Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Part 1: Data Transformation

Data used for Part 1: Scale4

1A. Transform lineorder.tbl table into a csv (comma-separated file)

Method 1: 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;
LOAD DATA LOCAL INPATH '/home/ec2-user/lineorder.tbl'
OVERWRITE INTO TABLE lineorder;

INSERT OVERWRITE DIRECTORY 'lineorder'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
SELECT * FROM lineorder;
```

Table create and Load:

```
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: 1.177 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/lineorder.tbl'
    > OVERWRITE INTO TABLE lineorder;
Loading data to table default.lineorder
OK
Time taken: 44.309 seconds
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Transforming the lineorder.tbl to a csv file: Time taken – 54.594 seconds.

```
hive> INSERT OVERWRITE DIRECTORY 'lineorder'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > 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_20161119174817_594d4c8b-7abc-466a-8dcd-168f680c69c6
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_1479510364858_0010, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0010/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0010
Hadoop job information for Stage-1: number of mappers: 10; number of reducers: 0
2016-11-19 17:48:24,077 Stage-1 map = 0%,  reduce = 0%
2016-11-19 17:48:39,313 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 24.8 sec
2016-11-19 17:48:48,193 Stage-1 map = 15%,  reduce = 0%, Cumulative CPU 70.78 sec
2016-11-19 17:48:49,328 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 75.99 sec
2016-11-19 17:48:50,468 Stage-1 map = 30%,  reduce = 0%, Cumulative CPU 84.5 sec
2016-11-19 17:48:52,651 Stage-1 map = 35%,  reduce = 0%, Cumulative CPU 89.81 sec
2016-11-19 17:48:53,751 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 98.68 sec
2016-11-19 17:48:55,953 Stage-1 map = 55%,  reduce = 0%, Cumulative CPU 105.7 sec
2016-11-19 17:48:58,209 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 116.97 sec
2016-11-19 17:49:00,481 Stage-1 map = 70%,  reduce = 0%, Cumulative CPU 123.69 sec
2016-11-19 17:49:01,555 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 125.36 sec
2016-11-19 17:49:04,777 Stage-1 map = 80%,  reduce = 0%, Cumulative CPU 133.87 sec
2016-11-19 17:49:07,925 Stage-1 map = 90%,  reduce = 0%, Cumulative CPU 140.71 sec
2016-11-19 17:49:08,955 Stage-1 map = 95%,  reduce = 0%, Cumulative CPU 142.74 sec
2016-11-19 17:49:11,024 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 145.78 sec
MapReduce Total cumulative CPU time: 2 minutes 25 seconds 780 msec
Ended Job = job_1479510364858_0010
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://172.31.46.250/user/ec2-user/lineorder/.hive-staging_hive_2016-11-19_17-48-17_643_4221350397587353735-1/-ext-10000
Moving data to: lineorder
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10   Cumulative CPU: 145.78 sec   HDFS Read: 2417902180 HDFS Write: 2417756563 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 25 seconds 780 msec
OK
Time taken: 54.594 seconds
hive>
```

Output file(s):

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/lineorder
Found 10 items
-rwxr-xr-x   2 ec2-user supergroup  268435390 2016-11-19 17:49 /user/ec2-user/lineorder/000000_0
-rwxr-xr-x   2 ec2-user supergroup  268435553 2016-11-19 17:49 /user/ec2-user/lineorder/000001_0
-rwxr-xr-x   2 ec2-user supergroup  268435475 2016-11-19 17:49 /user/ec2-user/lineorder/000002_0
-rwxr-xr-x   2 ec2-user supergroup  268435446 2016-11-19 17:48 /user/ec2-user/lineorder/000003_0
-rwxr-xr-x   2 ec2-user supergroup  268435487 2016-11-19 17:48 /user/ec2-user/lineorder/000004_0
-rwxr-xr-x   2 ec2-user supergroup  268435435 2016-11-19 17:49 /user/ec2-user/lineorder/000005_0
-rwxr-xr-x   2 ec2-user supergroup  268435526 2016-11-19 17:48 /user/ec2-user/lineorder/000006_0
-rwxr-xr-x   2 ec2-user supergroup  268435355 2016-11-19 17:48 /user/ec2-user/lineorder/000007_0
-rwxr-xr-x   2 ec2-user supergroup  268435500 2016-11-19 17:49 /user/ec2-user/lineorder/000008_0
-rwxr-xr-x   2 ec2-user supergroup    1837396 2016-11-19 17:48 /user/ec2-user/lineorder/000009_0
[ec2-user@ip-172-31-46-250 ~]$
```

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/lineorder/000009_0
23982182,4,80857,335011,10879,19961209,3-MEDIUM,0,23,2405800,17586382,3,2333626,62760,7,19970118,AIR
23982182,5,80857,129113,29491,19961209,3-MEDIUM,0,6,685266,17586382,4,657855,68526,4,19970120,TRUCK
23982183,1,75700,216120,1197,19951202,1-URGENT,0,26,2693886,8475336,2,2640008,62166,5,19960115,AIR
23982183,2,75700,315456,30492,19951202,1-URGENT,0,19,2795736,8475336,10,2516162,88286,3,19960213,AIR
23982183,3,75700,259318,22241,19951202,1-URGENT,0,21,2682330,8475336,4,2575036,76638,6,19960219,RAIL
23982183,4,75700,348367,309,19951202,1-URGENT,0,3,424605,8475336,10,382144,84921,0,19960120,AIR
23982208,1,43952,354065,39505,19930126,3-MEDIUM,0,1,111905,16835761,4,107428,67143,1,19930307,SHIP
23982208,2,43952,365955,26654,19930126,3-MEDIUM,0,27,5456538,16835761,2,5347407,121256,5,19930225,FOB
23982208,3,43952,193085,6494,19930126,3-MEDIUM,0,33,3887664,16835761,10,3498897,70684,2,19930310,SHIP
23982208,4,43952,366973,15796,19930126,3-MEDIUM,0,36,7343856,16835761,4,7050101,122397,7,19930414,AIR
23982209,1,100858,312175,3873,19970520,5-LOW,0,33,3917628,10163071,1,3878451,71229,7,19970727,FOB
```

Method 2: PIG

```
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_comitdate:int,
lo_shipmode:chararray);
DESCRIBE lineorder;

STORE lineorder INTO '/user/ec2-user/pig_lineorder_csv/' USING PigStorage(',');
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Transforming lineorder.tbl into a csv file:

```
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_comitdate:int, lo_shipmode:chararray);
2016-11-19 19:37:45,017 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> DESCRIBE lineorder;
lineorder: {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_comitdate: int,lo_shipmode: chararray}
grunt> STORE lineorder INTO '/user/ec2-user/pig_lineorder_csv/' USING PigStorage(',');
2016-11-19 19:38:06,776 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:38:06,800 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2016-11-19 19:38:06,815 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2016-11-19 19:38:06,838 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:38:06,859 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-19 19:38:06,905 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-19 19:38:07,007 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-19 19:38:07,032 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2016-11-19 19:38:07,032 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2016-11-19 19:38:07,051 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:38:07,075 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:38:07,228 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-11-19 19:38:07,233 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2016-11-19 19:38:07,233 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-11-19 19:38:07,233 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2016-11-19 19:38:07,235 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-11-19 19:38:07,449 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/pig-0.15.0-core-h2.jar to DistributedCache t
hrough /tmp/temp-1358699710/tmp-1971117342/pig-0.15.0-core-h2.jar
2016-11-19 19:38:07,476 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/automaton-1.11-8.jar to DistributedCache
 through /tmp/temp-1358699710/tmp-252694523/automaton-1.11-8.jar
2016-11-19 19:38:07,504 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/antlr-runtime-3.4.jar to DistributedCach
e through /tmp/temp-1358699710/tmp-1621201452/antlr-runtime-3.4.jar
2016-11-19 19:38:07,530 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/joda-time-2.5.jar to DistributedCache th
rough /tmp/temp-1358699710/tmp882416738/joda-time-2.5.jar
2016-11-19 19:38:07,560 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2016-11-19 19:38:07,567 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-11-19 19:38:07,567 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
```

```
2016-11-19 19:39:34,394 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-19 19:39:34,395 [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        2016-11-19 19:38:07     2016-11-19 19:39:34     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1479510364858_0015  18      0       77      31      59      62      0       0       0       0       lineorder       MAP_ONLY        /user/ec2-user/pig_lineorder_csv,

Input(s):
Successfully read 23996604 records (2417832945 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 23996604 records (2417756563 bytes) in: "/user/ec2-user/pig_lineorder_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_1479510364858_0015


2016-11-19 19:39:34,397 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:39:34,404 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:39:34,492 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:39:34,496 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:39:34,524 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:39:34,529 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:39:34,561 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Output file:

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/pig_lineorder_csv
Found 19 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/_SUCCESS
-rw-r--r--   2 ec2-user supergroup  136055151 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00000
-rw-r--r--   2 ec2-user supergroup  134217768 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00001
-rw-r--r--   2 ec2-user supergroup  134217785 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00002
-rw-r--r--   2 ec2-user supergroup  134217685 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00003
-rw-r--r--   2 ec2-user supergroup  134217750 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00004
-rw-r--r--   2 ec2-user supergroup  134217705 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00005
-rw-r--r--   2 ec2-user supergroup  134217725 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00006
-rw-r--r--   2 ec2-user supergroup  134217716 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00007
-rw-r--r--   2 ec2-user supergroup  134217755 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00008
-rw-r--r--   2 ec2-user supergroup  134217732 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00009
-rw-r--r--   2 ec2-user supergroup  134217678 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00010
-rw-r--r--   2 ec2-user supergroup  134217730 2016-11-19 19:38 /user/ec2-user/pig_lineorder_csv/part-m-00011
-rw-r--r--   2 ec2-user supergroup  134217757 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00012
-rw-r--r--   2 ec2-user supergroup  134217782 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00013
-rw-r--r--   2 ec2-user supergroup  134217659 2016-11-19 19:38 /user/ec2-user/pig_lineorder_csv/part-m-00014
-rw-r--r--   2 ec2-user supergroup  134217744 2016-11-19 19:38 /user/ec2-user/pig_lineorder_csv/part-m-00015
-rw-r--r--   2 ec2-user supergroup  134217696 2016-11-19 19:39 /user/ec2-user/pig_lineorder_csv/part-m-00016
-rw-r--r--   2 ec2-user supergroup  134217745 2016-11-19 19:38 /user/ec2-user/pig_lineorder_csv/part-m-00017
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/pig_lineorder_csv/part-m-00017
21331457,4,72980,367634,10797,19921022,4-NOT SPECI,0,30,5104860,22118985,4,4900665,102097,0,19921201,FOB
21331457,5,72980,18723,8984,19921022,4-NOT SPECI,0,20,3283440,22118985,3,3184936,98503,0,19921202,TRUCK
21331457,6,72980,263926,27486,19921022,4-NOT SPECI,0,22,4157802,22118985,4,3991489,113394,1,19921201,AIR
21331458,1,43810,80161,13895,19950112,3-MEDIUM,0,35,3994060,10162421,8,3674535,68469,8,19950321,AIR
21331458,2,43810,62883,2435,19950112,3-MEDIUM,0,12,2215056,10162421,5,2104303,110752,7,19950304,SHIP
21331458,3,43810,63316,3513,19950112,3-MEDIUM,0,30,3837930,10162421,4,3684412,76758,7,19950402,SHIP
21331459,1,90397,252951,23773,19920806,4-NOT SPECI,0,3,571182,8550485,8,525487,114236,2,19921029,TRUCK
21331459,2,90397,134541,20165,19920806,4-NOT SPECI,0,22,3466188,8550485,7,3223554,94532,2,19920924,REG AIR
21331459,3,90397,20847,33071,19920806,4-NOT SPECI,0,22,3889248,8550485,0,3889248,106070,0,19920909,MAIL
21331459,4,90397,370220,23139,19920806,4-NOT SPECI,0,7,903147,8550485,10,812832,77412,3,19920926,RAIL
21331460,1,73843,240272,27875,19970202,3-MEDIUM,0,1,121226,16426060,3,117589,72735,2,19970501,TRUCK
21331460,2,73843,247879,39419,19970202,3-MEDIUM,0,41,7490126,16426060,1,7415224,109611,2,19970502,MAIL
21331460,3,73843,95218,22672,19970202,3-MEDIUM,0,38,4610198,16426060,2,4517994,72792,8,19970410,AIR
21331460,4,73843,319304,460,19970202,3-MEDIUM,0,2,264658,16426060,2,259364,79397,3,19970502,REG AIR
21331460,5,73843,133566,1743,19970202,3-MEDIUM,0,19,3039164,16426060,10,2735247,95973,6,19970311,SHIP
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

1B. Transform dwdate.tbl table into a csv (comma-separated file)

```
create table dwdate (
 d_datekey          int,
 d_date             varchar(19),
 d_dayofweek        varchar(10),
 d_month            varchar(10),
 d_year             int,
 d_yearmonthnum     int,
 d_yearmonth        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;

INSERT OVERWRITE DIRECTORY 'dwdate'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
SELECT * FROM dwdate;
```

```
hive> create table dwdate (
    >   d_datekey            int,
    >   d_date               varchar(19),
    >   d_dayofweek          varchar(10),
    >   d_month              varchar(10),
    >   d_year               int,
    >   d_yearmonthnum       int,
    >   d_yearmonth          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: 0.082 seconds
```

