_year >= 1992 and d_year <= 1997
group by c_city, s_city, d_year
order by d_year asc, revenue asc;
```

Hive Query 3.3 Execution (only beginning, start of each map/reduce, and end are shown):



```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
hive> select c_city, s_city, d_year, sum(lo_revenue) as revenue
> from lineorder, customer, supplier, dwdate
> where lo_custkey = c_custkey
> and lo_suppkey = s_suppkey
> and lo_orderdate = d_datekey
> and c_nation = 'UNITED KINGDOM'
> and c_city='UNITED KI1'
> and s_city='UNITED KI1'
> and s_nation = 'UNITED KINGDOM'
> and d_year between 1992 and 1997
> -- and d_year >= 1992 and d_year <= 1997
> group by c_city, s_city, d_year
> order by d_year asc, revenue asc;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20180311035026_c23aa853-ecbf-49f4-8a8b-722d106c2e41
Total jobs = 4
Stage-15 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar!/_org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/_org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20180311035026_c23aa853-ecbf-49f4-8a8b-722d106c2e41.log
2018-03-11 03:50:31      Starting to launch local task to process map join;      m v
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
2018-03-11 03:50:31      Starting to launch local task to process map join;    m^
maximum memory = 477102080
2018-03-11 03:50:33      Dump the side-table for tag: 1 with group count: 468 int
o file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11
03-50-26_219_5399002296560110591-1/-local-10010/HashTable-Stage-11/MapJoin-mapfi
le61--.hashtable
2018-03-11 03:50:33      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-407
5-a89d-414f0f513158/hive_2018-03-11_03-50-26_219_5399002296560110591-1/-local-10
010/HashTable-Stage-11/MapJoin-mapfile61--.hashtable (15540 bytes)
2018-03-11 03:50:33      End of local task; Time Taken: 1.624 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 4
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0029, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0029/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0029
Hadoop job information for Stage-11: number of mappers: 10; number of reducers:
0
2018-03-11 03:50:39,454 Stage-11 map = 0%,  reduce = 0%
2018-03-11 03:50:49,173 Stage-11 map = 10%,  reduce = 0%, Cumulative CPU 6.51 se
c
2018-03-11 03:50:51,349 Stage-11 map = 20%,  reduce = 0%, Cumulative CPU 18.27 s
ec
2018-03-11 03:50:52,452 Stage-11 map = 30%,  reduce = 0%, Cumulative CPU 19.79 s
ec
2018-03-11 03:50:53,511 Stage-11 map = 40%,  reduce = 0%, Cumulative CPU 33.04 s
ec
2018-03-11 03:50:56,664 Stage-11 map = 55%,  reduce = 0%, Cumulative CPU 46.7 se^
v
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
o file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11
03-50-26_219_5399002296560110591-1/-local-10008/HashTable-Stage-4/MapJoin-mapfil
e51--.hashtable
2018-03-11 03:51:09      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-407
5-a89d-414f0f513158/hive_2018-03-11_03-50-26_219_5399002296560110591-1/-local-10
008/HashTable-Stage-4/MapJoin-mapfile51--.hashtable (5759 bytes)
2018-03-11 03:51:09      End of local task; Time Taken: 1.656 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 4
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1520720845975_0030, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0030/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0030
Hadoop job information for Stage-4: number of mappers: 3; number of reducers: 1
2018-03-11 03:51:15,923 Stage-4 map = 0%,  reduce = 0%
2018-03-11 03:51:22,183 Stage-4 map = 33%,  reduce = 0%, Cumulative CPU 2.38 sec
2018-03-11 03:51:23,212 Stage-4 map = 67%,  reduce = 0%, Cumulative CPU 4.23 sec
2018-03-11 03:51:24,241 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 6.54 se
c
2018-03-11 03:51:27,367 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 6.54
sec
MapReduce Total cumulative CPU time: 6 seconds 540 msec
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
set mapreduce.job.reduces=<number>
Starting Job = job_1520720845975_0031, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0031/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0031
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2018-03-11 03:51:35,051 Stage-5 map = 0%,  reduce = 0%
2018-03-11 03:51:40,282 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 0.83 se
c
2018-03-11 03:51:46,525 Stage-5 map = 100%,  reduce = 100%, Cumulative CPU 2.01
sec
MapReduce Total cumulative CPU time: 2 seconds 10 msec
Ended Job = job_1520720845975_0031
MapReduce Jobs Launched:
Stage-Stage-11: Map: 10  Cumulative CPU: 62.08 sec  HDFS Read: 2417923371 HDFS
Write: 3774639 SUCCESS
Stage-Stage-4: Map: 3  Reduce: 1  Cumulative CPU: 7.69 sec  HDFS Read: 3813443
HDFS Write: 378 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 1  Cumulative CPU: 2.01 sec  HDFS Read: 6979 HD
FS Write: 222 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 11 seconds 780 msec
OK
UNITED KI1      UNITED KI1      1992      160175326
UNITED KI1      UNITED KI1      1993      280314496
UNITED KI1      UNITED KI1      1994      173023390
UNITED KI1      UNITED KI1      1995      232484022
UNITED KI1      UNITED KI1      1996      180373857
UNITED KI1      UNITED KI1      1997      232378038
Time taken: 81.364 seconds, Fetched: 6 row(s)
hive>
```

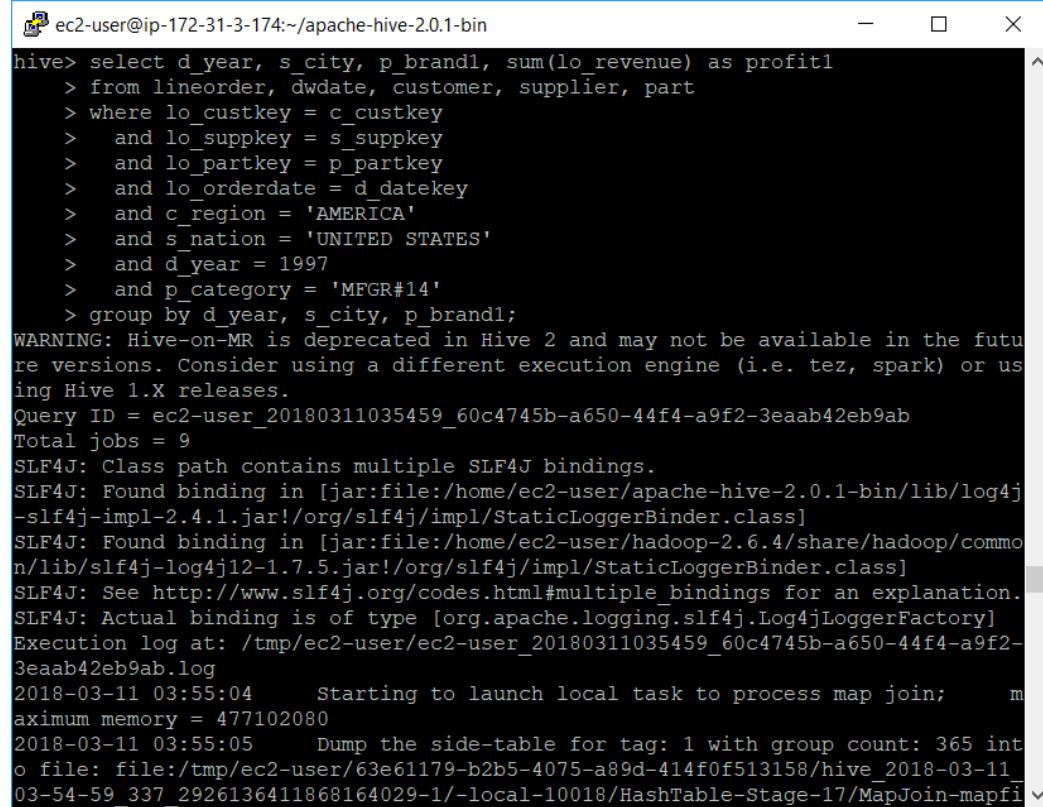
The Hive Query 3.3 took 81.364 secs to run on a 4-node cluster. Note that 2 Map/Reduce passes were executed.

Hive Query 4.3

Hive Query 4.3 Code:

```
select d_year, s_city, p_brand1, sum(lo_revenue) as profit1
from lineorder, dwdate, customer, supplier, part
where lo_custkey = c_custkey
and lo_suppkey = s_suppkey
and lo_partkey = p_partkey
and lo_orderdate = d_datekey
and c_region = 'AMERICA'
and s_nation = 'UNITED STATES'
and d_year = 1997
and p_category = 'MFGR#14'
group by d_year, s_city, p_brand1;
```

Hive Query 4.3 Execution (only beginning, start of each map/reduce, and end are shown):



The screenshot shows a terminal window titled "ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin". The window displays the execution of a Hive query. The query is identical to the one shown above. The output includes a warning about Hive-on-MR being deprecated, the query ID, total jobs (9), and various log messages from SLF4J and Log4j. It also shows the start of map and reduce tasks, and the final output file path.

```
hive> select d_year, s_city, p_brand1, sum(lo_revenue) as profit1
> from lineorder, dwdate, customer, supplier, part
> where lo_custkey = c_custkey
>   and lo_suppkey = s_suppkey
>   and lo_partkey = p_partkey
>   and lo_orderdate = d_datekey
>   and c_region = 'AMERICA'
>   and s_nation = 'UNITED STATES'
>   and d_year = 1997
>   and p_category = 'MFGR#14'
> group by d_year, s_city, p_brand1;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20180311035459_60c4745b-a650-44f4-a9f2-3eaab42eb9ab
Total jobs = 9
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20180311035459_60c4745b-a650-44f4-a9f2-3eaab42eb9ab.log
2018-03-11 03:55:04      Starting to launch local task to process map join;      m
aximum memory = 477102080
2018-03-11 03:55:05      Dump the side-table for tag: 1 with group count: 365 int
o file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11_03-54-59_337_2926136411868164029-1/-local-10018/HashTable-Stage-17/MapJoin-mapfi
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
03-54-59_337_2926136411868164029-1/-local-10018/HashTable-Stage-17/MapJoin-mapfile1e131--.hashtable
2018-03-11 03:55:05      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11_03-54-59_337_2926136411868164029-1/-local-10018/HashTable-Stage-17/MapJoin-mapfile1e131--.hashtable (8328 bytes)
2018-03-11 03:55:05      End of local task; Time Taken: 1.172 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 9
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0032, Tracking URL = http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0032/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0032
Hadoop job information for Stage-17: number of mappers: 10; number of reducers: 0
2018-03-11 03:55:12,321 Stage-17 map = 0%,  reduce = 0%
2018-03-11 03:55:23,015 Stage-17 map = 5%,  reduce = 0%, Cumulative CPU 7.24 sec
2018-03-11 03:55:24,110 Stage-17 map = 20%,  reduce = 0%, Cumulative CPU 10.23 sec
2018-03-11 03:55:29,367 Stage-17 map = 40%,  reduce = 0%, Cumulative CPU 51.29 sec
2018-03-11 03:55:30,432 Stage-17 map = 70%,  reduce = 0%, Cumulative CPU 60.54 sec
2018-03-11 03:55:32,550 Stage-17 map = 90%,  reduce = 0%, Cumulative CPU 71.57 sec
2018-03-11 03:55:33,579 Stage-17 map = 100%,  reduce = 0%, Cumulative CPU 72.29 sec
MapReduce Total cumulative CPU time: 1 minutes 12 seconds 290 msec
Ended Job = job_1520720845975_0032
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520720845975_0033
Hadoop job information for Stage-14: number of mappers: 3; number of reducers: 0
2018-03-11 03:55:47,133 Stage-14 map = 0%,  reduce = 0%
2018-03-11 03:55:55,361 Stage-14 map = 33%,  reduce = 0%, Cumulative CPU 4.54 sec
2018-03-11 03:55:57,425 Stage-14 map = 67%,  reduce = 0%, Cumulative CPU 4.54 sec
2018-03-11 03:55:59,498 Stage-14 map = 100%,  reduce = 0%, Cumulative CPU 15.54 sec
MapReduce Total cumulative CPU time: 15 seconds 540 msec
Ended Job = job_1520720845975_0033
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20180311035459_60c4745b-a650-44f4-a9f2-3eaab42eb9ab.log
2018-03-11 03:56:06      Starting to launch local task to process map join; maximum memory = 477102080
2018-03-11 03:56:08      Dump the side-table for tag: 1 with group count: 1598 into file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11_03-54-59_337_2926136411868164029-1/-local-10012/HashTable-Stage-13/MapJoin-mapfile1e101--.hashtable
2018-03-11 03:56:08      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11_03-54-59_337_2926136411868164029-1/-local-10012/HashTable-Stage-13/MapJoin-mapfile1e101--.hashtable (51624 bytes)
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
012/HashTable-Stage-13/MapJoin-mapfile101--.hashtable (51624 bytes)
2018-03-11 03:56:08      End of local task; Time Taken: 1.559 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 4 out of 9
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0034, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0034/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0034
Hadoop job information for Stage-13: number of mappers: 1; number of reducers: 0
2018-03-11 03:56:13,489 Stage-13 map = 0%,  reduce = 0%
2018-03-11 03:56:21,727 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 4.28 s
ec
MapReduce Total cumulative CPU time: 4 seconds 280 msec
Ended Job = job_1520720845975_0034
Stage-18 is filtered out by condition resolver.
Stage-19 is selected by condition resolver.
Stage-4 is filtered out by condition resolver.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-
-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/commo
n/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/ec2-user/ec2-user_20180311035459_60c4745b-a650-44f4-a9f2-
3aab42eb9ab.log
2018-03-11 03:56:27      Starting to launch local task to process map join;      m
aximum memory = 477102080
```

```
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
maximum memory = 477102080
2018-03-11 03:56:29      Dump the side-table for tag: 0 with group count: 27877 i
nto file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-1
1_03-54-59_337_2926136411868164029-1/-local-10010/HashTable-Stage-11/MapJoin-map
file90--.hashtable
2018-03-11 03:56:29      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-407
5-a89d-414f0f513158/hive_2018-03-11_03-54-59_337_2926136411868164029-1/-local-10
010/HashTable-Stage-11/MapJoin-mapfile90--.hashtable (1052926 bytes)
2018-03-11 03:56:29      End of local task; Time Taken: 1.658 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 6 out of 9
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0035, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0035/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0035
Hadoop job information for Stage-11: number of mappers: 1; number of reducers: 0
2018-03-11 03:56:34,748 Stage-11 map = 0%,  reduce = 0%
2018-03-11 03:56:42,985 Stage-11 map = 100%,  reduce = 0%, Cumulative CPU 4.0 se
c
MapReduce Total cumulative CPU time: 4 seconds 0 msec
Ended Job = job_1520720845975_0035
Launching Job 7 out of 9
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
```

```

ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1520720845975_0036, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0036/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0036
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2018-03-11 03:56:49,632 Stage-5 map = 0%,  reduce = 0%
2018-03-11 03:56:54,851 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 0.93 se
c
2018-03-11 03:57:01,032 Stage-5 map = 100%,  reduce = 100%, Cumulative CPU 2.19
sec
MapReduce Total cumulative CPU time: 2 seconds 190 msec
Ended Job = job_1520720845975_0036
MapReduce Jobs Launched:
Stage-Stage-17: Map: 10  Cumulative CPU: 72.29 sec  HDFS Read: 2417930601 HDFS
Write: 114944754 SUCCESS
Stage-Stage-14: Map: 3  Cumulative CPU: 15.54 sec  HDFS Read: 114962928 HDFS W
rite: 20535906 SUCCESS
Stage-Stage-13: Map: 1  Cumulative CPU: 4.28 sec  HDFS Read: 20541826 HDFS Wri
te: 1046664 SUCCESS
Stage-Stage-11: Map: 1  Cumulative CPU: 4.0 sec  HDFS Read: 51048588 HDFS Writ
e: 17528 SUCCESS
Stage-Stage-5: Map: 1 Reduce: 1  Cumulative CPU: 2.19 sec  HDFS Read: 24504 H
DFS Write: 13149 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 38 seconds 300 msec
OK
1997    UNITED ST0      MFGR#141        8701440
1997    UNITED ST0      MFGR#1410       19458985
1997    UNITED ST0      MFGR#1411       17945623

ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
1997    UNITED ST9      MFGR#1420       2413684
1997    UNITED ST9      MFGR#1421       2550278
1997    UNITED ST9      MFGR#1422       9591206
1997    UNITED ST9      MFGR#1423       26867073
1997    UNITED ST9      MFGR#1424       10087806
1997    UNITED ST9      MFGR#1425       7433355
1997    UNITED ST9      MFGR#1426       7106043
1997    UNITED ST9      MFGR#1427       9204537
1997    UNITED ST9      MFGR#1428       7214318
1997    UNITED ST9      MFGR#1429       18628842
1997    UNITED ST9      MFGR#143        13450938
1997    UNITED ST9      MFGR#1430       8471957
1997    UNITED ST9      MFGR#1431       17856795
1997    UNITED ST9      MFGR#1432       17284859
1997    UNITED ST9      MFGR#1433       11883991
1997    UNITED ST9      MFGR#1434       15821309
1997    UNITED ST9      MFGR#1435       19620964
1997    UNITED ST9      MFGR#1436       20504919
1997    UNITED ST9      MFGR#1437       5704978
1997    UNITED ST9      MFGR#1438       19864432
1997    UNITED ST9      MFGR#1439       1990404
1997    UNITED ST9      MFGR#144        2664223
1997    UNITED ST9      MFGR#1440       1140522
1997    UNITED ST9      MFGR#145        18030070
1997    UNITED ST9      MFGR#146        4855718
1997    UNITED ST9      MFGR#147        6612886
1997    UNITED ST9      MFGR#148        6820646
1997    UNITED ST9      MFGR#149        19862295
Time taken: 122.753 seconds, Fetched: 384 row(s)
hive>

```

The Hive Query 4.3 took 122.753 secs to run on a 4-node cluster. Note that 5 Map/Reduce passes were executed.

B)

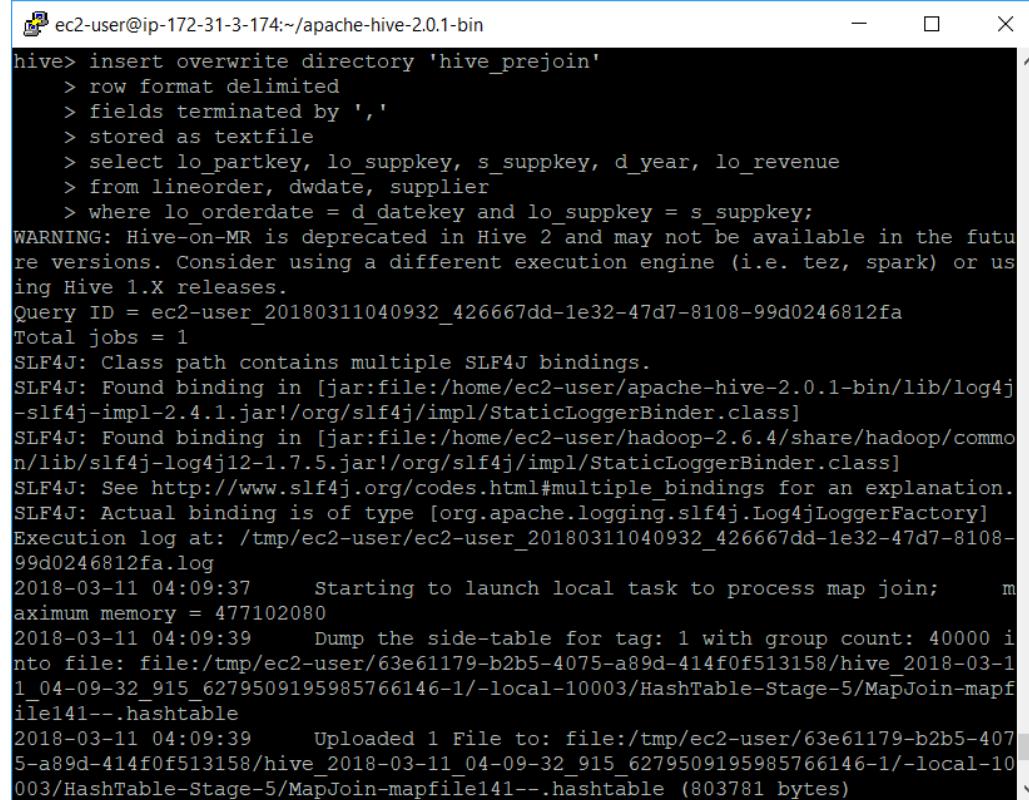
Note that since no delimiter was specified to be used for the pre-join file, I used a comma delimiter.

Hive Pre-Join

Hive Pre-Join Code:

```
insert overwrite directory 'hive_prejoin'
row format delimited
fields terminated by ','
stored as textfile
select lo_partkey, lo_suppkey, s_suppkey, d_year, lo_revenue
from lineorder, dwdate, supplier
where lo_orderdate = d_datekey and lo_suppkey = s_suppkey;
```

Hive Pre-Join Execution:



The screenshot shows a terminal window titled "ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin". The window displays the execution of a Hive query to create a pre-join file. The command is:

```
hive> insert overwrite directory 'hive_prejoin'
> row format delimited
> fields terminated by ','
> stored as textfile
> select lo_partkey, lo_suppkey, s_suppkey, d_year, lo_revenue
> from lineorder, dwdate, supplier
> where lo_orderdate = d_datekey and lo_suppkey = s_suppkey;
```

Output messages include:

- WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
- Query ID = ec2-user_20180311040932_426667dd-1e32-47d7-8108-99d0246812fa
- Total jobs = 1
- SLF4J: Class path contains multiple SLF4J bindings.
- SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
- SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
- SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
- SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
- Execution log at: /tmp/ec2-user/ec2-user_20180311040932_426667dd-1e32-47d7-8108-99d0246812fa.log
- 2018-03-11 04:09:37 Starting to launch local task to process map join; maximum memory = 477102080
- 2018-03-11 04:09:39 Dump the side-table for tag: 1 with group count: 40000 into file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11_04-09-32_915_6279509195985766146-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile141--.hashtable
- 2018-03-11 04:09:39 Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11_04-09-32_915_6279509195985766146-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile141--.hashtable (803781 bytes)

```

ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
003/HashTable-Stage-5/MapJoin-mapfile141--.hashtable (803781 bytes)
2018-03-11 04:09:39      Dump the side-table for tag: 1 with group count: 2556 in
to file: file:/tmp/ec2-user/63e61179-b2b5-4075-a89d-414f0f513158/hive_2018-03-11
_04-09-32_915_6279509195985766146-1/-local-10003/HashTable-Stage-5/MapJoin-mapfi
le151--.hashtable
2018-03-11 04:09:39      Uploaded 1 File to: file:/tmp/ec2-user/63e61179-b2b5-407
5-a89d-414f0f513158/hive_2018-03-11_04-09-32_915_6279509195985766146-1/-local-10
003/HashTable-Stage-5/MapJoin-mapfile151--.hashtable (67039 bytes)
2018-03-11 04:09:39      End of local task; Time Taken: 1.562 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520720845975_0037, Tracking URL = http://ip-172-31-3-174.us-
east-2.compute.internal:8088/proxy/application_1520720845975_0037/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_15207208459
75_0037
Hadoop job information for Stage-5: number of mappers: 10; number of reducers: 0
2018-03-11 04:09:45,031 Stage-5 map = 0%, reduce = 0%
2018-03-11 04:09:54,628 Stage-5 map = 10%, reduce = 0%, Cumulative CPU 2.97 sec
2018-03-11 04:10:00,073 Stage-5 map = 20%, reduce = 0%, Cumulative CPU 52.87 se
c
2018-03-11 04:10:03,233 Stage-5 map = 25%, reduce = 0%, Cumulative CPU 72.25 se
c
2018-03-11 04:10:06,445 Stage-5 map = 30%, reduce = 0%, Cumulative CPU 91.72 se
c
2018-03-11 04:10:07,518 Stage-5 map = 40%, reduce = 0%, Cumulative CPU 92.84 se
c
2018-03-11 04:10:08,594 Stage-5 map = 70%, reduce = 0%, Cumulative CPU 104.07 s
ec

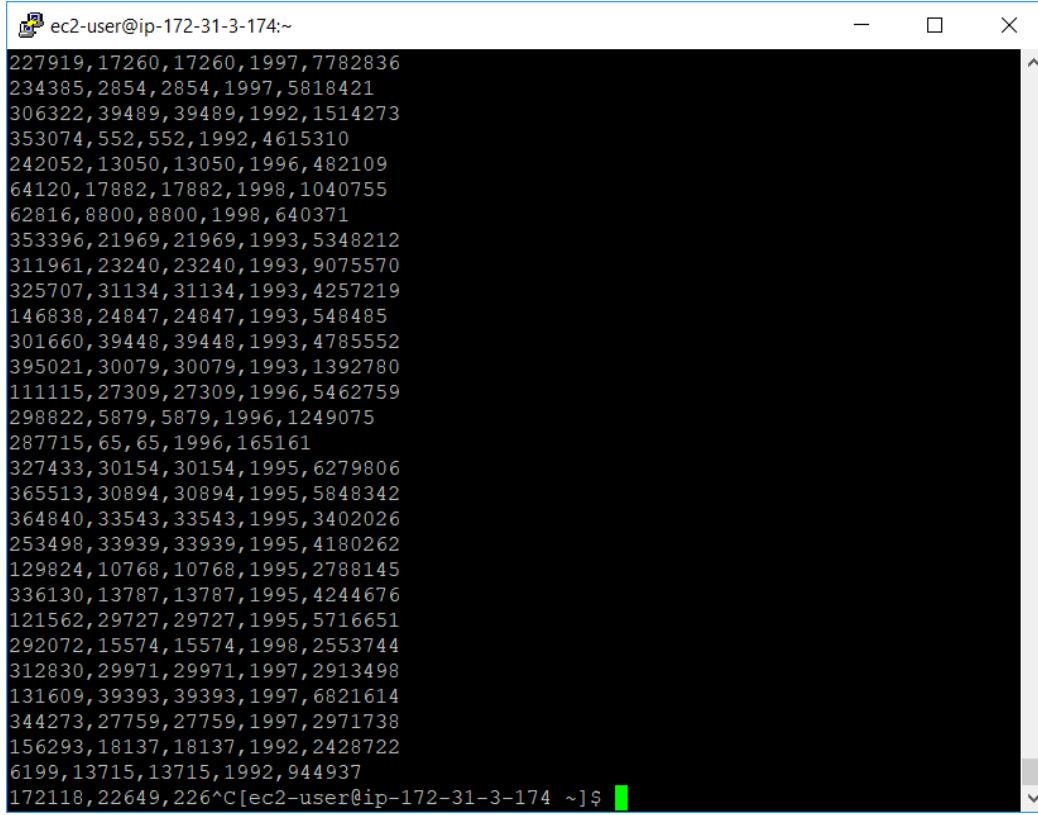
ec2-user@ip-172-31-3-174:~/apache-hive-2.0.1-bin
2018-03-11 04:09:45,031 Stage-5 map = 0%, reduce = 0%
2018-03-11 04:09:54,628 Stage-5 map = 10%, reduce = 0%, Cumulative CPU 2.97 sec
2018-03-11 04:10:00,073 Stage-5 map = 20%, reduce = 0%, Cumulative CPU 52.87 se
c
2018-03-11 04:10:03,233 Stage-5 map = 25%, reduce = 0%, Cumulative CPU 72.25 se
c
2018-03-11 04:10:06,445 Stage-5 map = 30%, reduce = 0%, Cumulative CPU 91.72 se
c
2018-03-11 04:10:07,518 Stage-5 map = 40%, reduce = 0%, Cumulative CPU 92.84 se
c
2018-03-11 04:10:08,594 Stage-5 map = 70%, reduce = 0%, Cumulative CPU 104.07 s
ec
2018-03-11 04:10:15,989 Stage-5 map = 75%, reduce = 0%, Cumulative CPU 127.73 s
ec
2018-03-11 04:10:17,055 Stage-5 map = 85%, reduce = 0%, Cumulative CPU 131.42 s
ec
2018-03-11 04:10:19,252 Stage-5 map = 95%, reduce = 0%, Cumulative CPU 136.27 s
ec
2018-03-11 04:10:20,278 Stage-5 map = 100%, reduce = 0%, Cumulative CPU 137.23
sec
MapReduce Total cumulative CPU time: 2 minutes 17 seconds 230 msec
Ended Job = job_1520720845975_0037
Moving data to: hive_prejoin
MapReduce Jobs Launched:
Stage-Stage-5: Map: 10   Cumulative CPU: 137.23 sec   HDFS Read: 2417956971 HDFS
Write: 744597239 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 17 seconds 230 msec
OK
Time taken: 48.47 seconds
hive>

```

The Hive Pre-Join code took 48.47 secs to execute on a 4-node cluster.

Hive Pre-Join File Output Size and File Contents:

```
ec2-user@ip-172-31-3-174:~$ vielens
-rw-r--r--  2 ec2-user supergroup  51039483 2018-03-11 00:12 /user/ec2-user/part.tbl
drwxr-xr-x - ec2-user supergroup      0 2018-03-05 00:56 /user/ec2-user/recommendations
-rw-r--r--  2 ec2-user supergroup  3344696 2018-03-11 00:12 /user/ec2-user/supplier.tbl
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -ls /user/ec2-user/hive_prejoin/
Found 10 items
-rwxr-xr-x  2 ec2-user supergroup  82895307 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000000_0
-rwxr-xr-x  2 ec2-user supergroup  83110590 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000001_0
-rwxr-xr-x  2 ec2-user supergroup  83049269 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000002_0
-rwxr-xr-x  2 ec2-user supergroup  83114382 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000003_0
-rwxr-xr-x  2 ec2-user supergroup  82291942 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000004_0
-rwxr-xr-x  2 ec2-user supergroup  82695897 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000005_0
-rwxr-xr-x  2 ec2-user supergroup  82295923 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000006_0
-rwxr-xr-x  2 ec2-user supergroup  82292046 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000007_0
-rwxr-xr-x  2 ec2-user supergroup  82288827 2018-03-11 04:10 /user/ec2-user/hive_prejoin/000008_0
-rwxr-xr-x  2 ec2-user supergroup  563056 2018-03-11 04:09 /user/ec2-user/hive_prejoin/000009_0
[ec2-user@ip-172-31-3-174 ~]$ 
ec2-user@ip-172-31-3-174:~$ 
-rw-r--r--  2 ec2-user supergroup  288374 2018-03-13 03:22 /user/ec2-user/synthetic_control.data
drwxr-xr-x - ec2-user supergroup      0 2018-03-13 03:43 /user/ec2-user/testdata
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -ls /user/ec2-user
Found 23 items
drwxr-xr-x - ec2-user supergroup      0 2018-03-05 00:51 /user/ec2-user/als
-rw-r--r--  2 ec2-user supergroup  11279031 2018-03-11 00:11 /user/ec2-user/customer.tbl
drwxr-xr-x - ec2-user supergroup      0 2018-03-05 00:37 /user/ec2-user/dataset
-rw-r--r--  2 ec2-user supergroup  227409 2018-03-11 00:11 /user/ec2-user/dwdate.tbl
drwxr-xr-x - ec2-user supergroup      0 2018-03-11 04:10 /user/ec2-user/hive_prejoin
-rw-r--r--  2 ec2-user supergroup 2417756563 2018-03-11 00:11 /user/ec2-user/lineorder.tbl
drwxr-xr-x - ec2-user supergroup      0 2018-03-11 01:18 /user/ec2-user/lo_hive_3col
drwxr-xr-x - ec2-user supergroup      0 2018-03-11 01:11 /user/ec2-user/lo_hive_csv
drwxr-xr-x - ec2-user supergroup      0 2018-03-11 03:20 /user/ec2-user/lo_pig_3col
drwxr-xr-x - ec2-user supergroup      0 2018-03-11 01:26 /user/ec2-user/lo_pig_csv
-rw-r--r--  2 ec2-user supergroup  9500 2018-03-11 00:15 /user/ec2-user/lo_sample.tbl
drwxr-xr-x - ec2-user supergroup      0 2018-03-05 00:37 /user/ec2-user/ml_dataset
drwxr-xr-x - ec2-user supergroup      0 2018-03-05 00:36 /user/ec2-user/movielens
drwxr-xr-x - ec2-user supergroup      0 2018-03-13 03:35 /user/ec2-user/output
drwxr-xr-x - ec2-user supergroup      0 2018-03-13 23:10 /user/ec2-user/output11
drwxr-xr-x - ec2-user supergroup      0 2018-03-13 03:44 /user/ec2-user/output11_orig
-rw-r--r--  2 ec2-user supergroup  89519809 2018-03-13 03:23 /user/ec2-user/part-00000
-rw-r--r--  2 ec2-user supergroup  51039483 2018-03-11 00:12 /user/ec2-user/part.tbl
drwxr-xr-x - ec2-user supergroup      0 2018-03-11 04:45 /user/ec2-user/pig_prejoin
drwxr-xr-x - ec2-user supergroup      0 2018-03-05 00:56 /user/ec2-user/recommendations
-rw-r--r--  2 ec2-user supergroup  3344696 2018-03-11 00:12 /user/ec2-user/supplier.tbl
-rw-r--r--  2 ec2-user supergroup  288374 2018-03-13 03:22 /user/ec2-user/synthetic_control.data
drwxr-xr-x - ec2-user supergroup      0 2018-03-13 03:43 /user/ec2-user/testdata
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -dus /user/ec2-user/hive_prejoin
dus: DEPRECATED: Please use 'du -s' instead.
744597239 /user/ec2-user/hive_prejoin
[ec2-user@ip-172-31-3-174 ~]$ 
```



A screenshot of a terminal window titled "ec2-user@ip-172-31-3-174:~". The window contains a large list of numerical values, each consisting of two parts separated by a comma. The first part is a 5-digit number, and the second part is a 7-digit number. The list is as follows:

```
227919,17260,17260,1997,7782836
234385,2854,2854,1997,5818421
306322,39489,39489,1992,1514273
353074,552,552,1992,4615310
242052,13050,13050,1996,482109
64120,17882,17882,1998,1040755
62816,8800,8800,1998,640371
353396,21969,21969,1993,5348212
311961,23240,23240,1993,9075570
325707,31134,31134,1993,4257219
146838,24847,24847,1993,548485
301660,39448,39448,1993,4785552
395021,30079,30079,1993,1392780
111115,27309,27309,1996,5462759
298822,5879,5879,1996,1249075
287715,65,65,1996,165161
327433,30154,30154,1995,6279806
365513,30894,30894,1995,5848342
364840,33543,33543,1995,3402026
253498,33939,33939,1995,4180262
129824,10768,10768,1995,2788145
336130,13787,13787,1995,4244676
121562,29727,29727,1995,5716651
292072,15574,15574,1998,2553744
312830,29971,29971,1997,2913498
131609,39393,39393,1997,6821614
344273,27759,27759,1997,2971738
156293,18137,18137,1992,2428722
6199,13715,13715,1992,944937
172118,22649,22649,1997,2428722
```

The size of the Hive pre-join output directory is 744597239 bytes or approximately 744,597 MB.

Pig Pre-Join

Pig Pre-Join Code:

```
lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_suppkey:INT,
lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARRAY, lo_quantity:INT,
lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:INT, lo_revenue:INT, lo_supplycost:INT,
lo_tax:INT, lo_commitdate:INT, lo_shipmode:CHARARRAY);

dwdate = LOAD '/user/ec2-user/dwdate.tbl' USING PigStorage('|')
AS (d_datekey:INT, d_date:CHARARRAY, d_dayofweek:CHARARRAY, d_month:CHARARRAY, d_year:INT,
d_yeарmonthnum:INT, d_yeарmonth:CHARARRAY, d_daynuminweek:INT, d_daynuminmonth:INT,
d_daynuminyear:INT, d_monthnuminyear:INT, d_weeknuminyear:INT, d_sellingseason:CHARARRAY,
d_lastdayinweekfl:CHARARRAY, d_lastdayinmonthfl:CHARARRAY, d_holidayfl:CHARARRAY,
d_weekdayfl:CHARARRAY);

supplier = LOAD '/user/ec2-user/supplier.tbl' USING PigStorage('|')
AS (s_suppkey:INT, s_name:CHARARRAY, s_address:CHARARRAY, s_city:CHARARRAY,
s_nation:CHARARRAY, s_region:CHARARRAY, s_phone:CHARARRAY);

first_join = JOIN lineorder BY lo_orderdate, dwdate BY d_datekey;
lo_and_date = FOREACH first_join GENERATE lineorder:::lo_partkey AS lo_partkey, lineorder:::lo_suppkey
AS lo_suppkey, lineorder:::lo_revenue AS lo_revenue, dwdate:::d_year AS d_year;

second_join = JOIN lo_and_date BY lo_suppkey, supplier BY s_suppkey;
supp_lo_date = FOREACH second_join GENERATE lo_and_date:::lo_partkey AS lo_partkey,
lo_and_date:::lo_suppkey AS lo_suppkey, lo_and_date:::lo_revenue AS lo_revenue, lo_and_date:::d_year
AS d_year, supplier:::s_suppkey AS s_suppkey;

store supp_lo_date into 'pig_prejoin' using PigStorage(',');
```

Pig Pre-Join Execution (start, map/reduce beginning, and end shown due to length):

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
grunt> lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
>> AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_suppkey:INT, lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARRAY, lo_quantity:INT, lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:INT, lo_revenue:INT, lo_supplycost:INT, lo_tax:INT, lo_commitdate:INT, lo_shipmode:CHARARRAY);
2018-03-11 04:40:52,999 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> dwdate = LOAD '/user/ec2-user/dwdate.tbl' USING PigStorage('|')
>> AS (d_datekey:INT, d_date:CHARARRAY, d_dayofweek:CHARARRAY, d_month:CHARARRAY, d_year:INT, d_yeарmonthnum:INT, d_yeарmonth:CHARARRAY, d_daynuminweek:INT, d_dаynuminmonth:INT, d_daynuminyear:INT, d_monthnuminyear:INT, d_weeknuminyear:INT, d_sellingseason:CHARARRAY, d_lastdayinweekfl:CHARARRAY, d_lastdayinmonthfl:CHARARRAY, d_holidayfl:CHARARRAY, d_weekdayfl:CHARARRAY);
2018-03-11 04:40:58,392 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> supplier = LOAD '/user/ec2-user/supplier.tbl' USING PigStorage('|')
>> AS (s_suppkey:INT, s_name:CHARARRAY, s_address:CHARARRAY, s_city:CHARARRAY, s_nation:CHARARRAY, s_region:CHARARRAY, s_phone:CHARARRAY);
2018-03-11 04:41:03,545 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> first_join = JOIN lineorder BY lo_orderdate, dwdate BY d_datekey;
grunt> lo_and_date = FOREACH first_join GENERATE lineorder:::lo_partkey AS lo_partkey, lineorder:::lo_suppkey AS lo_suppkey, lineorder:::lo_revenue AS lo_revenue, dwdate:::d_year AS d_year;
grunt> second_join = JOIN lo_and_date BY lo_suppkey, supplier BY s_suppkey;
grunt> supp_lo_date = FOREACH second_join GENERATE lo_and_date:::lo_partkey AS lo_partkey, lo_and_date:::lo_suppkey AS lo_suppkey, lo_and_date:::lo_revenue AS lo_revenue, lo_and_date:::d_year AS d_year, supplier:::s_suppkey AS s_suppkey;
grunt> store supp_lo_date into 'pig_prejoin' using PigStorage(',');
-
```

```
ec2-user@ip-172-31-3-174:~/pig-0.15.0
grunt> store supp_lo_date into 'pig_prejoin' using PigStorage(',');
2018-03-11 04:41:20,773 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-11 04:41:20,819 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2018-03-11 04:41:20,853 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2018-03-11 04:41:20,876 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-11 04:41:20,880 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-03-11 04:41:20,910 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-11 04:41:20,939 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for supplier: $1, $2, $3, $4, $5, $6
2018-03-11 04:41:20,942 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for lineorder: $0, $1, $2, $6, $7, $8, $9, $10, $11, $12, $13, $14, $15, $16
2018-03-11 04:41:20,943 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for dwdate: $1, $2, $3, $5, $6, $7, $8, $9, $10, $11, $12, $13, $14, $15, $16
2018-03-11 04:41:21,035 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-11 04:41:21,094 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->P
```

```

ec2-user@ip-172-31-3-174:~/pig-0.15.0
2018-03-11 04:41:21,805 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2018-03-11 04:41:21,809 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 18
2018-03-11 04:41:21,862 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:19
2018-03-11 04:41:22,041 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1520720845975_0038
2018-03-11 04:41:22,188 [JobControl] INFO org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2018-03-11 04:41:22,246 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1520720845975_0038
2018-03-11 04:41:22,279 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520720845975_0038/
2018-03-11 04:41:22,280 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1520720845975_0038
2018-03-11 04:41:22,280 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases dwd date, first_join, lineorder, lo_and_date
2018-03-11 04:41:22,280 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: lineorder[1,12],lineorder[-1,-1],first_join[7,13],dwd date[3,9],dwd date[-1,-1],first_join[7,13] C: R: lo_and_date[8,14]
2018-03-11 04:41:22,287 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2018-03-11 04:41:22,287 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1520720845975_0038]
2018-03-11 04:41:52,438 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 4% complete

ec2-user@ip-172-31-3-174:~/pig-0.15.0
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,385 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,400 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,466 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-11 04:45:23,484 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.4 0.15.0 ec2-user 2018-03-11 04:41:21 2018-03-11 04:45:23 HASH_JOIN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime
pTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_1520720845975_0038 19 3 44 19 41 44 52 5
0 51 51 dwd date, first_join, lineorder, lo_and_date HASH_JOIN
job_1520720845975_0039 5 1 60 7 40 47 135 1
35 135 135 second_join, supp_lo_date, supplier HASH_JOIN
dfs://172.31.3.174/user/ec2-user/pig_prejoin

Input(s):
Successfully read 2556 records from: "/user/ec2-user/dwd date.tbl"

```

```

ec2-user@ip-172-31-3-174:~/pig-0.15.0
Input(s):
Successfully read 2556 records from: "/user/ec2-user/dwdate.tbl"
Successfully read 23996604 records from: "/user/ec2-user/lineorder.tbl"
Successfully read 40000 records from: "/user/ec2-user/supplier.tbl"

Output(s):
Successfully stored 23996604 records (744597239 bytes) in: "hdfs://172.31.3.174/
user/ec2-user/pig_prejoin"

Counters:
Total records written : 23996604
Total bytes written : 744597239
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520720845975_0038 -> job_1520720845975_0039,
job_1520720845975_0039

2018-03-11 04:45:23,485 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,498 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,591 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,595 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,591 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,595 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,630 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,634 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,669 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,672 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,716 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,726 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,764 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.3.174:8032
2018-03-11 04:45:23,771 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-11 04:45:23,799 [main] INFO org.apache.pig.backend.hadoop.executionengi ne.mapReduceLayer.MapReduceLauncher - Success!
grunt>

```

The Pig Pre-Join code started at 4:41:21 and ended at 4:45:23, so it took 4 min 4 secs (or 244 secs) to execute on a 4-node cluster.

Pig Pre-Join File Output Size and File Contents:

```
ec2-user@ip-172-31-3-174:~  
ls  
neorder.tbl  
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 01:18 /user/ec2-user/lo  
_hive_3col  
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 01:11 /user/ec2-user/lo  
_hive_csv  
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 03:20 /user/ec2-user/lo  
_pig_3col  
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 01:26 /user/ec2-user/lo  
_pig_csv  
-rw-r--r-- 2 ec2-user supergroup 9500 2018-03-11 00:15 /user/ec2-user/lo  
_sample.tbl  
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:37 /user/ec2-user/ml  
_dataset  
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:36 /user/ec2-user/ml  
vielens  
-rw-r--r-- 2 ec2-user supergroup 51039483 2018-03-11 00:12 /user/ec2-user/pa  
rt.tbl  
drwxr-xr-x - ec2-user supergroup 0 2018-03-11 04:45 /user/ec2-user/pi  
g_prejoin  
drwxr-xr-x - ec2-user supergroup 0 2018-03-05 00:56 /user/ec2-user/re  
commendations  
-rw-r--r-- 2 ec2-user supergroup 3344696 2018-03-11 00:12 /user/ec2-user/su  
pplier.tbl  
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -ls /user/ec2-user/pig_prejoin/  
Found 2 items  
-rw-r--r-- 2 ec2-user supergroup 0 2018-03-11 04:45 /user/ec2-user/pi  
g_prejoin/_SUCCESS  
-rw-r--r-- 2 ec2-user supergroup 744597239 2018-03-11 04:45 /user/ec2-user/pi  
g_prejoin/part-r-00000  
[ec2-user@ip-172-31-3-174 ~]$  
  
ec2-user@ip-172-31-3-174:~  
cat  
183187,81,3080186,1994,81  
300128,81,1326657,1993,81  
379095,81,1296184,1994,81  
137920,81,5991235,1994,81  
321863,81,3223093,1995,81  
385459,81,1596950,1994,81  
132318,81,4060382,1997,81  
285455,81,1875452,1996,81  
150712,81,3014234,1996,81  
356452,81,6938824,1995,81  
127604,81,7208408,1997,81  
111812,81,5777830,1996,81  
238772,81,1950266,1997,81  
162892,81,6451137,1995,81  
227947,81,4640451,1996,81  
294737,81,2961241,1996,81  
323279,81,3436664,1996,81  
311692,81,7601820,1997,81  
85066,81,191292,1997,81  
91493,81,5217982,1996,81  
162448,81,3439271,1997,81  
270530,81,3601248,1996,81  
189001,81,3670030,1997,81  
262092,81,4396567,1995,81  
96649,81,3330775,1995,81  
18803,81,7761874,1995,81  
23878,81,1982057,1997,81  
121176,81,1185198,1995,81  
114537,81,2558472,1995,81  
71047,8^C[ec2-user@ip-172-31-3-174 ~]$
```

The size of the Pig pre-join output file is 744597239 bytes or approximately 744,597 MB. This matches the size of the Hive pre-join output directory.

Part 3

A)

Mahout Synthetic Clustering with Number of Clusters = 11

I first created a testdata directory in the HDFS /user/ec2-user/ directory. I then copied the part-00000 Hadoop Streaming output of 1B (the Hadoop Streaming 3 column extraction output) from the HDFS /data/lo_streaming_3col/ directory into the new /user/ec2-user/testdata directory. In Linux, I then went into the /home/ec2-user/apache-mahout-distribution-0.11.2/ directory and ran the Mahout synthetic clustering command below.

Mahout 11 Cluster Synthetic KMeans Command:

```
time bin/mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job --numClusters 11 --t1 100 -  
-t2 75 --maxIter 10 --input testdata --output output11
```

I set the t1 value to 100 and t2 to 75 because I was able to find the default values and set them close (the defaults were t1 = 80 and t2 = 55). I set the maxIter to 10 because this was the default.

t1, t2, and maxIter default values in org.apache.mahout.clustering.syntheticcontrol.kmeans.Job code:

The screenshot shows the GrepCode.com interface for the `org.apache.mahout.clustering.syntheticcontrol.kmeans.Job` class. The code is annotated with red boxes highlighting specific lines of code:

- A red box surrounds the line `double t1 = Double.parseDouble(cmdLine.getValue(t1Opt, "80").toString());`
- A red box surrounds the line `double t2 = Double.parseDouble(cmdLine.getValue(t2Opt, "55").toString());`
- A red box surrounds the line `int maxIterations = Integer.parseInt(cmdLine.getValue(maxIterationsOpt, 10).toString());`

The code snippet is as follows:

```
81     Parser parser = new Parser();
82     parser.setGroup(group);
83     CommandLine cmdline = parser.parse(args);
84
85     if (cmdLine.hasOption(helpOpt)) {
86         CommandLineUtil.printHelp(group);
87         return;
88     }
89     String input = cmdLine.getValue(inputOpt, "testdata").toString();
90     String output = cmdLine.getValue(outputOpt, "output").toString();
91     String measureClass = cmdLine.getValue(measureClassOpt,
92         "org.apache.mahout.common.distance.EuclideanDistanceMeasure").toString();
93     double t1 = Double.parseDouble(cmdLine.getValue(t1Opt, "80").toString());
94     double t2 = Double.parseDouble(cmdLine.getValue(t2Opt, "55").toString());
95
96     double convergenceDelta = Double.parseDouble(cmdLine.getValue(convergenceDeltaOpt, "0.5").toString());
97     int maxIterations = Integer.parseInt(cmdLine.getValue(maxIterationsOpt, 10).toString());
98     // String className = cmdLine.getValue(vectorClassOpt,
99     // "org.apache.mahout.math.RandomAccessSparseVector").toString();
100    // Class<? extends Vector> vectorClass = Class.forName(className).asSubclass(Vector.class);
101
102    runJob(input, output, measureClass, t1, t2, convergenceDelta, maxIterations);
103 } catch (OptionException e) {
104     log.error("Exception", e);
105     CommandLineUtil.printHelp(group);
106 }
```

Below the code, a note states: "Run the kmeans clustering job on an input dataset using the given distance measure, t1, t2 and iteration parameters. All output data will be written to the output directory, which will be initially deleted if it exists."

/user/ec2-user/testdata/part-00000 File Contents Used For Clustering:

```
ec2-user@ip-172-31-3-174:~/apache-mahout-distribution-0.11.2
1 36 6089141
1 39 4513579
1 41 7419259
1 39 5368747
1 3 418802
1 25 4029002
1 15 2839641
1 43 5631089
1 8 855076
1 43 5887253
1 32 4968588
1 1 129188
1 40 5094187
1 7 623651
1 46 5841768
1 50 5800693
1 34 4909416
1 13 1739025
1 17 2566990
1 4 581863
1 45 4281464
1 21 3183094
1 3 471406
1 45 5785009
1 38 4750897
1 46 7821078
1 38 6643491
1 50 5399347
[ec2-user@ip-172-31-3-174 apache-mahout-distribution-0.11.2]$ ^C
[ec2-user@ip-172-31-3-174 apache-mahout-distribution-0.11.2]$
```