Transform dwdate.tbl table into a csv: Time taken 12.333 seconds

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/dwdate.tbl'
    > OVERWRITE INTO TABLE dwdate;
Loading data to table default.dwdate
OK
Time taken: 0.229 seconds
hive> INSERT OVERWRITE DIRECTORY 'dwdate'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT * FROM dwdate;
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_20161119175117_ef419316-7ce3-4c27-b0ce-f9c31f72174a
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_1479510364858_0011, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0011/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2016-11-19 17:51:22,870 Stage-1 map = 0%,  reduce = 0%
2016-11-19 17:51:29,086 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.6 sec
MapReduce Total cumulative CPU time: 1 seconds 600 msec
Ended Job = job_1479510364858_0011
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://172.31.46.250/user/ec2-user/dwdate/.hive-staging_hive_2016-11-19_17-51-17_964_6107063020845902 76-1/-ext-10000
Moving data to: dwdate
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 1.6 sec   HDFS Read: 234917 HDFS Write: 227409 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 600 msec
OK
Time taken: 12.233 seconds
hive>
```

Output file -->

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/dwdate
Found 1 items
-rwxr-xr-x   2 ec2-user supergroup     227409 2016-11-19 17:51 /user/ec2-user/dwdate/000000_0
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/dwdate/000000_0
19920101,January 1, 1992,Thursday,January,1992,199201,Jan1992,5,1,1,1,1,Winter,0,1,1,1
19920102,January 2, 1992,Friday,January,1992,199201,Jan1992,6,2,2,1,1,Winter,0,1,0,1
19920103,January 3, 1992,Saturday,January,1992,199201,Jan1992,7,3,3,1,1,Winter,1,1,0,0
19920104,January 4, 1992,Sunday,January,1992,199201,Jan1992,1,4,4,1,1,Winter,0,1,0,0
19920105,January 5, 1992,Monday,January,1992,199201,Jan1992,2,5,5,1,1,Winter,0,1,0,1
19920106,January 6, 1992,Tuesday,January,1992,199201,Jan1992,3,6,6,1,1,Winter,0,1,0,1
19920107,January 7, 1992,Wednesday,January,1992,199201,Jan1992,4,7,7,1,2,Winter,0,1,0,1
19920108,January 8, 1992,Thursday,January,1992,199201,Jan1992,5,8,8,1,2,Winter,0,1,0,1
19920109,January 9, 1992,Friday,January,1992,199201,Jan1992,6,9,9,1,2,Winter,0,1,0,1
19920110,January 10, 1992,Saturday,January,1992,199201,Jan1992,7,10,10,1,2,Winter,1,1,0,0
19920111,January 11, 1992,Sunday,January,1992,199201,Jan1992,1,11,11,1,2,Winter,0,1,0,0
19920112,January 12, 1992,Monday,January,1992,199201,Jan1992,2,12,12,1,2,Winter,0,1,0,1
19920113,January 13, 1992,Tuesday,January,1992,199201,Jan1992,3,13,13,1,2,Winter,0,1,0,1
19920114,January 14, 1992,Wednesday,January,1992,199201,Jan1992,4,14,14,1,3,Winter,0,1,0,1
19920115,January 15, 1992,Thursday,January,1992,199201,Jan1992,5,15,15,1,3,Winter,0,1,0,1
19920116,January 16, 1992,Friday,January,1992,199201,Jan1992,6,16,16,1,3,Winter,0,1,0,1
19920117,January 17, 1992,Saturday,January,1992,199201,Jan1992,7,17,17,1,3,Winter,1,1,0,0
19920118,January 18, 1992,Sunday,January,1992,199201,Jan1992,1,18,18,1,3,Winter,0,1,0,0
```

Method 2: PIG

```
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_yearmonthnum:int, d_yearmonth: 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);
DESCRIBE dwdate;

STORE dwdate INTO '/user/ec2-user/pig_dwdate_csv/' USING PigStorage(',');
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
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_yearmonthnum:int, d_yearmonth: chararray, d_daynuminweek:int, d_daynuminmonth:int, d_daynuminyear:int, d_monthnuminyear:int, d_weeknuminyear:int, d_sellingseason: chararray, d_lastdayinweekfl: charar
ray,
>> d_lastdayinmonthfl: chararray, d_holidayfl: chararray, d_weekdayfl: chararray);
2016-11-19 19:47:30,830 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:47:30,908 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> DESCRIBE dwdate;
dwdate: {d_datekey: int,d_date: chararray,d_dayofweek: chararray,d_month: chararray,d_year: int,d_yearmonthnum: int,d_yearmonth: chararray,d_daynuminweek: int,d_daynuminmonth: int,d_daynuminyear: int,d_mo
nthnuminyear: int,d_weeknuminyear: int,d_sellingseason: chararray,d_lastdayinweekfl: chararray,d_lastdayinmonthfl: chararray,d_holidayfl: chararray,d_weekdayfl: chararray}
grunt> STORE dwdate INTO '/user/ec2-user/pig_dwdate_csv/' USING PigStorage(',');
2016-11-19 19:48:09,244 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:48:09,282 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2016-11-19 19:48:09,297 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:48:09,298 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-19 19:48:09,298 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-19 19:48:09,305 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-19 19:48:09,307 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
```

Transforming dwdate.tbl to a csv: output

```
2016-11-19 19:48:25,229 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-19 19:48:25,229 [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        2016-11-19 19:48:09     2016-11-19 19:48:25     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1479510364858_0016  1       0       3       3       3       3       0       0       0       0       dwdate  MAP_ONLY        /user/ec2-user/pig_dwdate_csv,

Input(s):
Successfully read 2556 records (227781 bytes) from: "/user/ec2-user/dwdate.tbl"

Output(s):
Successfully stored 2556 records (227409 bytes) in: "/user/ec2-user/pig_dwdate_csv"

Counters:
Total records written : 2556
Total bytes written : 227409
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1479510364858_0016


2016-11-19 19:48:25,231 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:48:25,235 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:48:25,258 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:48:25,262 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:48:25,280 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:48:25,285 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:48:25,308 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Output file -->

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/pig_dwdate_csv
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-19 19:48 /user/ec2-user/pig_dwdate_csv/_SUCCESS
-rw-r--r--   2 ec2-user supergroup     227409 2016-11-19 19:48 /user/ec2-user/pig_dwdate_csv/part-m-00000
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/pig_dwdate_csv/part-m-00000
19920101,January 1, 1992,Thursday,January,1992,199201,Jan1992,5,1,1,1,1,Winter,0,1,1,1
19920102,January 2, 1992,Friday,January,1992,199201,Jan1992,6,2,2,1,1,Winter,0,1,0,1
19920103,January 3, 1992,Saturday,January,1992,199201,Jan1992,7,3,3,1,1,Winter,1,1,0,0
19920104,January 4, 1992,Sunday,January,1992,199201,Jan1992,1,4,4,1,1,Winter,0,1,0,0
19920105,January 5, 1992,Monday,January,1992,199201,Jan1992,2,5,5,1,1,Winter,0,1,0,1
19920106,January 6, 1992,Tuesday,January,1992,199201,Jan1992,3,6,6,1,1,Winter,0,1,0,1
19920107,January 7, 1992,Wednesday,January,1992,199201,Jan1992,4,7,7,1,2,Winter,0,1,0,1
19920108,January 8, 1992,Thursday,January,1992,199201,Jan1992,5,8,8,1,2,Winter,0,1,0,1
19920109,January 9, 1992,Friday,January,1992,199201,Jan1992,6,9,9,1,2,Winter,0,1,0,1
19920110,January 10, 1992,Saturday,January,1992,199201,Jan1992,7,10,10,1,2,Winter,1,1,0,0
19920111,January 11, 1992,Sunday,January,1992,199201,Jan1992,1,11,11,1,2,Winter,0,1,0,0
19920112,January 12, 1992,Monday,January,1992,199201,Jan1992,2,12,12,1,2,Winter,0,1,0,1
19920113,January 13, 1992,Tuesday,January,1992,199201,Jan1992,3,13,13,1,2,Winter,0,1,0,1
19920114,January 14, 1992,Wednesday,January,1992,199201,Jan1992,4,14,14,1,3,Winter,0,1,0,1
19920115,January 15, 1992,Thursday,January,1992,199201,Jan1992,5,15,15,1,3,Winter,0,1,0,1
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

C. Extract all of the numeric columns from lineorder.tbl into a space-separated file

Method 1: 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;

LOAD DATA LOCAL INPATH '/home/ec2-user/lineorder.tbl'
OVERWRITE INTO TABLE lineorder;

INSERT OVERWRITE DIRECTORY 'lineorder_numeric'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ' '
SELECT lo_orderkey, lo_linenumber, lo_custkey, lo_partkey, lo_suppkey, lo_orderdate, lo_quantity, lo_extendedprice,
lo_ordertotalprice, lo_discount, lo_revenue, lo_supplycost, lo_tax, lo_commitdate FROM lineorder;
```

Time taken: 44.734 seconds

```
hive> INSERT OVERWRITE DIRECTORY 'lineorder_numeric'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ' '
    > SELECT lo_orderkey, lo_linenumber, lo_custkey, lo_partkey, lo_suppkey, lo_orderdate, lo_quantity, lo_extendedprice, lo_ordertotalprice, lo_discount, lo_revenue, lo_supplycost, lo_tax, lo_commitdate
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_20161119175317_9f6b3828-d299-4152-abbb-b3cb3f743021
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_1479510364858_0012, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0012/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0012
Hadoop job information for Stage-1: number of mappers: 10; number of reducers: 0
2016-11-19 17:53:21,940 Stage-1 map = 0%,  reduce = 0%
2016-11-19 17:53:37,329 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 24.7 sec
2016-11-19 17:53:42,728 Stage-1 map = 15%,  reduce = 0%, Cumulative CPU 55.76 sec
2016-11-19 17:53:45,908 Stage-1 map = 30%,  reduce = 0%, Cumulative CPU 68.56 sec
2016-11-19 17:53:46,976 Stage-1 map = 35%,  reduce = 0%, Cumulative CPU 70.28 sec
2016-11-19 17:53:48,047 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 75.8 sec
2016-11-19 17:53:49,116 Stage-1 map = 55%,  reduce = 0%, Cumulative CPU 81.63 sec
2016-11-19 17:53:51,227 Stage-1 map = 70%,  reduce = 0%, Cumulative CPU 91.99 sec
2016-11-19 17:53:52,318 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 95.6 sec
2016-11-19 17:53:55,513 Stage-1 map = 80%,  reduce = 0%, Cumulative CPU 101.96 sec
2016-11-19 17:53:57,596 Stage-1 map = 85%,  reduce = 0%, Cumulative CPU 107.03 sec
2016-11-19 17:53:58,626 Stage-1 map = 90%,  reduce = 0%, Cumulative CPU 108.0 sec
2016-11-19 17:53:59,657 Stage-1 map = 95%,  reduce = 0%, Cumulative CPU 110.7 sec
2016-11-19 17:54:00,691 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 113.16 sec
MapReduce Total cumulative CPU time: 1 minutes 53 seconds 160 msec
Ended Job = job_1479510364858_0012
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://172.31.46.250/user/ec2-user/lineorder_numeric/.hive-staging_hive_2016-11-19_17-53-17_137_8676491540899457203-1/-ext-10000
Moving data to: lineorder_numeric
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10   Cumulative CPU: 113.16 sec   HDFS Read: 2417896940 HDFS Write: 2036536712 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 53 seconds 160 msec
OK
Time taken: 44.734 seconds
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Output files:

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/lineorder_numeric
Found 10 items
-rwxr-xr-x   2 ec2-user supergroup  225887276 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000000_0
-rwxr-xr-x   2 ec2-user supergroup  225718254 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000001_0
-rwxr-xr-x   2 ec2-user supergroup  225889752 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000002_0
-rwxr-xr-x   2 ec2-user supergroup  226073943 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000003_0
-rwxr-xr-x   2 ec2-user supergroup  226206648 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000004_0
-rwxr-xr-x   2 ec2-user supergroup  226307887 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000005_0
-rwxr-xr-x   2 ec2-user supergroup  226301651 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000006_0
-rwxr-xr-x   2 ec2-user supergroup  226301532 2016-11-19 17:54 /user/ec2-user/lineorder_numeric/000007_0
-rwxr-xr-x   2 ec2-user supergroup  226301456 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000008_0
-rwxr-xr-x   2 ec2-user supergroup    1548313 2016-11-19 17:53 /user/ec2-user/lineorder_numeric/000009_0
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/lineorder_numeric/000009_0
23982182 4 80857 335011 10879 19961209 23 2405800 17586382 3 2333626 62760 7 19970118
23982182 5 80857 129113 29491 19961209 6 685266 17586382 4 657855 68526 4 19970120
23982183 1 75700 216120 1197 19951202 26 2693886 8475336 2 2640008 62166 5 19960115
23982183 2 75700 315456 30492 19951202 19 2795736 8475336 10 2516162 88286 3 19960213
23982183 3 75700 259318 22241 19951202 21 2682330 8475336 4 2575036 76638 6 19960219
23982183 4 75700 348367 309 19951202 3 424605 8475336 10 382144 84921 0 19960120
23982208 1 43952 354065 39505 19930126 1 111905 16835761 4 107428 67143 1 19930307
23982208 2 43952 365955 26654 19930126 27 5456538 16835761 2 5347407 121256 5 19930225
23982208 3 43952 193085 6494 19930126 33 3887664 16835761 10 3498897 70684 2 19930310
23982208 4 43952 366973 15796 19930126 36 7343856 16835761 4 7050101 122397 7 19930414
23982209 1 100858 312175 3873 19970520 33 3917628 10163071 1 3878451 71229 7 19970727
23982209 2 100858 320906 10194 19970520 32 6166048 10163071 8 5672764 115613 6 19970806
23982210 1 107642 384172 14532 19980331 45 5652720 41831639 8 5200502 75369 7 19980530
23982210 2 107642 372703 22884 19980331 37 6570053 41831639 0 6570053 106541 7 19980609
```