Start of Mahout Synthetic KMeans Execution (first and last map/reduce passes shown):

```
[ec2-user@ip-172-31-3-174:~/apache-mahout-distribution-0.11.2]$ time bin/mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job --numClusters 11 --t1 100 --t2 75 --maxIter 10 --input testdata --output output11
Running on hadoop, using /home/ec2-user/hadoop-2.6.4/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /home/ec2-user/apache-mahout-distribution-0.11.2/mahout-examples-0.11.2-job.jar
18/03/13 03:47:26 WARN MahoutDriver: No org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.pr
ops found on classpath, will use command-line arguments only
18/03/13 03:47:26 INFO Job: Running with only user-supplied arguments
18/03/13 03:47:26 INFO AbstractJob: Command line arguments: {--convergenceDelta=[0.5], --distanceMe
asure=[org.apache.mahout.common.distance.SquaredEuclideanDistanceMeasure], --endPhase=[2147483647],
--input=[testdata], --maxIter=[10], --numClusters=[11], --output=[output11], --startPhase=[0], --t
1=[100], --t2=[75], --tempDir=[temp]}
18/03/13 03:47:26 INFO Job: Preparing Input
18/03/13 03:47:27 INFO RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 03:47:27 WARN JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/03/13 03:47:28 INFO FileInputFormat: Total input paths to process : 1
18/03/13 03:47:28 INFO JobSubmitter: number of splits:1
18/03/13 03:47:28 INFO JobSubmitter: Submitting tokens for job: job_1520879226572_0144
18/03/13 03:47:29 INFO YarnClientImpl: Submitted application application_1520879226572_0144
18/03/13 03:47:29 INFO Job: The url to track the job: http://ip-172-31-3-174.us-east-2.compute.inte
rnal:8088/proxy/application_1520879226572_0144/
18/03/13 03:47:29 INFO Job: Running job: job_1520879226572_0144
18/03/13 03:47:35 INFO Job: Job job_1520879226572_0144 running in uber mode : false
18/03/13 03:47:35 INFO Job: map 0% reduce 0%
18/03/13 03:47:46 INFO Job: map 34% reduce 0%
18/03/13 03:47:49 INFO Job: map 54% reduce 0%
18/03/13 03:47:52 INFO Job: map 75% reduce 0%
18/03/13 03:47:55 INFO Job: map 100% reduce 0%
18/03/13 03:47:56 INFO Job: Job job_1520879226572_0144 completed successfully
18/03/13 03:47:56 INFO Job: Counters: 30
    File System Counters
        FILE: Number of bytes read=0
        FILE: Number of bytes written=106545
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=89519927
        HDFS: Number of bytes written=264390130
        HDFS: Number of read operations=5
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=1
        Rack-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=18437
        Total time spent by all reduces in occupied slots (ms)=0
        Total time spent by all map tasks (ms)=18437
        Total vcore-milliseconds taken by all map tasks=18437
```

```
ec2-user@ip-172-31-3-174:~/apache-mahout-distribution-0.11.2
      Total vcore-milliseconds taken by all map tasks=18437
      Total megabyte-milliseconds taken by all map tasks=18879488
Map-Reduce Framework
  Map input records=6544308
  Map output records=6544308
  Input split bytes=118
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=255
  CPU time spent (ms)=17600
  Physical memory (bytes) snapshot=186580992
  Virtual memory (bytes) snapshot=984342528
  Total committed heap usage (bytes)=112721920
File Input Format Counters
  Bytes Read=89519809
File Output Format Counters
  Bytes Written=264390130
18/03/13 03:47:56 INFO Job: Running random seed to get initial clusters
18/03/13 03:48:13 INFO ZlibFactory: Successfully loaded & initialized native-zlib library
18/03/13 03:48:13 INFO CodecPool: Got brand-new compressor [.deflate]
18/03/13 03:48:13 INFO RandomSeedGenerator: Wrote 11 Klusters to output11/random-seeds/part-randomSeed
18/03/13 03:48:13 INFO Job: Running KMeans with k = 11
18/03/13 03:48:13 INFO KMeansDriver: Input: output11/data Clusters In: output11/random-seeds/part-randomSeed Out: output11
18/03/13 03:48:13 INFO KMeansDriver: convergence: 0.5 max Iterations: 10
18/03/13 03:48:13 INFO CodecPool: Got brand-new decompressor [.deflate]
18/03/13 03:48:13 INFO RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 03:48:14 INFO FileInputFormat: Total input paths to process : 1
18/03/13 03:48:14 INFO JobSubmitter: number of splits:2
18/03/13 03:48:14 INFO JobSubmitter: Submitting tokens for job: job_1520879226572_0145
18/03/13 03:48:14 INFO YarnClientImpl: Submitted application application_1520879226572_0145
18/03/13 03:48:14 INFO Job: The url to track the job: http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520879226572_0145/
18/03/13 03:48:14 INFO Job: Running job: job_1520879226572_0145
18/03/13 03:48:19 INFO Job: Job job_1520879226572_0145 running in uber mode : false
18/03/13 03:48:19 INFO Job: map 0% reduce 0%
18/03/13 03:48:31 INFO Job: map 17% reduce 0%
18/03/13 03:48:34 INFO Job: map 31% reduce 0%
18/03/13 03:48:37 INFO Job: map 45% reduce 0%
18/03/13 03:48:40 INFO Job: map 59% reduce 0%
18/03/13 03:48:42 INFO Job: map 100% reduce 0%
18/03/13 03:48:48 INFO Job: map 100% reduce 100%
18/03/13 03:48:48 INFO Job: Job job_1520879226572_0145 completed successfully
18/03/13 03:48:48 INFO Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=5308
    FILE: Number of bytes written=333508
```

```
ec2-user@ip-172-31-3-174:~/apache-mahout-distribution-0.11.2
18/03/13 03:54:03 INFO YarnClientImpl: Submitted application application_1520879226572_0155
18/03/13 03:54:03 INFO Job: The url to track the job: http://ip-172-31-3-174.us-east-2.compute.intern
rnal:8088/proxy/application_1520879226572_0155/
18/03/13 03:54:03 INFO Job: Running job: job_1520879226572_0155
18/03/13 03:54:09 INFO Job: Job job_1520879226572_0155 running in uber mode : false
18/03/13 03:54:09 INFO Job: map 0% reduce 0%
18/03/13 03:54:20 INFO Job: map 19% reduce 0%
18/03/13 03:54:23 INFO Job: map 32% reduce 0%
18/03/13 03:54:26 INFO Job: map 44% reduce 0%
18/03/13 03:54:29 INFO Job: map 57% reduce 0%
18/03/13 03:54:32 INFO Job: map 70% reduce 0%
18/03/13 03:54:35 INFO Job: map 83% reduce 0%
18/03/13 03:54:38 INFO Job: map 96% reduce 0%
18/03/13 03:54:39 INFO Job: map 100% reduce 0%
18/03/13 03:54:39 INFO Job: Job job_1520879226572_0155 completed successfully
18/03/13 03:54:39 INFO Job: Counters: 31
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=214690
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=264397180
    HDFS: Number of bytes written=588212225
    HDFS: Number of read operations=26
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
  Job Counters
    Launched map tasks=2
    Data-local map tasks=1
    Rack-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=55535
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=55535
    Total vcore-milliseconds taken by all map tasks=55535
    Total megabyte-milliseconds taken by all map tasks=56867840
  Map-Reduce Framework
    Map input records=6544308
    Map output records=6544308
    Input split bytes=250
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=551
    CPU time spent (ms)=55310
    Physical memory (bytes) snapshot=375996416
    Virtual memory (bytes) snapshot=1996333056
    Total committed heap usage (bytes)=189267968
  File Input Format Counters
```

```
ec2-user@ip-172-31-3-174:~/apache-mahout-distribution-0.11.2$ bin/mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job --numClusters 11
Running on hadoop, using /home/ec2-user/hadoop-2.6.4/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /home/ec2-user/apache-mahout-distribution-0.11.2/mahout-examples-0.11.2-job.jar

File Input Format Counters
    Bytes Read=264391458
File Output Format Counters
    Bytes Written=588212225
18/03/13 03:54:39 INFO Job: Dumping out clusters from clusters: output11/clusters--final and clusteredPoints: output11/clusteredPoints
Exception in thread "main" java.lang.OutOfMemoryError: GC overhead limit exceeded
    at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:887)
    at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:696)
    at java.io.DataInputStream.readInt(DataInputStream.java:390)
    at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2433)
    at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2333)
    at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2379)
    at org.apache.mahout.common.iterator.sequencefile.SequenceFileIterator.computeNext(SequenceFileIterator.java:101)
    at org.apache.mahout.common.iterator.sequencefile.SequenceFileIterator.computeNext(SequenceFileIterator.java:40)
    at com.google.common.collect.AbstractIterator.tryToComputeNext(AbstractIterator.java:143)
    at com.google.common.collect.AbstractIterator.hasNext(AbstractIterator.java:138)
    at com.google.common.collect.Iterators$5.hasNext(Iterators.java:543)
    at com.google.common.collect.ForwardingIterator.hasNext(ForwardingIterator.java:43)
    at org.apache.mahout.utils.clustering.ClusterDumper.readPoints(ClusterDumper.java:311)
    at org.apache.mahout.utils.clustering.ClusterDumper.init(ClusterDumper.java:262)
    at org.apache.mahout.utils.clustering.ClusterDumper.<init>(ClusterDumper.java:92)
    at org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.run(Job.java:141)
    at org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.run(Job.java:95)
    at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
    at org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.main(Job.java:54)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(ProgramDriver.java:71)
    at org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:144)
    at org.apache.hadoop.util.ProgramDriver.driver(ProgramDriver.java:152)
    at org.apache.mahout.driver.MahoutDriver.main(MahoutDriver.java:195)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:221)

real    7m48.578s
user    1m27.476s
sys     0m3.560s
[ec2-user@ip-172-31-3-174 apache-mahout-distribution-0.11.2]$
```

The clustering output had 12 map/reduce passes present (not sure why 12 when maxIter was 10). The last map/reduce pass was missing the reduce stage. It also had 0 for the number of merged map outputs, where all other passes had 2. The number of job counters of the last map/reduce pass was 30 versus the 49 to 50 for most map/reduce passes. Just after the last pass ended, an “Exception in thread “main” java.lang.OutOfMemoryError: GC overhead limit exceeded” error was received.

I checked the HDFS report to see how much of the storage was used. I found that only 29.43% of the DFS storage was used. So, this is not the source of the issue.

```
[ec2-user@ip-172-31-3-174 ~]$ hdfs dfsadmin -report
Configured Capacity: 126282006528 (117.61 GB)
Present Capacity: 109708722176 (102.17 GB)
DFS Remaining: 77422796800 (72.11 GB)
DFS Used: 32285925376 (30.07 GB)
DFS Used%: 29.43%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-----
Live datanodes (4):

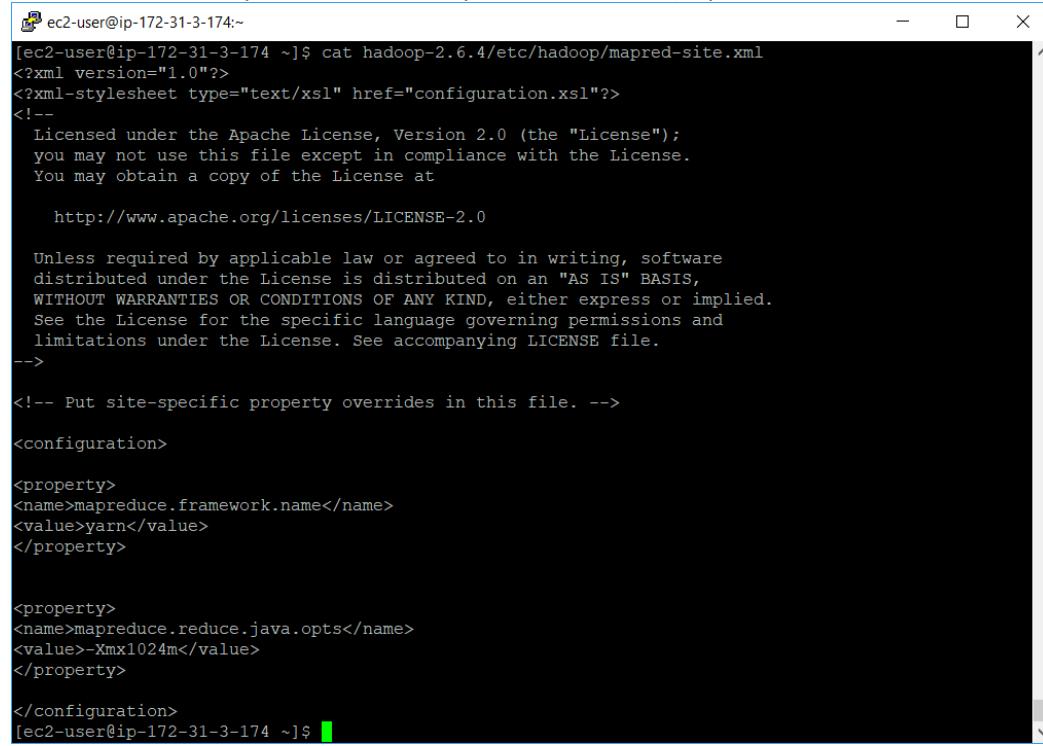
Name: 172.31.5.189:50010 (ip-172-31-5-189.us-east-2.compute.internal)
Hostname: ip-172-31-5-189.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 31570501632 (29.40 GB)
DFS Used: 9351499776 (8.71 GB)
Non DFS Used: 1997041664 (1.86 GB)
DFS Remaining: 20221960192 (18.83 GB)
DFS Used%: 29.62%
DFS Remaining%: 64.05%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue Mar 13 23:48:03 UTC 2018

Name: 172.31.3.174:50010 (ip-172-31-3-174.us-east-2.compute.internal)
Hostname: ip-172-31-3-174.us-east-2.compute.internal
Decommission Status : Normal
Configured Capacity: 31570501632 (29.40 GB)
DFS Used: 9962610688 (9.28 GB)
Non DFS Used: 9964974080 (9.28 GB)
DFS Remaining: 11642916864 (10.84 GB)
DFS Used%: 31.56%
DFS Remaining%: 36.88%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue Mar 13 23:48:02 UTC 2018
```

```
ec2-user@ip-172-31-3-174:~  
DFS Remaining: 11642916864 (10.84 GB)  
DFS Used%: 31.56%  
DFS Remaining%: 36.88%  
Configured Cache Capacity: 0 (0 B)  
Cache Used: 0 (0 B)  
Cache Remaining: 0 (0 B)  
Cache Used%: 100.00%  
Cache Remaining%: 0.00%  
Xceivers: 1  
Last contact: Tue Mar 13 23:48:02 UTC 2018  
  
Name: 172.31.8.219:50010 (ip-172-31-8-219.us-east-2.compute.internal)  
Hostname: ip-172-31-8-219.us-east-2.compute.internal  
Decommission Status : Normal  
Configured Capacity: 31570501632 (29.40 GB)  
DFS Used: 7056220160 (6.57 GB)  
Non DFS Used: 2305433600 (2.15 GB)  
DFS Remaining: 22208847872 (20.68 GB)  
DFS Used%: 22.35%  
DFS Remaining%: 70.35%  
Configured Cache Capacity: 0 (0 B)  
Cache Used: 0 (0 B)  
Cache Remaining: 0 (0 B)  
Cache Used%: 100.00%  
Cache Remaining%: 0.00%  
Xceivers: 1  
Last contact: Tue Mar 13 23:48:03 UTC 2018  
  
Name: 172.31.13.181:50010 (ip-172-31-13-181.us-east-2.compute.internal)  
Hostname: ip-172-31-13-181.us-east-2.compute.internal  
Decommission Status : Normal  
Configured Capacity: 31570501632 (29.40 GB)  
DFS Used: 5915594752 (5.51 GB)  
Non DFS Used: 2305835008 (2.15 GB)  
DFS Remaining: 23349071872 (21.75 GB)  
DFS Used%: 18.74%  
DFS Remaining%: 73.96%  
Configured Cache Capacity: 0 (0 B)  
Cache Used: 0 (0 B)  
Cache Remaining: 0 (0 B)  
Cache Used%: 100.00%  
Cache Remaining%: 0.00%  
Xceivers: 1  
Last contact: Tue Mar 13 23:48:03 UTC 2018  
  
[ec2-user@ip-172-31-3-174 ~]$
```

I took the advice of the project assignment and added the mapreduce.reduce.java.opts with a value of -Xmx1024 in the mapred-site.xml file of all nodes of the cluster. The screenshot below is from the master node. I copied this file to all of the workers using scp, then copied the updated file into each /hadoop-2.6.4/etc/hadoop/ folder (removing the original mapred-site.xml file first). I then shut down Hadoop (stopped the history server, yarn, and dfs). I then restarted Hadoop (started the history server, yarn, and dfs).

Screenshot of the updated master mapred-site.xml file (copied to and verified to be on all clusters):



```
[ec2-user@ip-172-31-3-174 ~]$ cat hadoop-2.6.4/etc/hadoop/mapred-site.xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>

<property>
<name>mapreduce.reduce.java.opts</name>
<value>-Xmx1024m</value>
</property>
</configuration>
[ec2-user@ip-172-31-3-174 ~]$
```

I then went back into the apache mahout directory and re-ran the clustering command for 11 clusters. I received the same error at the last map/reduce pass that I did before updating the mapred-site.xml file. It occurred during the 12th map/reduce pass again and had the same characteristics (no reduce pass, 0 merged map output setting, and number of job counters 30 versus the 49 to 50 for most map/reduce passes).

Error during pass with updated mapred-site.xml file:

```
ec2-user@ip-172-31-3-174:~/apache-mahout-distribution-0.11.2
Virtual memory (bytes) snapshot=1988952064
Total committed heap usage (bytes)=220200960
File Input Format Counters
Bytes Read=264391458
File Output Format Counters
Bytes Written=588113276
18/03/13 04:40:17 INFO Job: Dumping out clusters from clusters: output11/clusters--final and clusteredPoints: output11/clusteredPoints
Exception in thread "main" java.lang.OutOfMemoryError: GC overhead limit exceeded
    at java.lang.String.toCharArray(String.java:2748)
    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 org.apache.mahout.clustering.classify.WeightedPropertyVectorWritable.readFields(WeightedPropertyVectorWritable.java:61)
    at org.apache.hadoop.io.SequenceFile$Reader.getCurrentValue(SequenceFile.java:2254)
    at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2382)
    at org.apache.mahout.common.iterator.sequencefile.SequenceFileIterator.computeNext(SequenceFileIterator.java:101)
    at org.apache.mahout.common.iterator.sequencefile.SequenceFileIterator.computeNext(SequenceFileIterator.java:40)
    at com.google.common.collect.AbstractIterator.tryToComputeNext(AbstractIterator.java:143)
    at com.google.common.collect.AbstractIterator.hasNext(AbstractIterator.java:138)
    at com.google.common.collect.Iterators$5.hasNext(Iterators.java:543)
    at com.google.common.collect.ForwardingIterator.hasNext(ForwardingIterator.java:43)
    at org.apache.mahout.utils.clustering.ClusterDumper.readPoints(ClusterDumper.java:311)
    at org.apache.mahout.utils.clustering.ClusterDumper.init(ClusterDumper.java:262)
    at org.apache.mahout.utils.clustering.ClusterDumper.<init>(ClusterDumper.java:92)
    at org.apache.mahout.clustering.syntheticcontrol.Kmeans.Job.run(Job.java:141)
    at org.apache.mahout.clustering.syntheticcontrol.Kmeans.Job.run(Job.java:95)
    at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
    at org.apache.mahout.clustering.syntheticcontrol.Kmeans.Job.main(Job.java:54)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(ProgramDriver.java:71)
    at org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:144)
    at org.apache.hadoop.util.ProgramDriver.driver(ProgramDriver.java:152)
    at org.apache.mahout.driver.MahoutDriver.main(MahoutDriver.java:195)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)

real    7m58.034s
user    1m24.096s
sys     0m3.440s
[ec2-user@ip-172-31-3-174 apache-mahout-distribution-0.11.2]$
```

Next, I stopped Hadoop, exited the cluster, and stopped all of the instances in AWS. I then increased all of the nodes in AWS to be t1-large instances with 8G of memory. After this, I started the instances again. I then updated the mapred-site.xml file to set the max memory available for mapping and reducing to 4G each and the map and reduce java.opts to 3G each (found this to be a way online to work around the out of memory error). I again made the same updates to all of the mapred-site.xml files on all of the nodes. The screenshot below is of the file on the master node.

Master node new mapred-site.xml with max 4G RAM (t1-large instances):

```
ec2-user@ip-172-31-3-174:~$ nano 2.5.3      File: hadoop-2.6.4/etc/hadoop/mapred-site.xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
 Licensed under the Apache License, Version 2.0 (the "License");
 you may not use this file except in compliance with the License.
 You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>

<property>
<name>mapreduce.map.memory.mb</name>
<value>4096</value>
</property>

<property>
<name>mapreduce.reduce.memory.mb</name>
<value>4096</value>
</property>

<property>
<name>mapreduce.map.java.opts</name>
<value>-Xmx3072m</value>
</property>

^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos
^X Exit       ^R Read File  ^\ Replace   ^U Uncut Text ^T To Spell  ^  Go To Line
```

The screenshot shows a terminal window with a black background and white text. At the top, it says "ec2-user@ip-172-31-3-174:~". Below that, it says "GNU nano 2.5.3 File: hadoop-2.6.4/etc/mapred-site.xml". The main area contains XML code for configuration properties. At the bottom, there is a menu bar with various keyboard shortcuts.

```
</property>

<property>
<name>mapreduce.map.memory.mb</name>
<value>4096</value>
</property>

<property>
<name>mapreduce.reduce.memory.mb</name>
<value>4096</value>
</property>

<property>
<name>mapreduce.map.java.opts</name>
<value>-Xmx3072m</value>
</property>

<property>
<name>mapreduce.reduce.java.opts</name>
<value>-Xmx3072m</value>
</property>

</configuration>
```

^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify ^C Cur Pos
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Spell ^ Up Go To Line

I then restarted Hadoop (started the history server, yarn, and dfs). I then went back into the apache mahout directory and re-ran the clustering command for 11 clusters. I again received the same error at the last map/reduce pass that I did on the previous 2 tries). I shutdown Hadoop, stopped the instances, and changed them back to t1 medium. I also reverted all of the mapred-site.xml files on all nodes to the original (without any mapred memory or java.opts entries). I then started Hadoop again for use in the rest of the project (the cluster was returned to the same state it was in before part 3A).