Method 2: PIG

```
lineorder_numeric = FOREACH lineorder GENERATE lo_orderkey, lo_linenumber, lo_custkey, lo_partkey,
lo_suppkey, lo_orderdate, lo_quantity, lo_extendedprice, lo_ordertotalprice, lo_discount, lo_revenue,
lo_supplycost, lo_tax, lo_comitdate;
STORE lineorder_numeric INTO  '/user/ec2-user/pig_lineorder_numeric/' USING PigStorage(' ');
```

Transforming numeric columns of lineorder.tbl to a space separated file:

```
grunt> lineorder_numeric = FOREACH lineorder GENERATE lo_orderkey, lo_linenumber, lo_custkey, lo_partkey, lo_suppkey, lo_orderdate, lo_quantity, lo_extendedprice, lo_ordertotalprice, lo_discount, lo_reven
ue, lo_supplycost, lo_tax, lo_comitdate;
grunt> STORE lineorder_numeric INTO  '/user/ec2-user/pig_lineorder_numeric/' USING PigStorage(' ');
2016-11-19 19:51:43,639 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:51:43,689 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2016-11-19 19:51:43,705 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:51:43,706 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-19 19:51:43,706 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-19 19:51:43,713 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for lineorder: $6, $7, $16
2016-11-19 19:51:43,720 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-19 19:51:43,722 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2016-11-19 19:51:43,722 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2016-11-19 19:51:43,730 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-19 19:51:43,731 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:51:43,733 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-11-19 19:51:43,734 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-11-19 19:51:43,734 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-11-19 19:51:43,796 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/pig-0.15.0-core-h2.jar to DistributedCache t
hrough /tmp/temp-1358699710/tmp1000846687/pig-0.15.0-core-h2.jar
2016-11-19 19:51:43,811 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/automaton-1.11-8.jar to DistributedCache
 through /tmp/temp-1358699710/tmp1318347148/automaton-1.11-8.jar
2016-11-19 19:51:43,830 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/antlr-runtime-3.4.jar to DistributedCach
e through /tmp/temp-1358699710/tmp-1669271574/antlr-runtime-3.4.jar
2016-11-19 19:51:43,864 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/joda-time-2.5.jar to DistributedCache th
rough /tmp/temp-1358699710/tmp961288013/joda-time-2.5.jar
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
2016-11-19 19:53:04,959 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-19 19:53:04,959 [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        2016-11-19 19:51:43     2016-11-19 19:53:04     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1479510364858_0017  18      0       69      29      53      54      0       0       0       0       lineorder,lineorder_numeric     MAP_ONLY        /user/ec2-user/pig_lineorder_numeric,

Input(s):
Successfully read 23996604 records (2417832945 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 23996604 records (2036536712 bytes) in: "/user/ec2-user/pig_lineorder_numeric"

Counters:
Total records written : 23996604
Total bytes written : 2036536712
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1479510364858_0017


2016-11-19 19:53:04,960 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:53:04,963 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:53:05,027 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:53:05,030 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:53:05,049 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-19 19:53:05,052 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-19 19:53:05,074 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Output files:

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/pig_lineorder_numeric/
Found 19 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/_SUCCESS
-rw-r--r--   2 ec2-user supergroup  114697553 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00000
-rw-r--r--   2 ec2-user supergroup  112774578 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00001
-rw-r--r--   2 ec2-user supergroup  112943676 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00002
-rw-r--r--   2 ec2-user supergroup  112940914 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00003
-rw-r--r--   2 ec2-user supergroup  112948247 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00004
-rw-r--r--   2 ec2-user supergroup  112946362 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00005
-rw-r--r--   2 ec2-user supergroup  112941505 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00006
-rw-r--r--   2 ec2-user supergroup  112930373 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00007
-rw-r--r--   2 ec2-user supergroup  113062054 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00008
-rw-r--r--   2 ec2-user supergroup  113144594 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00009
-rw-r--r--   2 ec2-user supergroup  113154124 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00010
-rw-r--r--   2 ec2-user supergroup  113143570 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00011
-rw-r--r--   2 ec2-user supergroup  113153763 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00012
-rw-r--r--   2 ec2-user supergroup  113150604 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00013
-rw-r--r--   2 ec2-user supergroup  113153694 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00014
-rw-r--r--   2 ec2-user supergroup  113151047 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00015
-rw-r--r--   2 ec2-user supergroup  113147838 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00016
-rw-r--r--   2 ec2-user supergroup  113152216 2016-11-19 19:52 /user/ec2-user/pig_lineorder_numeric/part-m-00017
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/pig_lineorder_numeric/part-m-00017/
^C[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/pig_lineorder_numeric/part-m-00017/
21331457 4 72980 367634 10797 19921022 30 5104860 22118985 4 4900665 102097 0 19921201
21331457 5 72980 18723 8984 19921022 20 3283440 22118985 3 3184936 98503 0 19921202
21331457 6 72980 263926 27486 19921022 22 4157802 22118985 4 3991489 113394 1 19921201
21331458 1 43810 80161 13895 19950112 35 3994060 10162421 8 3674535 68469 8 19950321
21331458 2 43810 62883 2435 19950112 12 2215056 10162421 5 2104303 110752 7 19950304
21331458 3 43810 63316 3513 19950112 30 3837930 10162421 4 3684412 76758 7 19950402
21331459 1 90397 252951 23773 19920806 3 571182 8550485 8 525487 114236 2 19921029
21331459 2 90397 134541 20165 19920806 22 3466188 8550485 7 3223554 94532 2 19920924
21331459 3 90397 20847 33071 19920806 22 3889248 8550485 0 3889248 106070 0 19920909
21331459 4 90397 370220 23139 19920806 7 903147 8550485 10 812832 77412 3 19920926
21331460 1 73843 240272 27875 19970202 1 121226 16426060 3 117589 72735 2 19970501
21331460 2 73843 247879 39419 19970202 41 7490126 16426060 1 7415224 109611 2 19970502
21331460 3 73843 95218 22672 19970202 38 4610198 16426060 2 4517994 72792 8 19970410
21331460 4 73843 319304 460 19970202 2 264658 16426060 2 259364 79397 3 19970502
21331460 5 73843 133566 1743 19970202 19 3039164 16426060 10 2735247 95973 6 19970311
21331460 6 73843 34136 13115 19970202 7 749091 16426060 7 696654 64207 0 19970310
21331461 1 31052 259398 15154 19930428 16 2171808 2083849 5 2063217 81442 1 19930711
```

D. Create a pre-join (i.e. a new data file) that corresponds to (all the columns needed for Q2.1 that you ran before). What is the size of the new file? (lo_partkey and lo_suppkey will let you join it with other 2 tables from Q2.1
SELECT lo_partkey, lo_suppkey, d_year, lo_revenue
FROM lineorder, dwdate
WHERE lo_orderdate = d_datekey;

Method 1: HIVE

```
CREATE TABLE prejoin1 AS
SELECT lo_partkey, lo_suppkey, d_year, lo_revenue
FROM lineorder, dwdate WHERE lo_orderdate = d_datekey;
```

Time taken: 53.892 seconds

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
hive> CREATE TABLE prejoin1 AS
    > SELECT lo_partkey, lo_suppkey, d_year, lo_revenue
    > FROM lineorder, dwdate WHERE lo_orderdate = d_datekey;
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_20161119205853_a1173098-2c9e-4e15-8409-54827f9984ad
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_20161119205853_a1173098-2c9e-4e15-8409-54827f9984ad.log
2016-11-19 20:59:00     Starting to launch local task to process map join;      maximum memory = 477102080
2016-11-19 20:59:01     Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/a8f03bb2-cfa9-49bf-96b5-c737812b93a4/hive_2016-11-19_20-58-53_878_5537446486159651032-1/-local-
10004/HashTable-Stage-4/MapJoin-mapfile01--.hashtable
2016-11-19 20:59:01     Uploaded 1 File to: file:/tmp/ec2-user/a8f03bb2-cfa9-49bf-96b5-c737812b93a4/hive_2016-11-19_20-58-53_878_5537446486159651032-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile01--.h
ashtable (67039 bytes)
2016-11-19 20:59:01     End of local task; Time Taken: 1.223 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_1479510364858_0020, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0020/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0020
Hadoop job information for Stage-4: number of mappers: 10; number of reducers: 0
2016-11-19 20:59:07,809 Stage-4 map = 0%,  reduce = 0%
2016-11-19 20:59:21,392 Stage-4 map = 5%,  reduce = 0%, Cumulative CPU 7.16 sec
2016-11-19 20:59:23,601 Stage-4 map = 10%,  reduce = 0%, Cumulative CPU 18.49 sec
2016-11-19 20:59:30,390 Stage-4 map = 20%,  reduce = 0%, Cumulative CPU 46.92 sec
2016-11-19 20:59:33,670 Stage-4 map = 25%,  reduce = 0%, Cumulative CPU 56.34 sec
2016-11-19 20:59:35,839 Stage-4 map = 35%,  reduce = 0%, Cumulative CPU 63.73 sec
2016-11-19 20:59:36,887 Stage-4 map = 40%,  reduce = 0%, Cumulative CPU 66.27 sec
2016-11-19 20:59:37,930 Stage-4 map = 45%,  reduce = 0%, Cumulative CPU 67.44 sec
2016-11-19 20:59:39,018 Stage-4 map = 60%,  reduce = 0%, Cumulative CPU 73.55 sec
2016-11-19 20:59:44,398 Stage-4 map = 70%,  reduce = 0%, Cumulative CPU 87.48 sec
2016-11-19 20:59:45,453 Stage-4 map = 85%,  reduce = 0%, Cumulative CPU 90.88 sec
2016-11-19 20:59:46,483 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 92.9 sec
MapReduce Total cumulative CPU time: 1 minutes 32 seconds 900 msec
Ended Job = job_1479510364858_0020
Moving data to: hdfs://172.31.46.250/user/hive/warehouse/prejoin1
MapReduce Jobs Launched:
Stage-Stage-4: Map: 10   Cumulative CPU: 92.9 sec   HDFS Read: 2417935110 HDFS Write: 607282429 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 32 seconds 900 msec
OK
Time taken: 53.892 seconds
hive>
```

Output files:

```
[ec2-user@ip-172-31-46-250 apache-hive-2.0.1-bin]$ hadoop fs -ls /user/hive/warehouse/prejoin1
Found 10 items
-rwxrwxr-x   2 ec2-user supergroup   68069134 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000000_0
-rwxrwxr-x   2 ec2-user supergroup   67446798 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000001_0
-rwxrwxr-x   2 ec2-user supergroup   67786769 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000002_0
-rwxrwxr-x   2 ec2-user supergroup   67786023 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000003_0
-rwxrwxr-x   2 ec2-user supergroup   67115774 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000004_0
-rwxrwxr-x   2 ec2-user supergroup   67269861 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000005_0
-rwxrwxr-x   2 ec2-user supergroup   67117101 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000006_0
-rwxrwxr-x   2 ec2-user supergroup   67114805 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000007_0
-rwxrwxr-x   2 ec2-user supergroup   67116224 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000008_0
-rwxrwxr-x   2 ec2-user supergroup     459114 2016-11-19 20:59 /user/hive/warehouse/prejoin1/000009_0
[ec2-user@ip-172-31-46-250 apache-hive-2.0.1-bin]$ hadoop fs -tail /user/hive/warehouse/prejoin1/000009_0
7203660
1758991582019962063760
3194263465719962523685
2944604975199714399905
3042902973199514094 70
1926504127199516729 4
1786412696919935853654
1628124414199325861 24
2193239061199348925 21
2139291821919959214 55
2880213491719953450814
3680193463919957913 36
7966930781995615115 0
1907833628919958102224
8244818574199567287 89
3838803219119951571096
2665771883819955478094
3562293989219957171 47
```

Method 2: PIG