Error during pass with new mapred-site.xml with max 4G RAM (t1-large instances):

```
ec2-user@ip-172-31-3-174:~/apache-mahout-distribution-0.11.2
Rack-local map tasks=1
Total time spent by all maps in occupied slots (ms)=220496
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=55124
Total vcore-milliseconds taken by all map tasks=55124
Total megabyte-milliseconds taken by all map tasks=225787904
Map-Reduce Framework
  Map input records=6544308
  Map output records=6544308
  Input split bytes=250
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=361
  CPU time spent (ms)=54690
  Physical memory (bytes) snapshot=770191360
  Virtual memory (bytes) snapshot=8234065920
  Total committed heap usage (bytes)=281018368
File Input Format Counters
  Bytes Read=264391458
File Output Format Counters
  Bytes Written=588192883
18/03/13 23:10:35 INFO Job: Dumping out clusters from clusters: output11/clusters-*_final and clusteredPoints: output11/clusteredPoints
Exception in thread "main" java.lang.OutOfMemoryError: GC overhead limit exceeded
        at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:887)
        at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:696)
        at java.io.DataInputStream.readInt(DataInputStream.java:390)
        at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2433)
        at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2333)
        at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2379)
        at org.apache.mahout.common.iterator.sequencefile.SequenceFileIterator.computeNext(SequenceFileIterator.java:101)
        at org.apache.mahout.common.iterator.sequencefile.SequenceFileIterator.computeNext(SequenceFileIterator.java:40)
        at com.google.common.collect.AbstractIterator.tryToComputeNext(AbstractIterator.java:143)
        at com.google.common.collect.AbstractIterator.hasNext(AbstractIterator.java:138)
        at com.google.common.collect.Iterators$5.hasNext(Iterators.java:543)
        at com.google.common.collect.ForwardingIterator.hasNext(ForwardingIterator.java:43)
        at org.apache.mahout.utils.clustering.ClusterDumper.readPoints(ClusterDumper.java:311)
        at org.apache.mahout.utils.clustering.ClusterDumper.init(ClusterDumper.java:262)
        at org.apache.mahout.utils.clustering.ClusterDumper.<init>(ClusterDumper.java:92)
```

B)

3 Passes of Manual Hadoop Streaming Clustering With Number of Clusters = 11

I created a centers.txt file in the /home/ec2-user/ Linux directory. I populated it with 11 points I chose from the 3 column data in the HDFS /user/ec2-user/testdata/part-00000 file. This centers file was read in by the mapper code. The mapper code determined which center each line of the current block was closest to and printed the cluster number as the key and the current point (the three values of the current line) as the value, separated by underscores. The reducer then received all of the points assigned with the same cluster number in order. The reducer looped over the points and added each of the column entries together (generated a sum of the first column, second column, and third column), as well as updated a counter to track the number of points in the current cluster number. When a new cluster number was reached, the sum of each of the columns was divided by the number of points to create a mean point (mean of the first column, mean of the second column, and mean of the third column are the mean point). The keys of each reducer output were set to the cluster number corresponding to each center and the value was the three entries of the new cluster mean, separated by spaces. The output of all of the reducers was the new cluster centers to be used in the next round of clustering. After the new cluster centers were created, I copied the file from HDFS to the Linux ec2-user home directory, renamed the old centers.txt file to reflect the pass it was used for, then renamed the clustering output to centers.txt. I then updated the centers.txt file to delete the cluster number keys, so only the center points remained. I repeated this entire process until a total of 3 passes were reached.

Cluster Mapper Code:

The screenshot shows a terminal window titled "GNU nano 2.5.3" with the file "cluster_mapper.py" open. The code implements a k-means clustering algorithm. It starts by reading center points from "centers.txt", then iterates over input points, calculating distances to each center and determining the nearest one.

```
ec2-user@ip-172-31-3-174:~$ GNU nano 2.5.3          File: cluster_mapper.py
#!/usr/bin/python

import sys
import math

center_file = open('centers.txt', 'r')
center_lst = center_file.readlines()
center_file.close()

def calc_dist(v1, v2):
    if len(v1) != len(v2):
        return

    sum = 0
    for x in range(len(v1)):
        sum += (float(v2[x]) - float(v1[x]))**2

    return math.sqrt(sum)

for line in sys.stdin:
    curr_pt = line.strip().split(' ')
    min_dist_ndx = 0
    min_dist = 999999999999
    for y in range(len(center_lst)):
        entry = center_lst[y]
        center = entry.strip().split(' ')
        tmp_dist = calc_dist(center, curr_pt)
        if tmp_dist < min_dist:
            min_dist = tmp_dist
            min_dist_ndx = y
    print '%s\t%s' % (min_dist_ndx, ' '.join(curr_pt))
```

At the bottom of the terminal window, there is a menu bar with various keyboard shortcuts:

- ^G Get Help
- ^O Write Out
- ^W Where Is
- ^K Cut Text
- ^J Justify
- ^C Cur Pos
- ^X Exit
- ^R Read File
- ^V Replace
- ^U Uncut Text
- ^T To Linter
- ^L Go To Line

Cluster Reducer Code:

The screenshot shows a terminal window titled "GNU nano 2.5.3" with the file "cluster_reducer.py" open. The code implements a reducer for a map-reduce job, specifically for calculating means. It reads input from standard input, processes it by key, and prints the mean for each key.

```
ec2-user@ip-172-31-3-174:~$ GNU nano 2.5.3          File: cluster_reducer.py
#!/usr/bin/python

import sys

curr_key = None
line_key = None
sum_pts = []
num_pts = 0

for line in sys.stdin:
    vals = line.strip().split('\t')
    line_key = vals[0]
    curr_pt = vals[1].split(' ')

    if curr_key == line_key:
        num_pts += 1
        for x in range(len(sum_pts)):
            sum_pts[x] += float(curr_pt[x])

    else:
        if curr_key:
            mean_pts = []
            for y in range(len(sum_pts)):
                curr_mean = sum_pts[y] / num_pts
                mean_pts.append(str(curr_mean))

            mean_str = ' '.join(mean_pts)
            print '%s\t%s' % (curr_key, mean_str)

        curr_key = line_key
        num_pts = 1
        sum_pts = []
        for x in range(len(curr_pt)):
            sum_pts.append(float(curr_pt[x]))

if curr_key == line_key:
    mean_pts = []
    for y in range(len(sum_pts)):
        curr_mean = sum_pts[y] / num_pts
        mean_pts.append(str(curr_mean))

    mean_str = ' '.join(mean_pts)
    print '%s\t%s' % (curr_key, mean_str)
```

The terminal also displays a set of keyboard shortcuts at the bottom:

- ^G Get Help
- ^O Write Out
- ^W Where Is
- ^K Cut Text
- ^J Justify
- ^C Cur Pos
- ^X Exit
- ^R Read File
- ^N Replace
- ^U Uncut Text
- ^T To Linter
- ^ Go To Line

Initial Cluster centers.txt File:

The screenshot shows a terminal window titled "GNU nano 2.5.3" with the file "centers.txt" open. The window displays a list of 11 numerical coordinates, each consisting of three values separated by spaces. The coordinates are:

```
4 26 3443188
5 30 5105503
6 4 419627
1 23 4158023
2 8 1282644
3 32 4806492
5 50 7103386
6 25 2641893
1 36 6206656
2 42 4299490
4 50 5904187
```

At the bottom of the terminal window, there is a menu bar with various keyboard shortcuts for text manipulation.

Initial Hadoop Streaming Command Run:

```
time hadoop jar hadoop-streaming-2.6.4.jar -D  
mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D  
mapred.text.key.comparator.options=-n -input /user/ec2-user/testdata/ -output  
/data/streaming_cluster1 -mapper cluster_mapper.py -reducer cluster_reducer.py -file /home/ec2-  
user/cluster_mapper.py -file /home/ec2-user/cluster_reducer.py -file /home/ec2-user/centers.txt
```

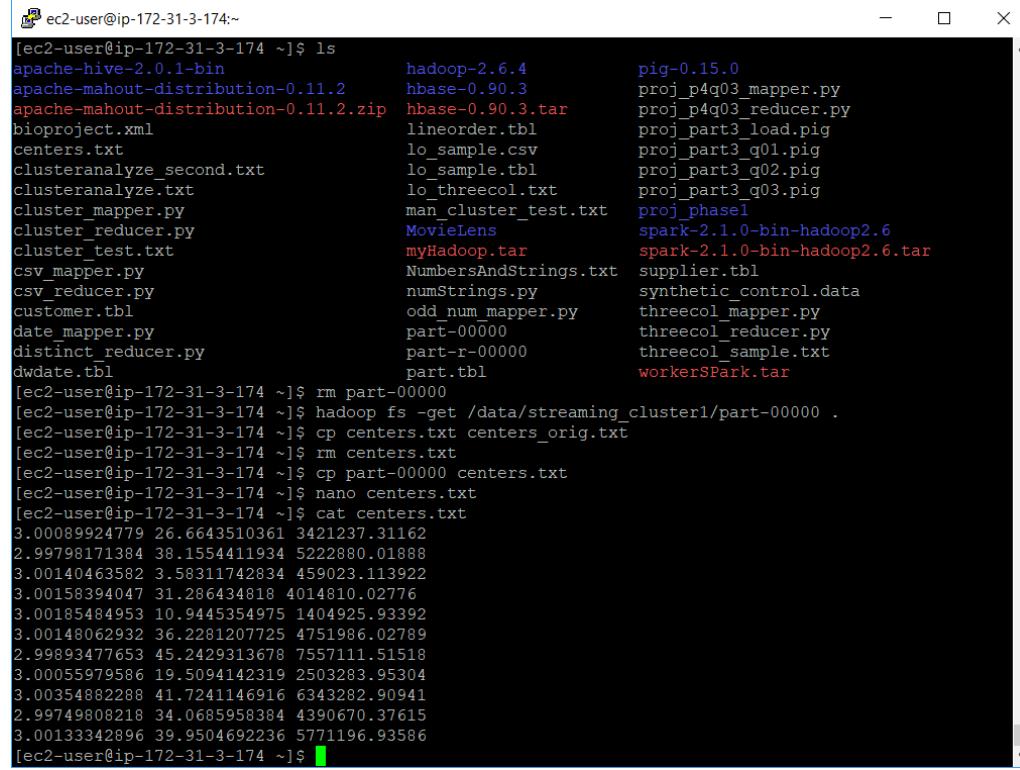
Initial Clustering Round Execution (start and end shown due to length):

```
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]$ time hadoop jar hadoop-streaming-2.6.4.jar -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D mapred.text.key.comparator.options=-n -input /user/ec2-user/testdata/ -output /data/streaming_cluster1 -mapper cluster_mapper.py -reducer cluster_reducer.py -file /home/ec2-user/cluster_mapper.py -file /home/ec2-user/cluster_reducer.py -file /home/ec2-user/centers.txt
18/03/13 01:12:16 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/ec2-user/cluster_mapper.py, /home/ec2-user/cluster_reducer.py, /home/ec2-user/centers.txt, /tmp/hadoop-unjar2759196249170212317/] [] /tmp/streamjob881182556622296568.jar tmpDir=null
18/03/13 01:12:17 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 01:12:17 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 01:12:17 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/13 01:12:17 INFO mapreduce.JobSubmitter: number of splits:2
18/03/13 01:12:17 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Instead, use mapreduce.job.output.key.comparator.class
18/03/13 01:12:17 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Instead, use mapreduce.partition.keycomparator.options
18/03/13 01:12:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1520879226572_0129
18/03/13 01:12:18 INFO impl.YarnClientImpl: Submitted application application_1520879226572_0129
18/03/13 01:12:18 INFO mapreduce.Job: The url to track the job: http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520879226572_0129/
18/03/13 01:12:18 INFO mapreduce.Job: Running job: job_1520879226572_0129
18/03/13 01:12:24 INFO mapreduce.Job: Job job_1520879226572_0129 running in uber mode : false
18/03/13 01:12:24 INFO mapreduce.Job: map 0% reduce 0%
18/03/13 01:12:35 INFO mapreduce.Job: map 1% reduce 0%
18/03/13 01:12:36 INFO mapreduce.Job: map 3% reduce 0%
18/03/13 01:12:38 INFO mapreduce.Job: map 4% reduce 0%
18/03/13 01:12:39 INFO mapreduce.Job: map 5% reduce 0%
18/03/13 01:12:42 INFO mapreduce.Job: map 6% reduce 0%
18/03/13 01:12:44 INFO mapreduce.Job: map 7% reduce 0%
18/03/13 01:12:45 INFO mapreduce.Job: map 8% reduce 0%
18/03/13 01:12:47 INFO mapreduce.Job: map 9% reduce 0%
18/03/13 01:12:50 INFO mapreduce.Job: map 10% reduce 0%
18/03/13 01:12:51 INFO mapreduce.Job: map 11% reduce 0%
18/03/13 01:12:53 INFO mapreduce.Job: map 12% reduce 0%
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
Map output records=6544308
Map output bytes=96440590
Map output materialized bytes=109529218
Input split bytes=210
Combine input records=0
Combine output records=0
Reduce input groups=11
Reduce shuffle bytes=109529218
Reduce input records=6544308
Reduce output records=11
Spilled Records=19632924
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=734
CPU time spent (ms)=281150
Physical memory (bytes) snapshot=685965312
Virtual memory (bytes) snapshot=2978676736
Total committed heap usage (bytes)=438304768
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=89520993
File Output Format Counters
Bytes Written=484
18/03/13 01:15:03 INFO streaming.StreamJob: Output directory: /data/streaming_cluster1
real    2m48.006s
user    0m4.280s
sys     0m0.232s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$
```

Initial Clustering Round File Output Size and Contents:

```
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
Spilled Records=19632924
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=734
CPU time spent (ms)=281150
Physical memory (bytes) snapshot=685965312
Virtual memory (bytes) snapshot=2978676736
Total committed heap usage (bytes)=438304768
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=89520993
File Output Format Counters
Bytes Written=484
18/03/13 01:15:03 INFO streaming.StreamJob: Output directory: /data/streaming_cluster1
real    2m48.006s
user    0m4.280s
sys     0m0.232s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/
Found 4 items
drwxr-xr-x  - ec2-user supergroup      0 2018-03-11 00:51 /data/lo_streaming_3col
drwxr-xr-x  - ec2-user supergroup      0 2018-03-11 00:40 /data/lo_streaming_csv
drwxr-xr-x  - ec2-user supergroup      0 2018-03-04 23:57 /data/p4q03_output
drwxr-xr-x  - ec2-user supergroup      0 2018-03-13 01:15 /data/streaming_cluster1
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/streaming_cluster1
Found 2 items
-rw-r--r--  2 ec2-user supergroup      0 2018-03-13 01:15 /data/streaming_cluster1/_SUCCESS
-rw-r--r--  2 ec2-user supergroup    484 2018-03-13 01:15 /data/streaming_cluster1/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ 
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=89520993
File Output Format Counters
Bytes Written=484
18/03/13 01:15:03 INFO streaming.StreamJob: Output directory: /data/streaming_cluster1
real    2m48.006s
user    0m4.280s
sys     0m0.232s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/
Found 4 items
drwxr-xr-x  - ec2-user supergroup      0 2018-03-11 00:51 /data/lo_streaming_3col
drwxr-xr-x  - ec2-user supergroup      0 2018-03-11 00:40 /data/lo_streaming_csv
drwxr-xr-x  - ec2-user supergroup      0 2018-03-04 23:57 /data/p4q03_output
drwxr-xr-x  - ec2-user supergroup      0 2018-03-13 01:15 /data/streaming_cluster1
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/streaming_cluster1
Found 2 items
-rw-r--r--  2 ec2-user supergroup      0 2018-03-13 01:15 /data/streaming_cluster1/_SUCCESS
-rw-r--r--  2 ec2-user supergroup    484 2018-03-13 01:15 /data/streaming_cluster1/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -cat /data/streaming_cluster1/part-00000
0      3.00089924779 26.6643510361 3421237.31162
1      2.99798171384 38.1554411934 5222880.01888
2      3.00140463582 3.58311742834 459023.113922
3      3.00158394047 31.286434818 4014810.02776
4      3.00185484953 10.9445354975 1404925.93392
5      3.00148062932 36.2281207725 4751986.02789
6      2.99893477653 45.2429313678 7557111.51518
7      3.00055979586 19.5094142319 2503283.95304
8      3.00354882288 41.7241146916 6343282.90941
9      2.99749808218 34.0685958384 4390670.37615
10     3.00133342896 39.9504692236 5771196.93586
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ 
```

Creation of new centers.txt file from the initial round clustering output (including nano to remove center assignments from the file):



```
[ec2-user@ip-172-31-3-174 ~]$ ls
apache-hive-2.0.1-bin          hadoop-2.6.4           pig-0.15.0
apache-mahout-distribution-0.11.2 hbase-0.90.3       proj_p4q03_mapper.py
apache-mahout-distribution-0.11.2.zip hbbase-0.90.3.tar proj_p4q03_reducer.py
bioproject.xml                 lineorder.tbl      proj_part3_load.pig
centers.txt                    lo_sample.csv     proj_part3_q01.pig
clusteranalyze_second.txt       lo_sample.tbl     proj_part3_q02.pig
clusteranalyze.txt              lo_threecol.txt   proj_part3_q03.pig
cluster_mapper.py               man_cluster_test.txt proj_phase1
cluster_reducer.py              MovieLens          spark-2.1.0-bin-hadoop2.6
cluster_test.txt                myHadoop.tar     spark-2.1.0-bin-hadoop2.6.tar
csv_mapper.py                  NumbersAndStrings.txt supplier.tbl
csv_reducer.py                 numStrings.py    synthetic_control.data
customer.tbl                   odd_num_mapper.py threecol_mapper.py
date_mapper.py                 part-00000       threecol_reducer.py
distinct_reducer.py            part-r-00000     threecol_sample.txt
dwdate.tbl                     part.tbl         workerSPark.tar
[ec2-user@ip-172-31-3-174 ~]$ rm part-00000
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -get /data/streaming_cluster1/part-00000 .
[ec2-user@ip-172-31-3-174 ~]$ cp centers.txt centers_orig.txt
[ec2-user@ip-172-31-3-174 ~]$ rm centers.txt
[ec2-user@ip-172-31-3-174 ~]$ cp part-00000 centers.txt
[ec2-user@ip-172-31-3-174 ~]$ nano centers.txt
[ec2-user@ip-172-31-3-174 ~]$ cat centers.txt
3.00089924779 26.6643510361 3421237.31162
2.99798171384 38.1554411934 5222880.01888
3.00140463582 3.58311742834 459023.113922
3.00158394047 31.286434818 4014810.02776
3.00185484953 10.9445354975 1404925.93392
3.00148062932 36.2281207725 4751986.02789
2.99893477653 45.2429313678 7557111.51518
3.00055979586 19.5094142319 2503283.95304
3.00354882288 41.7241146916 6343282.90941
2.99749808218 34.0685958384 4390670.37615
3.00133342896 39.9504692236 5771196.93586
[ec2-user@ip-172-31-3-174 ~]$
```

Second Hadoop Streaming Command Run:

```
time hadoop jar hadoop-streaming-2.6.4.jar -D
mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D
mapred.text.key.comparator.options=-n -input /user/ec2-user/testdata/ -output
/data/streaming_cluster2 -mapper cluster_mapper.py -reducer cluster_reducer.py -file /home/ec2-
user/cluster_mapper.py -file /home/ec2-user/cluster_reducer.py -file /home/ec2-user/centers.txt
```

The only change to this command is the output folder being streaming_cluster2 instead of streaming_cluster1.

Second Clustering Round Execution (start and end shown due to length):

```
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]$ time hadoop jar hadoop-streaming-2.6.4.jar -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D mapred.text.key.comparator.options=-n -input /user/ec2-user/testdata/ -output /data/streaming_cluster2 -mapper cluster_mapper.py -reducer cluster_reducer.py -file /home/ec2-user/cluster_mapper.py -file /home/ec2-user/cluster_reducer.py -file /home/ec2-user/centers.txt
18/03/13 01:25:42 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/ec2-user/cluster_mapper.py, /home/ec2-user/cluster_reducer.py, /home/ec2-user/centers.txt, /tmp/hadoop-unjar6651573805778507053/] [] /tmp/streamjob2787387856201493050.jar tmpDir=null
18/03/13 01:25:43 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 01:25:43 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 01:25:43 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/13 01:25:43 INFO mapreduce.JobSubmitter: number of splits:2
18/03/13 01:25:43 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Instead, use mapreduce.job.output.key.comparator.class
18/03/13 01:25:43 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Instead, use mapreduce.partition.keycomparator.options
18/03/13 01:25:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1520879226572_0130
18/03/13 01:25:44 INFO impl.YarnClientImpl: Submitted application application_1520879226572_0130
18/03/13 01:25:44 INFO mapreduce.Job: The url to track the job: http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520879226572_0130/
18/03/13 01:25:44 INFO mapreduce.Job: Running job: job_1520879226572_0130
18/03/13 01:25:49 INFO mapreduce.Job: Job job_1520879226572_0130 running in uber mode : false
18/03/13 01:25:49 INFO mapreduce.Job: map 0% reduce 0%
18/03/13 01:25:59 INFO mapreduce.Job: map 2% reduce 0%
18/03/13 01:26:00 INFO mapreduce.Job: map 3% reduce 0%
18/03/13 01:26:03 INFO mapreduce.Job: map 5% reduce 0%
18/03/13 01:26:05 INFO mapreduce.Job: map 6% reduce 0%
18/03/13 01:26:08 INFO mapreduce.Job: map 7% reduce 0%
18/03/13 01:26:09 INFO mapreduce.Job: map 8% reduce 0%
18/03/13 01:26:11 INFO mapreduce.Job: map 9% reduce 0%
18/03/13 01:26:12 INFO mapreduce.Job: map 10% reduce 0%
18/03/13 01:26:15 INFO mapreduce.Job: map 11% reduce 0%
18/03/13 01:26:17 INFO mapreduce.Job: map 12% reduce 0%
18/03/13 01:26:18 INFO mapreduce.Job: map 13% reduce 0%
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
Map output records=6544308
Map output bytes=96447605
Map output materialized bytes=109536233
Input split bytes=210
Combine input records=0
Combine output records=0
Reduce input groups=11
Reduce shuffle bytes=109536233
Reduce input records=6544308
Reduce output records=11
Spilled Records=19632924
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=439
CPU time spent (ms)=290090
Physical memory (bytes) snapshot=699539456
Virtual memory (bytes) snapshot=2987638784
Total committed heap usage (bytes)=436207616
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=89520993
File Output Format Counters
Bytes Written=478
18/03/13 01:28:24 INFO streaming.StreamJob: Output directory: /data/streaming_cluster2
real    2m43.075s
user    0m4.292s
sys     0m0.256s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$
```

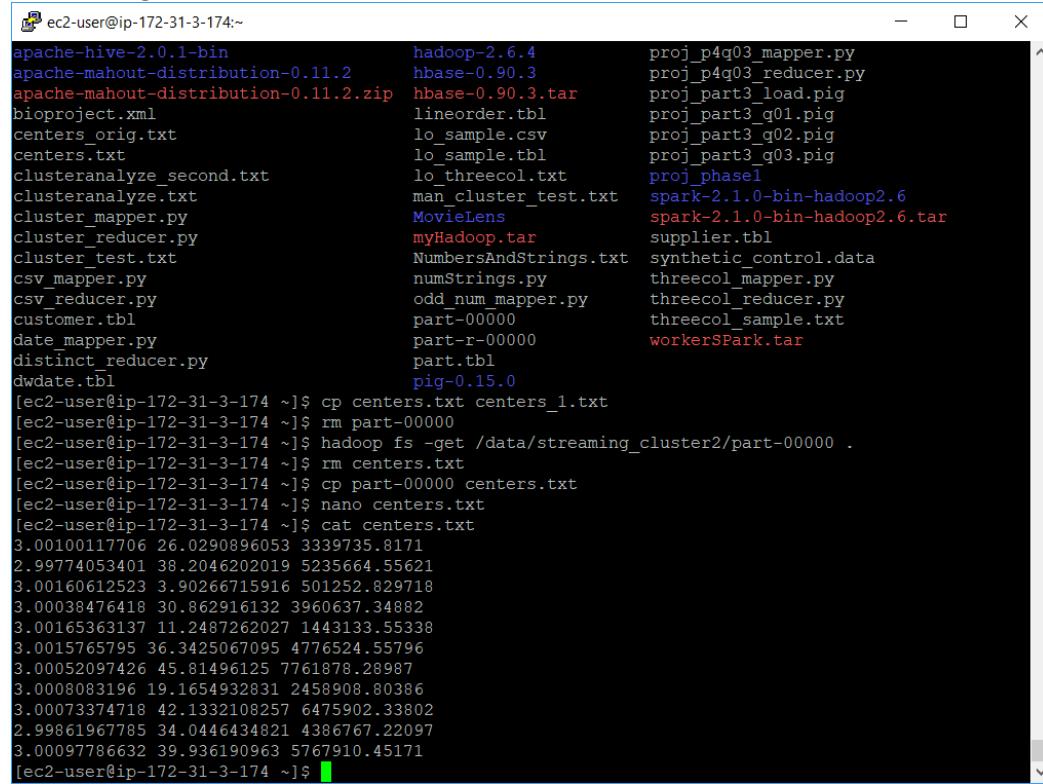
Second Clustering Round File Output Size and Contents:

```
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
Combine input records=0
Combine output records=0
Reduce input groups=11
Reduce shuffle bytes=109536233
Reduce input records=6544308
Reduce output records=11
Spilled Records=19632924
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=439
CPU time spent (ms)=290090
Physical memory (bytes) snapshot=699539456
Virtual memory (bytes) snapshots=2987638784
Total committed heap usage (bytes)=436207616
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=89520993
File Output Format Counters
Bytes Written=478
18/03/13 01:28:24 INFO streaming.StreamJob: Output directory: /data/streaming_cluster2

real    2m43.075s
user    0m4.292s
sys     0m0.256s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/streaming_cluster2
Found 2 items
-rw-r--r--  2 ec2-user supergroup          0 2018-03-13 01:28 /data/streaming_cluster2/_SUCCESS
-rw-r--r--  2 ec2-user supergroup        478 2018-03-13 01:28 /data/streaming_cluster2/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ [ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
Physical memory (bytes) snapshot=699539456
Virtual memory (bytes) snapshot=2987638784
Total committed heap usage (bytes)=436207616
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=89520993
File Output Format Counters
Bytes Written=478
18/03/13 01:28:24 INFO streaming.StreamJob: Output directory: /data/streaming_cluster2

real    2m43.075s
user    0m4.292s
sys     0m0.256s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/streaming_cluster2
Found 2 items
-rw-r--r--  2 ec2-user supergroup          0 2018-03-13 01:28 /data/streaming_cluster2/_SUCCESS
-rw-r--r--  2 ec2-user supergroup        478 2018-03-13 01:28 /data/streaming_cluster2/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -cat /data/streaming_cluster2/part-00000
0      3.00100117706 26.0290896053 3339735.8171
1      2.99774053401 38.2046202019 5235664.55621
2      3.00160612523 3.90266715916 501252.829718
3      3.00038476418 30.862916132 3960637.34882
4      3.00165363137 11.2487262027 1443133.55338
5      3.0015765795 36.3425067095 4776524.55796
6      3.00052097426 45.81496125 7761878.28987
7      3.0008083196 19.1654932831 2458908.80386
8      3.00073374718 42.1332108257 6475902.33802
9      2.99861967785 34.0446434821 4386767.22097
10     3.00097786632 39.936190963 5767910.45171
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$
```

Creation of new centers.txt file from the second round clustering output (including nano to remove center assignments from the file):



```
ec2-user@ip-172-31-3-174:~$ cp centers.txt centers_1.txt
[ec2-user@ip-172-31-3-174 ~]$ rm part-00000
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -get /data/streaming_cluster2/part-00000 .
[ec2-user@ip-172-31-3-174 ~]$ rm centers.txt
[ec2-user@ip-172-31-3-174 ~]$ cp part-00000 centers.txt
[ec2-user@ip-172-31-3-174 ~]$ nano centers.txt
[ec2-user@ip-172-31-3-174 ~]$ cat centers.txt
3.00100117706 26.0290896053 3339735.8171
2.99774053401 38.2046202019 5235664.55621
3.00160612523 3.90266715916 501252.829718
3.00038476418 30.862916132 3960637.34982
3.00165363137 11.2487262027 1443133.55338
3.0015765795 36.3425067095 4776524.55796
3.00052097426 45.81496125 7761878.28987
3.0008083196 19.1654932831 2458908.80386
3.00073374718 42.1332108257 6475902.33802
2.99861967785 34.0446434821 4386767.22097
3.00097786632 39.936190963 5767910.45171
[ec2-user@ip-172-31-3-174 ~]$
```

Third Hadoop Streaming Command Run:

```
time hadoop jar hadoop-streaming-2.6.4.jar -D
mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D
mapred.text.key.comparator.options=-n -input /user/ec2-user/testdata/ -output
/data/streaming_cluster3 -mapper cluster_mapper.py -reducer cluster_reducer.py -file /home/ec2-
user/cluster_mapper.py -file /home/ec2-user/cluster_reducer.py -file /home/ec2-user/centers.txt
```

The only change to this command is the output folder being streaming_cluster3 instead of streaming_cluster2.

Third Clustering Round Execution (start and end shown due to length):

```
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]$ time hadoop jar hadoop-streaming-2.6.4.jar -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D mapred.text.key.comparator.options=-n -input /user/ec2-user/testdata/ -output /data/streaming_cluster3 -mapper cluster_mapper.py -reducer cluster_reducer.py -file /home/ec2-user/cluster_mapper.py -file /home/ec2-user/cluster_reducer.py -file /home/ec2-user/centers.txt
18/03/13 01:37:47 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/ec2-user/cluster_mapper.py, /home/ec2-user/cluster_reducer.py, /home/ec2-user/centers.txt, /tmp/hadoop-unjar4151107630645041209/] [] /tmp/streamjob4060819480079093062.jar tmpDir=null
18/03/13 01:37:48 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 01:37:48 INFO client.RMProxy: Connecting to ResourceManager at /172.31.3.174:8032
18/03/13 01:37:49 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/13 01:37:49 INFO mapreduce.JobSubmitter: number of splits:2
18/03/13 01:37:49 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Instead, use mapreduce.job.output.key.comparator.class
18/03/13 01:37:49 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Instead, use mapreduce.partition.keycomparator.options
18/03/13 01:37:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1520879226572_0131
18/03/13 01:37:49 INFO impl.YarnClientImpl: Submitted application application_1520879226572_0131
18/03/13 01:37:49 INFO mapreduce.Job: The url to track the job: http://ip-172-31-3-174.us-east-2.compute.internal:8088/proxy/application_1520879226572_0131/
18/03/13 01:37:49 INFO mapreduce.Job: Running job: job_1520879226572_0131
18/03/13 01:37:54 INFO mapreduce.Job: Job job_1520879226572_0131 running in uber mode : false
18/03/13 01:37:54 INFO mapreduce.Job: map 0% reduce 0%
18/03/13 01:38:06 INFO mapreduce.Job: map 3% reduce 0%
18/03/13 01:38:09 INFO mapreduce.Job: map 4% reduce 0%
18/03/13 01:38:12 INFO mapreduce.Job: map 6% reduce 0%
18/03/13 01:38:15 INFO mapreduce.Job: map 7% reduce 0%
18/03/13 01:38:18 INFO mapreduce.Job: map 9% reduce 0%
18/03/13 01:38:21 INFO mapreduce.Job: map 11% reduce 0%
18/03/13 01:38:24 INFO mapreduce.Job: map 12% reduce 0%
18/03/13 01:38:27 INFO mapreduce.Job: map 14% reduce 0%
18/03/13 01:38:30 INFO mapreduce.Job: map 15% reduce 0%
18/03/13 01:38:33 INFO mapreduce.Job: map 17% reduce 0%
18/03/13 01:38:36 INFO mapreduce.Job: map 18% reduce 0%
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]$ 
Map output records=6544308
Map output bytes=96483475
Map output materialized bytes=109572103
Input split bytes=210
Combine input records=0
Combine output records=0
Reduce input groups=11
Reduce shuffle bytes=109572103
Reduce input records=6544308
Reduce output records=11
Spilled Records=19632924
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=606
CPU time spent (ms)=291970
Physical memory (bytes) snapshot=694722560
Virtual memory (bytes) snapshot=2980106240
Total committed heap usage (bytes)=449314816
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=89520993
File Output Format Counters
Bytes Written=480
18/03/13 01:40:40 INFO streaming.StreamJob: Output directory: /data/streaming_cluster3
real    2m54.099s
user    0m4.344s
sys     0m0.224s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ 
```

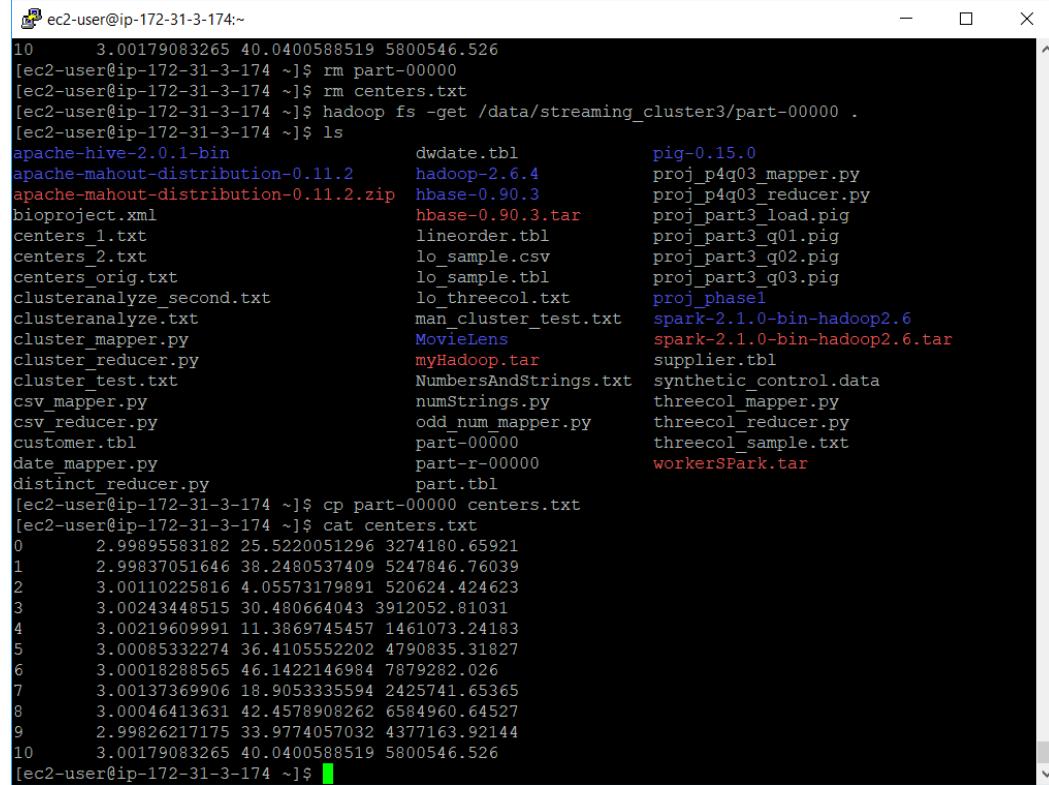
Third Clustering Round File Output Size and Contents:

```
[ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
Combine input records=0
Combine output records=0
Reduce input groups=11
Reduce shuffle bytes=109572103
Reduce input records=6544308
Reduce output records=11
Spilled Records=19632924
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=606
CPU time spent (ms)=291970
Physical memory (bytes) snapshot=694722560
Virtual memory (bytes) snapshots=2980106240
Total committed heap usage (bytes)=449314816
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=89520993
File Output Format Counters
Bytes Written=480
18/03/13 01:40:40 INFO streaming.StreamJob: Output directory: /data/streaming_cluster3

real    2m54.099s
user    0m4.344s
sys     0m0.224s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/streaming_cluster3
Found 2 items
-rw-r--r--  2 ec2-user supergroup          0 2018-03-13 01:40 /data/streaming_cluster3/_SUCCESS
-rw-r--r--  2 ec2-user supergroup        480 2018-03-13 01:40 /data/streaming_cluster3/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ [ec2-user@ip-172-31-3-174:~/hadoop-2.6.4]
Physical memory (bytes) snapshot=694722560
Virtual memory (bytes) snapshot=2980106240
Total committed heap usage (bytes)=449314816
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=89520993
File Output Format Counters
Bytes Written=480
18/03/13 01:40:40 INFO streaming.StreamJob: Output directory: /data/streaming_cluster3

real    2m54.099s
user    0m4.344s
sys     0m0.224s
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -ls /data/streaming_cluster3
Found 2 items
-rw-r--r--  2 ec2-user supergroup          0 2018-03-13 01:40 /data/streaming_cluster3/_SUCCESS
-rw-r--r--  2 ec2-user supergroup        480 2018-03-13 01:40 /data/streaming_cluster3/part-00000
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$ hadoop fs -cat /data/streaming_cluster3/part-00000
0      2.99895583182 25.5220051296 3274180.65921
1      2.99837051646 38.2480537409 5247846.76039
2      3.00110225816 4.05573179891 520624.424623
3      3.00243448515 30.480664043 3912052.81031
4      3.00219609991 11.3869745457 1461073.24183
5      3.00085332274 36.4105552202 4790835.31827
6      3.00018288565 46.1422146984 7879282.026
7      3.00137369906 18.9053335594 2425741.65365
8      3.00046413631 42.4578908262 6584960.64527
9      2.99826217175 33.9774057032 4377163.92144
10     3.00179083265 40.0400588519 5800546.526
[ec2-user@ip-172-31-3-174 hadoop-2.6.4]$
```