```
join_date = JOIN lineorder by lo_orderdate, dwdate by d_datekey;
prejoin1 = FOREACH join_date GENERATE lineorder::lo_partkey, lineorder::lo_suppkey, dwdate::d_year, lineorder::lo_revenue;
STORE prejoin1 INTO '/user/ec2-user/prejoin1/' USING PigStorage('|');
```

Creating table prejoin1-->

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
grunt> join_date = JOIN lineorder by lo_orderdate, dwdate by d_datekey;
2016-11-22 05:25:34,525 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> prejoin1 = FOREACH join_date GENERATE lineorder::lo_partkey, lineorder::lo_suppkey, dwdate::d_year, lineorder::lo_revenue;
grunt> STORE prejoin1 INTO '/user/ec2-user/prejoin1/' USING PigStorage('|');
2016-11-22 05:25:38,287 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 05:25:38,347 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2016-11-22 05:25:38,363 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 05:25:38,366 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-22 05:25:38,367 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-22 05:25:38,378 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for lineorder: $0, $1, $2, $6, $7, $8, $9, $10, $11, $13, $14, $15, $16
2016-11-22 05:25:38,378 [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, $16
2016-11-22 05:25:38,385 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-22 05:25:38,386 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 05:25:38,387 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2016-11-22 05:25:38,387 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2016-11-22 05:25:38,395 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 05:25:38,396 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:25:38,398 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-11-22 05:25:38,399 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-11-22 05:25:38,399 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2016-11-22 05:25:38,399 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.InputSizeReducerEstimator
2016-11-22 05:25:38,402 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=2417983972
2016-11-22 05:25:38,402 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 3
2016-11-22 05:25:38,402 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-11-22 05:25:38,455 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/pig-0.15.0-core-h2.jar to DistributedCache t
hrough /tmp/temp451466183/tmp-117814215/pig-0.15.0-core-h2.jar
2016-11-22 05:25:38,500 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/automaton-1.11-8.jar to DistributedCache
 through /tmp/temp451466183/tmp-1547393316/automaton-1.11-8.jar
2016-11-22 05:25:38,513 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/antlr-runtime-3.4.jar to DistributedCach
e through /tmp/temp451466183/tmp1136015589/antlr-runtime-3.4.jar
2016-11-22 05:25:38,540 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/lib/joda-time-2.5.jar to DistributedCache th
rough /tmp/temp451466183/tmp-552169625/joda-time-2.5.jar
```

```
2016-11-22 05:25:39,069 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2016-11-22 05:25:39,069 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:06,156 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 5% complete
2016-11-22 05:26:06,156 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:09,160 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 9% complete
2016-11-22 05:26:09,160 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:13,167 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 18% complete
2016-11-22 05:26:13,167 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:16,173 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 22% complete
2016-11-22 05:26:16,173 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:21,180 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 28% complete
2016-11-22 05:26:21,180 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:23,683 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 36% complete
2016-11-22 05:26:23,683 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:26,186 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 42% complete
2016-11-22 05:26:26,186 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:28,189 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 49% complete
2016-11-22 05:26:28,189 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:38,207 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 59% complete
2016-11-22 05:26:38,207 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:41,211 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 65% complete
2016-11-22 05:26:41,211 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:43,214 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 73% complete
2016-11-22 05:26:43,214 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:46,217 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 83% complete
2016-11-22 05:26:46,217 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:26:58,733 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 87% complete
2016-11-22 05:26:58,733 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:27:11,247 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 92% complete
2016-11-22 05:27:11,247 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:27:21,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 96% complete
2016-11-22 05:27:21,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0068]
2016-11-22 05:27:33,777 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:27:33,782 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 05:27:33,916 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:27:33,920 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 05:27:33,966 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:27:33,969 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 05:27:33,992 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-22 05:27:33,992 [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        2016-11-22 05:25:38     2016-11-22 05:27:33     HASH_JOIN

Success!
```

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1479756709677_0068  19      3       52      10      40      38      77      75      77      77      dwdate,join_date,lineorder,prejoin1     HASH_JOIN       /user/ec2-user/prejoin1,

Input(s):
Successfully read 2556 records from: "/user/ec2-user/scale4/dwdate.tbl"
Successfully read 23996604 records from: "/user/ec2-user/scale4/lineorder.tbl"

Output(s):
Successfully stored 23996604 records (607281603 bytes) in: "/user/ec2-user/prejoin1"

Counters:
Total records written : 23996604
Total bytes written : 607281603
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1479756709677_0068
```

```
2016-11-22 05:27:33,993 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:27:33,996 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 05:27:34,059 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:27:34,062 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 05:27:34,087 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:27:34,090 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 05:27:34,112 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

prejoin1 – output files:

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/prejoin1
Found 4 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-22 05:27 /user/ec2-user/prejoin1/_SUCCESS
-rw-r--r--   2 ec2-user supergroup  202652299 2016-11-22 05:27 /user/ec2-user/prejoin1/part-r-00000
-rw-r--r--   2 ec2-user supergroup  202297701 2016-11-22 05:27 /user/ec2-user/prejoin1/part-r-00001
-rw-r--r--   2 ec2-user supergroup  202331603 2016-11-22 05:27 /user/ec2-user/prejoin1/part-r-00002
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -tail /user/ec2-user/prejoin1/part-r-00002
|1998|2894737
258834|19991|1998|6325068
222003|23415|1998|1106288
398848|2966|1998|3440048
22819|15673|1998|317009
317402|9342|1998|1405196
159607|2727|1998|4649814
143544|24106|1998|1539913
130786|15132|1998|6867428
8364|38181|1998|1145124
292519|10751|1998|1015728
357435|9339|1998|590998
193662|386|1998|3335754
69655|1485|1998|4617255
137336|29019|1998|3356418
204917|9656|1998|5471165
292081|36129|1998|3187017
309871|24979|1998|2941665
146100|10413|1998|4171804
299170|9569|1998|229155
80907|29951|1998|7559151
112927|26180|1998|1117393
363195|36867|1998|226472
80393|12309|1998|4214933
```

Part 2: Data Querying

Data used for Part 2: **Scale4**

2A. Run SSBM query 3.1

Method 1: Hive