Creation of new centers.txt file from the third round clustering output (including nano to remove center assignments from the file):



```
ec2-user@ip-172-31-3-174:~$ rm part-00000
[ec2-user@ip-172-31-3-174 ~]$ rm centers.txt
[ec2-user@ip-172-31-3-174 ~]$ hadoop fs -get /data/streaming_cluster3/part-00000 .
[ec2-user@ip-172-31-3-174 ~]$ ls
apache-hive-2.0.1-bin          dwdate.tbl           pig-0.15.0
apache-mahout-distribution-0.11.2 hadoop-2.6.4      proj_p4q03_mapper.py
apache-mahout-distribution-0.11.2.zip hbase-0.90.3    proj_p4q03_reducer.py
bioproject.xml                  hbase-0.90.3.tar   proj_part3_load.pig
centers_1.txt                   lineorder.tbl     proj_part3_q01.pig
centers_2.txt                   lo_sample.csv    proj_part3_q02.pig
centers_orig.txt                lo_sample.tbl    proj_part3_q03.pig
clusteranalyze_second.txt       lo_threecol.txt  proj_phase1
clusteranalyze.txt               man_cluster_test.txt spark-2.1.0-bin-hadoop2.6
cluster_mapper.py               MovieLens         spark-2.1.0-bin-hadoop2.6.tar
cluster_reducer.py              myHadoop.tar    supplier.tbl
cluster_test.txt                NumbersAndStrings.txt synthetic_control.data
csv_mapper.py                  numStrings.py   threecol_mapper.py
csv_reducer.py                 odd_num_mapper.py threecol_reducer.py
customer.tbl                   part-00000      threecol_sample.txt
date_mapper.py                 part-r-00000    workerSPark.tar
distinct_reducer.py            part.tbl
[ec2-user@ip-172-31-3-174 ~]$ cp part-00000 centers.txt
[ec2-user@ip-172-31-3-174 ~]$ cat centers.txt
0      2.99895583182 25.5220051296 3274180.65921
1      2.99837051646 38.2480537409 5247846.76039
2      3.00110225816 4.05573179891 520624.424623
3      3.00243448515 30.480664043 3912052.81031
4      3.00219609991 11.3869745457 1461073.24183
5      3.00085332274 36.4105552202 4790835.31827
6      3.00018288565 46.1422146984 7879282.026
7      3.00137369906 18.9053335594 2425741.65365
8      3.00046413631 42.4578908262 6584960.64527
9      2.99826217175 33.9774057032 4377163.92144
10     3.00179083265 40.0400588519 5800546.526
[ec2-user@ip-172-31-3-174 ~]$
```

Part 4

I ran all of the Pig, Hive, and Hadoop Streaming commands on a 1 node cluster to compare to the 4-node results. I did not include the code I used in this section because it is the same that was used in part 1 (thought I would make this section more compact and hopefully easier to read). I also did not include the screenshots of the output directory sizes or contents, since these also are the same as in part 1. I downloaded all of the scale 4 tables to the 1 node cluster Linux ec2-user home directory using wget. I then copied them into the HDFS /user/ec2-user/ directory. I used the same 1-node cluster that was used for assignments 1 through 3. It was a t2 medium instance with 12 GB hard drive space. The HDFS replication factor was set to 1 (since there was only 1 node).

Note that I had issues with Pig running after I had run the Hadoop Streaming and Hive code and had to reformat the HDFS in order to get Pig to work (ran a small vehicles data query to check, which failed). I then found that I couldn't run Hive code after the Pig code had worked and needed to reformat the HDFS again to get Hive to work again. I'm not sure what the issue is with the 1 node cluster because it worked fine for assignments 1 through 3. It is possible that the hard drive space was not adequate (was only set to 12G). Since I had already reformatted the HDFS to fix Pig and then Hive, I was not able to check the size present during the failing cases. Since the lineorder.tbl was a couple gigabytes in size (present in both the Linux file system and HDFS), and the outputs of the Hadoop Streaming and Hive csv file transformations were each a couple gigabytes, it is possible the issue was that the HDFS was running out of space.

A)

1 Node Hive lineorder.tbl Transformation From | Separated to CSV

1 Node Hive Transformation Execution:

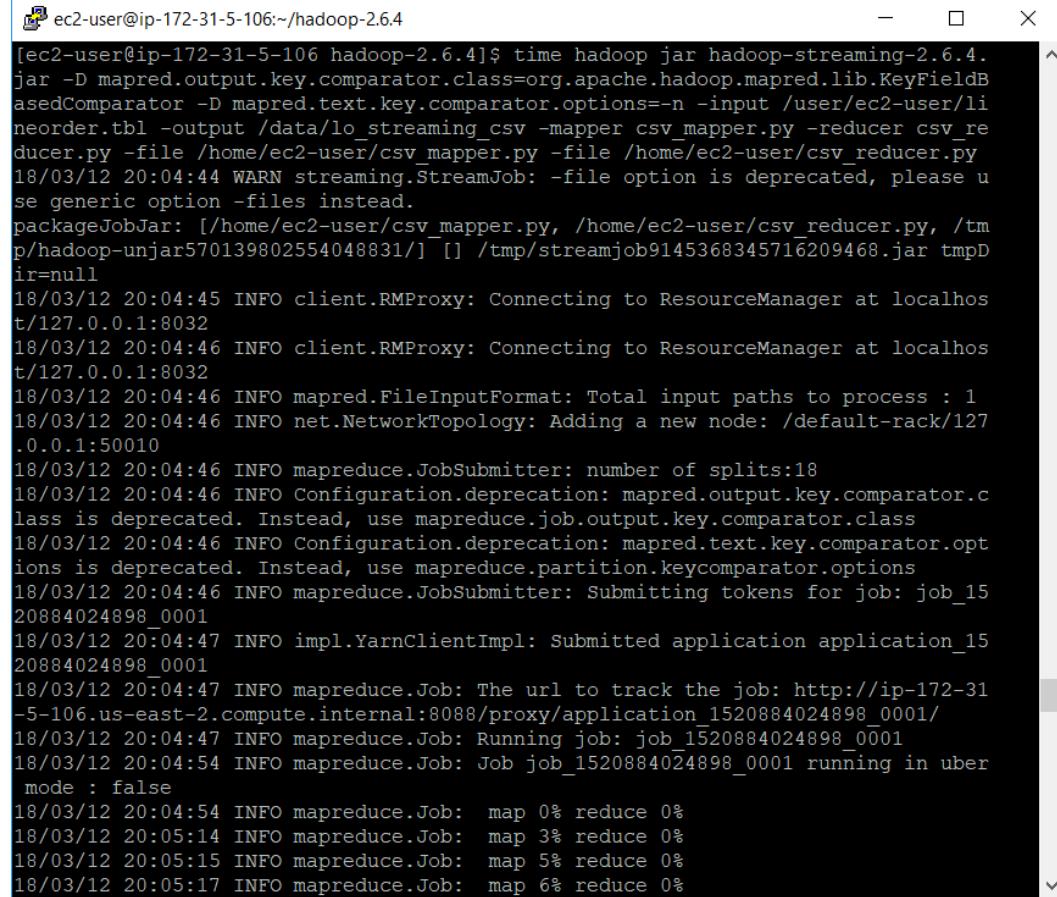
```
ec2-user@ip-172-31-5-106:~/apache-hive-2.0.1-bin
hive> insert overwrite directory 'lo_hive_csv'
  > row format delimited
  > fields terminated by ','
  > stored as textfile
  > select * from lineorder;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20180313025827_b5f56417-cd35-48b5-b0b1-ae74881101c9
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520909480307_0001, Tracking URL = http://ip-172-31-5-106.us-east-2.compute.internal:8088/proxy/application_1520909480307_0001/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520909480307_0001
Hadoop job information for Stage-1: number of mappers: 10; number of reducers: 0
2018-03-13 02:58:36,740 Stage-1 map = 0%,  reduce = 0%
2018-03-13 02:59:22,113 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 55.45 sec
2018-03-13 02:59:25,434 Stage-1 map = 30%,  reduce = 0%, Cumulative CPU 60.96 sec
2018-03-13 02:59:46,301 Stage-1 map = 45%,  reduce = 0%, Cumulative CPU 99.26 sec
2018-03-13 02:59:47,384 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 100.96 sec
2018-03-13 03:00:02,984 Stage-1 map = 70%,  reduce = 0%, Cumulative CPU 111.92 sec
2018-03-13 03:00:19,292 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 131.07 sec
2018-03-13 03:00:20,402 Stage-1 map = 85%,  reduce = 0%, Cumulative CPU 133.02 sec
2018-03-13 03:00:38,904 Stage-1 map = 90%,  reduce = 0%, Cumulative CPU 151.73 sec
2018-03-13 03:00:39,934 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 153.0 sec
MapReduce Total cumulative CPU time: 2 minutes 33 seconds 0 msec
Ended Job = job_1520909480307_0001
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://localhost/user/ec2-user/lo_hive_csv/.hive-staging_hive_2018-03-13_02-58-27_225_480077066791790743-1/-ext-10000
Moving data to: lo_hive_csv
MapReduce Jobs Launched:
```

```
ec2-user@ip-172-31-5-106:~/apache-hive-2.0.1-bin
rsions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20180313025827_b5f56417-cd35-48b5-b0b1-ae74881101c9
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520909480307_0001, Tracking URL = http://ip-172-31-5-106.us-east-2.compute.internal:8088/proxy/application_1520909480307_0001/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520909480307_0001
Hadoop job information for Stage-1: number of mappers: 10; number of reducers: 0
2018-03-13 02:58:36,740 Stage-1 map = 0%,  reduce = 0%
2018-03-13 02:59:22,113 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 55.45 sec
2018-03-13 02:59:25,434 Stage-1 map = 30%,  reduce = 0%, Cumulative CPU 60.96 sec
2018-03-13 02:59:46,301 Stage-1 map = 45%,  reduce = 0%, Cumulative CPU 99.26 sec
2018-03-13 02:59:47,384 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 100.96 sec
2018-03-13 03:00:02,984 Stage-1 map = 70%,  reduce = 0%, Cumulative CPU 111.92 sec
2018-03-13 03:00:19,292 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 131.07 sec
2018-03-13 03:00:20,402 Stage-1 map = 85%,  reduce = 0%, Cumulative CPU 133.02 sec
2018-03-13 03:00:38,904 Stage-1 map = 90%,  reduce = 0%, Cumulative CPU 151.73 sec
2018-03-13 03:00:39,934 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 153.0 sec
MapReduce Total cumulative CPU time: 2 minutes 33 seconds 0 msec
Ended Job = job_1520909480307_0001
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://localhost/user/ec2-user/lo_hive_csv/.hive-staging_hive_2018-03-13_02-58-27_225_480077066791790743-1-ext-10000
Moving data to: lo_hive_csv
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Cumulative CPU: 153.0 sec  HDFS Read: 2417901954 HDFS Write: 2417756563 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 33 seconds 0 msec
OK
Time taken: 133.913 seconds
hive>
```

The Hive transformation code took 133.913 secs to execute on the 1-node cluster. This is compared to the 50.157 secs it took on a 4-node cluster. It took 2.67 times longer to run on the 1-node cluster than the 4-node one.

1-Node Hadoop Streaming lineorder.tbl Transformation From | Separated to CSV

1-Node Hadoop Streaming Transformation Execution (first and last screenshots only due to length):



The terminal window shows the command being run:

```
[ec2-user@ip-172-31-5-106 hadoop-2.6.4]$ time hadoop jar hadoop-streaming-2.6.4.jar -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D mapred.text.key.comparator.options=-n -input /user/ec2-user/lineorder.tbl -output /data/lo_streaming_csv -mapper csv_mapper.py -reducer csv_reducer.py -file /home/ec2-user/csv_mapper.py -file /home/ec2-user/csv_reducer.py
```

Output logs from the job execution:

```
18/03/12 20:04:44 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/ec2-user/csv_mapper.py, /home/ec2-user/csv_reducer.py, /tmp/hadoop-unjar570139802554048831/] [] /tmp/streamjob9145368345716209468.jar tmpDir=null
18/03/12 20:04:45 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/12 20:04:46 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/12 20:04:46 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/12 20:04:46 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:50010
18/03/12 20:04:46 INFO mapreduce.JobSubmitter: number of splits:18
18/03/12 20:04:46 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Instead, use mapreduce.job.output.key.comparator.class
18/03/12 20:04:46 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Instead, use mapreduce.partition.keycomparator.options
18/03/12 20:04:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1520884024898_0001
18/03/12 20:04:47 INFO impl.YarnClientImpl: Submitted application application_1520884024898_0001
18/03/12 20:04:47 INFO mapreduce.Job: The url to track the job: http://ip-172-31-5-106.us-east-2.compute.internal:8088/proxy/application_1520884024898_0001/
18/03/12 20:04:47 INFO mapreduce.Job: Running job: job_1520884024898_0001
18/03/12 20:04:54 INFO mapreduce.Job: Job job_1520884024898_0001 running in uber mode : false
18/03/12 20:04:54 INFO mapreduce.Job: map 0% reduce 0%
18/03/12 20:05:14 INFO mapreduce.Job: map 3% reduce 0%
18/03/12 20:05:15 INFO mapreduce.Job: map 5% reduce 0%
18/03/12 20:05:17 INFO mapreduce.Job: map 6% reduce 0%
```

```
ec2-user@ip-172-31-5-106:~/hadoop-2.6.4
Map output materialized bytes=2465749879
Input split bytes=1728
Combine input records=0
Combine output records=0
Reduce input groups=6000000
Reduce shuffle bytes=2465749879
Reduce input records=23996604
Reduce output records=23996604
Spilled Records=83924039
Shuffled Maps =18
Failed Shuffles=0
Merged Map outputs=18
GC time elapsed (ms)=2841
CPU time spent (ms)=326540
Physical memory (bytes) snapshot=4742082560
Virtual memory (bytes) snapshot=18831822848
Total committed heap usage (bytes)=3377463296
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=2417826195
File Output Format Counters
Bytes Written=2441753167
18/03/12 20:09:48 INFO streaming.StreamJob: Output directory: /data/lo_streaming
_csv
real    5m4.985s
user    0m4.700s
sys     0m0.320s
[ec2-user@ip-172-31-5-106 hadoop-2.6.4]$
```

The Hadoop Streaming transformation code took 5 min 4.985 secs (or 304.985 secs) to execute on the 1-node cluster. This is compared to the 3 min and 17.27 secs (or 197.27 secs) on the 4-node cluster. It took 1.54 times longer to run on the 1-node cluster than the 4-node one.

1-Node Pig lineorder.tbl Transformation From | Separated to CSV

Pig Transformation Execution (only showing start and finish due to length):

```
ec2-user@ip-172-31-5-106:~/pig-0.15.0
grunt> lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
>> AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_suppkey:INT, lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARRAY, lo_quantity:INT, lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:INT, lo_revenue:INT, lo_supplycost:INT, lo_tax:INT, lo_commitdate:INT, lo_shipmode:CHARARRAY);
2018-03-13 02:03:32,501 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> store lineorder into 'lo_pig_csv'
>> using PigStorage(',');
2018-03-13 02:03:37,124 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-13 02:03:37,160 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2018-03-13 02:03:37,174 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2018-03-13 02:03:37,190 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-13 02:03:37,191 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-03-13 02:03:37,191 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-13 02:03:37,198 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-13 02:03:37,200 [main] INFO org.apache.pig.backend.hadoop.executionengi
```

```

ec2-user@ip-172-31-5-106:~/pig-0.15.0
2018-03-13 02:06:47,634 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:06:47,640 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-13 02:06:47,862 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:06:47,866 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-13 02:06:47,908 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:06:47,911 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-13 02:06:47,933 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-13 02:06:47,933 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.4 0.15.0 ec2-user 2018-03-13 02:03:37 2018-03-13 02:06:47 UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime
pTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1520905372677_0003 18 0 60 56 57 57 0 0
0 0 lineorder MAP_ONLY hdfs://localhost/user/ec2-user/lo_pig_csv,
Input(s):
Successfully read 23996604 records (2417832873 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 23996604 records (2417756563 bytes) in: "hdfs://localhost/user/ec2-user/lo_pig_csv"

Counters:
Total records written : 23996604
Total bytes written : 2417756563
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520905372677_0003

2018-03-13 02:06:47,934 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:06:47,937 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

```

```
ec2-user@ip-172-31-5-106:~/pig-0.15.0

Counters:
Total records written : 23996604
Total bytes written : 2417756563
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520905372677_0003

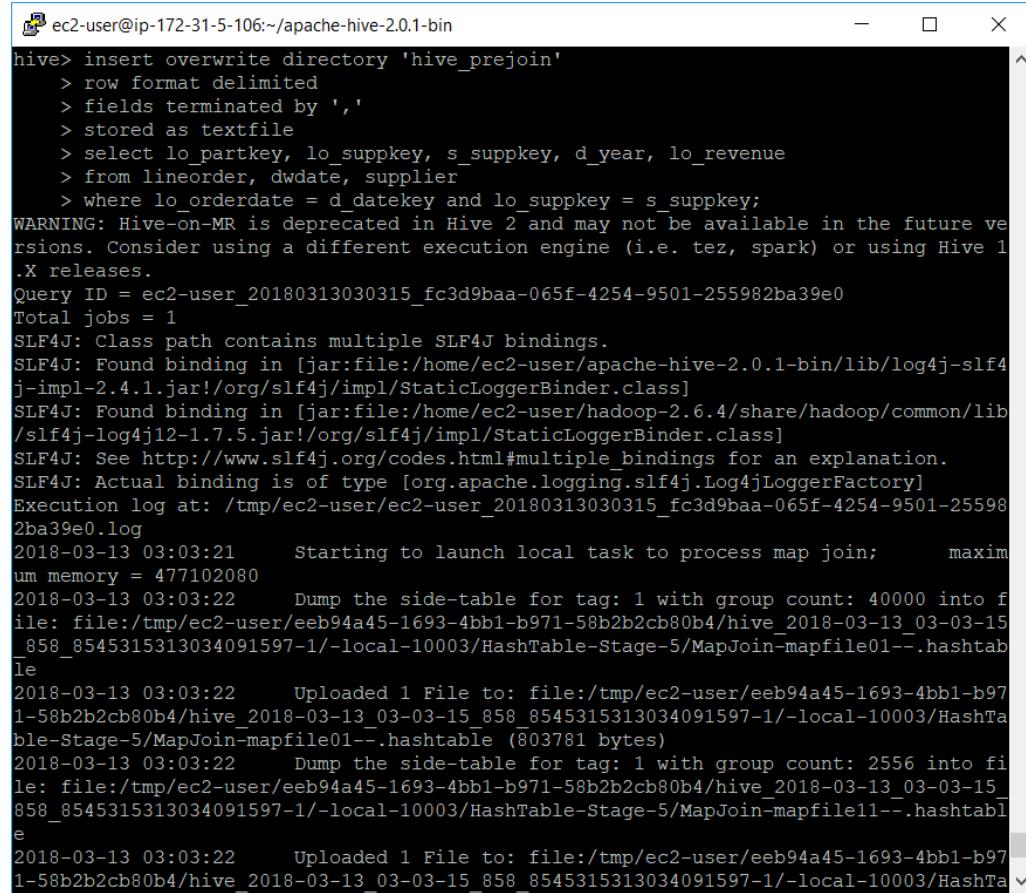
2018-03-13 02:06:47,934 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:06:47,937 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-13 02:06:48,034 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:06:48,039 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-13 02:06:48,059 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:06:48,064 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-03-13 02:06:48,088 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

The Pig transformation code started at 2:03:37 and ended at 2:06:47, so it took 3 min 10 secs (or 190 secs) to execute on the 1-node cluster. This is in comparison to the 1 min 13 secs (or 73 secs) it took on a 4-node cluster. It took 2.6 times longer to run on the 1-node cluster than the 4-node one.

B)

1-Node Hive Pre-Join

1-Node Hive Pre-Join Execution:



The screenshot shows a terminal window titled 'ec2-user@ip-172-31-5-106:~\$ apache-hive-2.0.1-bin'. The command entered is:

```
hive> insert overwrite directory 'hive_prejoin'
  > row format delimited
  > fields terminated by ','
  > stored as textfile
  > select lo_partkey, lo_suppkey, s_suppkey, d_year, lo_revenue
  > from lineorder, dwdate, supplier
  > where lo_orderdate = d_datekey and lo_suppkey = s_suppkey;
```

Output messages include:

- WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez, spark) or using Hive 1.X releases.
- Query ID = ec2-user_20180313030315_fc3d9baa-065f-4254-9501-255982ba39e0
- Total jobs = 1
- SLF4J: Class path contains multiple SLF4J bindings.
- SLF4J: Found binding in [jar:file:/home/ec2-user/apache-hive-2.0.1-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
- SLF4J: Found binding in [jar:file:/home/ec2-user/hadoop-2.6.4/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
- SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
- SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
- Execution log at: /tmp/ec2-user/ec2-user_20180313030315_fc3d9baa-065f-4254-9501-255982ba39e0.log
- 2018-03-13 03:03:21 Starting to launch local task to process map join; maximum memory = 477102080
- 2018-03-13 03:03:22 Dump the side-table for tag: 1 with group count: 40000 into file: file:/tmp/ec2-user/eec94a45-1693-4bb1-b971-58b2b2cb80b4/hive_2018-03-13_03-03-15_858_8545315313034091597-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile01--.hashtable
- 2018-03-13 03:03:22 Uploaded 1 File to: file:/tmp/ec2-user/eec94a45-1693-4bb1-b971-58b2b2cb80b4/hive_2018-03-13_03-03-15_858_8545315313034091597-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile01--.hashtable (803781 bytes)
- 2018-03-13 03:03:22 Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/eec94a45-1693-4bb1-b971-58b2b2cb80b4/hive_2018-03-13_03-03-15_858_8545315313034091597-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile11--.hashtable
- 2018-03-13 03:03:22 Uploaded 1 File to: file:/tmp/ec2-user/eec94a45-1693-4bb1-b971-58b2b2cb80b4/hive_2018-03-13_03-03-15_858_8545315313034091597-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile11--.hashtable (1024 bytes)

```
ec2-user@ip-172-31-5-106:~/apache-hive-2.0.1-bin
ble-Stage-5/MapJoin-mapfile01--.hashtable (803781 bytes)
2018-03-13 03:03:22      Dump the side-table for tag: 1 with group count: 2556 into fi
le: file:/tmp/ec2-user/eeb94a45-1693-4bb1-b971-58b2b2cb80b4/hive_2018-03-13_03-03-15_
858_8545315313034091597-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile11--.hashtable
2018-03-13 03:03:22      Uploaded 1 File to: file:/tmp/ec2-user/eeb94a45-1693-4bb1-b97
1-58b2b2cb80b4/hive_2018-03-13_03-03-15_858_8545315313034091597-1/-local-10003/HashTable-Stage-5/MapJoin-mapfile11--.hashtable (67039 bytes)
2018-03-13 03:03:22      End of local task; Time Taken: 1.552 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1520909480307_0002, Tracking URL = http://ip-172-31-5-106.us-east-
2.compute.internal:8088/proxy/application_1520909480307_0002/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1520909480307_00
02
Hadoop job information for Stage-5: number of mappers: 10; number of reducers: 0
2018-03-13 03:03:29,202 Stage-5 map = 0%,  reduce = 0%
2018-03-13 03:04:14,976 Stage-5 map = 30%,  reduce = 0%, Cumulative CPU 57.5 sec
2018-03-13 03:04:32,436 Stage-5 map = 45%,  reduce = 0%, Cumulative CPU 88.23 sec
2018-03-13 03:04:33,469 Stage-5 map = 60%,  reduce = 0%, Cumulative CPU 92.19 sec
2018-03-13 03:04:49,957 Stage-5 map = 70%,  reduce = 0%, Cumulative CPU 104.01 sec
2018-03-13 03:05:03,935 Stage-5 map = 75%,  reduce = 0%, Cumulative CPU 122.58 sec
2018-03-13 03:05:07,151 Stage-5 map = 85%,  reduce = 0%, Cumulative CPU 125.42 sec
2018-03-13 03:05:21,185 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 140.75 sec
MapReduce Total cumulative CPU time: 2 minutes 20 seconds 750 msec
Ended Job = job_1520909480307_0002
Moving data to: hive_prejoin
MapReduce Jobs Launched:
Stage-Stage-5: Map: 10  Cumulative CPU: 140.75 sec  HDFS Read: 2417955764 HDFS Writ
e: 744597239 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 20 seconds 750 msec
OK
Time taken: 126.464 seconds
hive>
```

The Hive Pre-Join code took 126.464 secs to execute on a 1-node cluster. This is compared to the 48.47 secs it took on a 4-node cluster. It took 2.61 times longer to run on the 1-node cluster than the 4-node one.

1-Node Pig Pre-Join

Pig Pre-Join Execution:

```
ec2-user@ip-172-31-5-106:~/pig-0.15.0
grunt> lineorder = LOAD '/user/ec2-user/lineorder.tbl' USING PigStorage('|')
>> AS (lo_orderkey:INT, lo_linenumber:INT, lo_custkey:INT, lo_partkey:INT, lo_suppkey:INT, lo_orderdate:INT, lo_orderpriority:CHARARRAY, lo_shippriority:CHARARRAY, lo_quantity:INT, lo_extendedprice:INT, lo_ordertotalprice:INT, lo_discount:INT, lo_revenue:INT, lo_supplycost:INT, lo_tax:INT, lo_commitdate:INT, lo_shipmode:CHARARRAY);
2018-03-13 02:14:55,145 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-13 02:14:55,184 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> dwdate = LOAD '/user/ec2-user/dwdate.tbl' USING PigStorage('|')
>> AS (d_datekey:INT, d_date:CHARARRAY, d_dayofweek:CHARARRAY, d_month:CHARARRAY, d_year:INT, d_yeарmonthnum:INT, d_yeарmonth:CHARARRAY, d_daynuminweek:INT, d_dаynuminmonth:INT, d_daynuminyear:INT, d_monthnuminyear:INT, d_weeknuminyear:INT, d_sellingseason:CHARARRAY, d_lastdayinweekfl:CHARARRAY, d_lastdayinmonthfl:CHARARRAY, d_holidayfl:CHARARRAY, d_weekdayfl:CHARARRAY);
2018-03-13 02:15:02,003 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> supplier = LOAD '/user/ec2-user/supplier.tbl' USING PigStorage('|')
>> AS (s_suppkey:INT, s_name:CHARARRAY, s_address:CHARARRAY, s_city:CHARARRAY, s_nation:CHARARRAY, s_region:CHARARRAY, s_phone:CHARARRAY);
2018-03-13 02:15:06,625 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> first_join = JOIN lineorder BY lo_orderdate, dwdate BY d_datekey;
grunt> lo_and_date = FOREACH first_join GENERATE lineorder:::lo_partkey AS lo_partkey, lineorder:::lo_suppkey AS lo_suppkey, lineorder:::lo_revenue AS lo_revenue, dwdate:::d_year AS d_year;
grunt> second_join = JOIN lo_and_date BY lo_suppkey, supplier BY s_suppkey;
grunt> supp_lo_date = FOREACH second_join GENERATE lo_and_date:::lo_partkey AS lo_partkey, lo_and_date:::lo_suppkey AS lo_suppkey, lo_and_date:::lo_revenue AS lo_r
```

```
ec2-user@ip-172-31-5-106:~/pig-0.15.0
_partkey, lo_and_date:::lo_suppkey AS lo_suppkey, lo_and_date:::lo_revenue AS lo_revenue, lo_and_date:::d_year AS d_year, supplier:::s_suppkey AS s_suppkey;
grunt> store supp_lo_date into 'pig_prejoin' using PigStorage(',');
2018-03-13 02:15:21,667 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-13 02:15:21,749 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2018-03-13 02:15:21,766 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-03-13 02:15:21,766 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2018-03-13 02:15:21,766 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2018-03-13 02:15:21,776 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for supplier: $1, $2, $3, $4, $5, $6
2018-03-13 02:15:21,777 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for lineorder: $0, $1, $2, $6, $7, $8, $9, $10, $11, $12, $13, $14, $15, $16
2018-03-13 02:15:21,781 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-03-13 02:15:21,786 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
```

```

ec2-user@ip-172-31-5-106:~/pig-0.15.0
ing to job history server
2018-03-13 02:21:58,681 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-03-13 02:21:58,690 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.4 0.15.0 ec2-user 2018-03-13 02:15:21 2018-03-13 02:21:58 H
ASH_JOIN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime
pTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime A
lias Feature Outputs
job_1520905372677_0004 19 3 39 6 32 37 99 9
7 97 97 dwdate,first_join,lineorder,lo_and_date HASH_JOIN
job_1520905372677_0005 5 1 101 15 77 99 174 1
74 174 174 second_join,supp_lo_date,supplier HASH_JOIN
hdfs://localhost/user/ec2-user/pig_prejoin,

Input(s):
Successfully read 23996604 records from: "/user/ec2-user/lineorder.tbl"
Successfully read 2556 records from: "/user/ec2-user/dwdate.tbl"
Successfully read 40000 records from: "/user/ec2-user/supplier.tbl"

Output(s):
Successfully stored 23996604 records (744597239 bytes) in: "hdfs://localhost/use
r/ec2-user/pig_prejoin"

ec2-user@ip-172-31-5-106:~/pig-0.15.0
r/ec2-user/pig_prejoin"

Counters:
Total records written : 23996604
Total bytes written : 744597239
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1520905372677_0004 -> job_1520905372677_0005,
job_1520905372677_0005

2018-03-13 02:21:58,691 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,694 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,792 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,795 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,835 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,847 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,873 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con

```

```
ec2-user@ip-172-31-5-106:~/pig-0.15.0
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,792 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,795 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,835 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,847 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,873 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,876 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,906 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,921 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,943 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2018-03-13 02:21:58,946 [main] INFO org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2018-03-13 02:21:58,969 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

The Pig Pre-Join code started at 2:15:21 and ended at 2:21:59, so it took 6 min 38 secs (398 secs) or to execute on a 1-node cluster. This is compared to the 4 min 4 secs (or 244 secs) it took on a 4-node cluster. It took 1.63 times longer to run on the 1-node cluster than the 4-node one.

c)

The Hive portion of part 1A ran 2.67 times slower on the 1-node cluster than the 4-node cluster (133.913 secs versus 50.157 secs). The Pig portion of part 1A ran 2.6 times slower on the 1-node cluster than the 4-node cluster (190 secs versus 73 secs). The Hadoop Streaming portion of part 1A ran 1.54 times slower on the 1-node cluster versus the 4-node cluster (304.985 secs versus 197.27 secs). The Hive portion of part 2B ran 2.61 times slower on the 1-node cluster than the 4-node cluster (126.464 secs versus 48.47 secs). The Pig portion of part 2B ran 1.63 times slower on the 1-node cluster than the 4-node cluster (398 secs versus 244 secs). The run times of Hive on the 1-node cluster to 4-node cluster had very similar relations to each other between running part 1A and part 2B (both ran approximately 2.6 times slower on 1-node than 4-node). The Pig run times on the 1-node cluster and 4-node cluster had different relations to each other for part 1A and part 2B (part 1A ran 2.6 times slower on the 1-node versus 4-node where part 2B ran 1.63 times slower).

Running the tasks on 4-node clusters showed significant improvement in run time in all cases. The improvement, however, does not match the theoretical improvement predictions. Theoretically, without any network traffic delays, a 4-node cluster should run 4 times faster (a 1-node task taking 12 secs would take $12/4 = 3$ secs and $12/3$ equals 4 times faster). Since none of the 4-node tasks were close the theoretical speeds, this means that network traffic delay makes up a significant portion of all of the times. The network traffic delay is made up of the time for mappers to read from HDFS, the time for reducers to write to HDFS (HDFS write with replication factor of 2 for the 4-node cluster used), and network bus delay between the cluster nodes (for example the delay from having to send the mapper output from one node's local disk to another node that has the reducer required). Another factor is that the 1-node cluster was created with 12 GB of hard drive space, where each instance in the 4-node cluster had 30 GB. There was an issue running the Pig tasks after running the Hive tasks (Pig froze at 0% of the map/reduce stage on all queries, even simple ones on a smaller dataset). HDFS had to be reformatted between the Hive and Pig runs in order to get Pig to work. When the Hive tasks were run again after Pig was run, the same behavior was seen in Hive (was frozen at start of map/reduce). HDFS had to be reformatted again to get Hive to run again. This behavior was most likely due to the fact that the 1-node cluster only had 12 GB of hard drive space. Each output of part 1A was approximately 2.4 GB (one for the Hadoop Streaming, one for Hive, and one for Pig). On top of that, the lineorder table was 2.4 GB in size as well. This table was present in the Linux file system and the HDFS. This means that after Hadoop Streaming and Hive were run, there was at least $2.4*4 = 9.6$ GB of hard drive space on the instance taken (streaming output, Hive output, Linux table, and HDFS table = 4 items 2.4 GB in size). This doesn't include the approximately 80 MB each for the part 2B outputs. It is possible that as the 1-node instance hard drive started to get filled, writing and reading to its HDFS could have been affected (could have been slowed down). The 4-node cluster HDFS was only 29% full, so this wouldn't have been an issue for the 4-node cluster. The 4-node cluster however did have a replication factor of 2, so the time to write to the HDFS would be double the time of the 1-node (without actual network bus delays for writing between nodes).

This experiment did show that running on a 4-node cluster significantly improves run time, but unfortunately since all of the variables were not held constant (4-node instances had more hard drive space than 1-node, 4-node cluster had replication of 2 where 1-node had replication of 1), it is hard to determine how much of the improvement was due to the presence of more nodes in the cluster alone. In other words, it is hard to determine exactly how increase in cluster size affected run time due to the other mitigating factors.