```
--Q3.1  No simpifications
select c_nation, s_nation, d_year, sum(lo_revenue) as revenue
from customer, lineorder, supplier, dwdate
where lo_custkey = c_custkey
  and lo_suppkey = s_suppkey
  and lo_orderdate = d_datekey
  and c_region = 'ASIA'
  and s_region = 'ASIA'
  and d_year between 1992 and 1997
group by c_nation, s_nation, d_year
order by d_year asc, revenue asc;
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Query 3.1 in Hive:

```
hive> select c_nation, s_nation, d_year, sum(lo_revenue) as revenue
    > from customer, lineorder, supplier, dwdate
    > where lo_custkey = c_custkey
    >   and lo_suppkey = s_suppkey
    >   and lo_orderdate = d_datekey
    >   and c_region = 'ASIA'
    >   and s_region = 'ASIA'
    >   and d_year between 1992 and 1997
    > group by c_nation, s_nation, 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_20161120015752_83402618-da8a-4cd0-8010-db34885a7cf3
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_20161120015752_83402618-da8a-4cd0-8010-db34885a7cf3.log
2016-11-20 01:57:58     Starting to launch local task to process map join;      maximum memory = 477102080
2016-11-20 01:58:00     Dump the side-table for tag: 0 with group count: 24183 into file: file:/tmp/ec2-user/b44253bf-36d6-48b0-9aa9-e69c1df99d9d/hive_2016-11-20_01-57-52_048_7601904054231283350-1/-local
-10010/HashTable-Stage-12/MapJoin-mapfile30--.hashtable
2016-11-20 01:58:00     Uploaded 1 File to: file:/tmp/ec2-user/b44253bf-36d6-48b0-9aa9-e69c1df99d9d/hive_2016-11-20_01-57-52_048_7601904054231283350-1/-local-10010/HashTable-Stage-12/MapJoin-mapfile30--.
hashtable (696855 bytes)
2016-11-20 01:58:00     End of local task; Time Taken: 1.818 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_1479510364858_0021, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0021/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0021
Hadoop job information for Stage-12: number of mappers: 10; number of reducers: 0
2016-11-20 01:58:06,968 Stage-12 map = 0%,  reduce = 0%
2016-11-20 01:58:29,576 Stage-12 map = 10%,  reduce = 0%, Cumulative CPU 34.0 sec
2016-11-20 01:58:32,830 Stage-12 map = 25%,  reduce = 0%, Cumulative CPU 50.6 sec
2016-11-20 01:58:34,915 Stage-12 map = 30%,  reduce = 0%, Cumulative CPU 55.24 sec
2016-11-20 01:58:35,960 Stage-12 map = 35%,  reduce = 0%, Cumulative CPU 61.12 sec
2016-11-20 01:58:37,012 Stage-12 map = 50%,  reduce = 0%, Cumulative CPU 63.13 sec
2016-11-20 01:58:38,055 Stage-12 map = 60%,  reduce = 0%, Cumulative CPU 66.4 sec
2016-11-20 01:58:39,130 Stage-12 map = 75%,  reduce = 0%, Cumulative CPU 68.81 sec
2016-11-20 01:58:44,426 Stage-12 map = 80%,  reduce = 0%, Cumulative CPU 75.48 sec
2016-11-20 01:58:45,486 Stage-12 map = 95%,  reduce = 0%, Cumulative CPU 77.8 sec
2016-11-20 01:58:46,528 Stage-12 map = 100%,  reduce = 0%, Cumulative CPU 79.34 sec
MapReduce Total cumulative CPU time: 1 minutes 19 seconds 340 msec
Ended Job = job_1479510364858_0021
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_20161120015752_83402618-da8a-4cd0-8010-db34885a7cf3.log
2016-11-20 01:58:52     Starting to launch local task to process map join;      maximum memory = 477102080
```

Time taken: 108.763 seconds

```
INDIA    INDONESIA        1996    22519200047
INDONESIA        INDONESIA        1996    22702436705
INDONESIA        INDIA    1996    22984340988
VIETNAM CHINA    1997    20205271121
VIETNAM VIETNAM 1997    20501834171
INDIA    CHINA    1997    20710386579
JAPAN    CHINA    1997    20716708109
INDIA    JAPAN    1997    20823378658
JAPAN    INDIA    1997    20879337247
INDIA    VIETNAM 1997    20906063484
CHINA    VIETNAM 1997    20945486746
JAPAN    JAPAN    1997    21002232981
JAPAN    INDONESIA        1997    21125190315
JAPAN    VIETNAM 1997    21147434582
VIETNAM JAPAN    1997    21211983035
INDONESIA        VIETNAM 1997    21249107957
VIETNAM INDONESIA        1997    21367241926
CHINA    CHINA    1997    21464890789
INDIA    INDONESIA        1997    21471229778
VIETNAM INDIA    1997    21728099866
INDONESIA        CHINA    1997    21758306872
CHINA    JAPAN    1997    21859870180
CHINA    INDONESIA        1997    21876152997
INDONESIA        INDIA    1997    22096226465
INDONESIA        JAPAN    1997    22236659441
INDIA    INDIA    1997    22328229227
CHINA    INDIA    1997    22342102708
INDONESIA        INDONESIA        1997    22488960888
Time taken: 108.763 seconds, Fetched: 150 row(s)
hive>
```

Method 2: PIG

```
dwdate= LOAD '/user/ec2-user/scale4/dwdate.tbl' USING PigStorage('|')
AS (d_datekey:int, d_date:chararray, d_dayofweek: chararray, d_month: chararray, d_year:int, d_yearmonthnum:int, d_yearmonth:
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);

DESCRIBE dwdate;

lineorder= LOAD '/user/ec2-user/scale4/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_comitdate:int, lo_shipmode:chararray);

DESCRIBE lineorder;

customer= LOAD '/user/ec2-user/scale4/customer.tbl' USING PigStorage('|')
AS (c_custkey:int, c_name:chararray, c_address:chararray, c_city:chararray, c_nation:chararray, c_region:chararray,
c_phone:chararray, c_mktsegment:chararray);

DESCRIBE customer;

supplier= LOAD '/user/ec2-user/scale4/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);

DESCRIBE supplier;

lo_sample = SAMPLE lineorder 0.3;
dwdate_sample = SAMPLE dwdate 0.3;
customer_sample = SAMPLE customer 0.3;
supplier_sample = SAMPLE supplier 0.3;
join1 = JOIN lo_sample by lo_custkey, customer_sample by c_custkey;
join2 = JOIN join1 by lo_suppkey, supplier_sample by s_suppkey;
join3 = JOIN join2 by lo_orderdate, dwdate_sample by d_datekey;
filtereddata = FILTER join3 BY c_region == 'ASIA' AND s_region == 'ASIA' AND d_year >= 1992 AND d_year <= 1997;
groupeddata = GROUP filtereddata BY (c_nation, s_nation, d_year);
query31 = FOREACH groupeddata GENERATE group.$0, group.$1, group.$2 as d_year, SUM(filtereddata.lo_revenue) as revenue;
ordered_result = ORDER query31 BY d_year ASC,revenue ASC;

STORE ordered_result INTO '/user/ec2-user/Part2A_pig' USING PigStorage('|');
```

Running Query 3.1 on Pig with scale4 data resulted in the following error:

ERROR 0: org.apache.pig.backend.executionengine.ExecException: ERROR 2997: Unable to recreate exception from backed error: Error: GC overhead limit exceeded
2016-11-22 05:44:13,060 [main] ERROR org.apache.pig.tools.pigstats.mapreduce.MRPigStatsUtil - 1 map reduce job(s) failed!

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
[ec2-user@ip-172-31-46-250 ~]$ $PIG_HOME/bin/pig -f Part2A_pig.pig
16/11/22 05:36:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/22 05:36:49 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/11/22 05:36:49 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-11-22 05:36:49,081 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-11-22 05:36:49,082 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig_1479793009079.log
2016-11-22 05:36:49,717 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-11-22 05:36:49,830 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-11-22 05:36:49,830 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 05:36:49,830 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.31.46.250/
2016-11-22 05:36:50,766 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
dwdate: {d_datekey: int,d_date: chararray,d_dayofweek: chararray,d_month: chararray,d_year: int,d_yearmonthnum: int,d_yearmonth: chararray,d_daynuminweek: int,d_daynuminmonth: int,d_daynuminyear: int,d_mo
nthnuminyear: int,d_weeknuminyear: int,d_sellingseason: chararray,d_lastdayinweekfl: chararray,d_lastdayinmonthfl: chararray,d_holidayfl: chararray,d_weekdayfl: chararray}
2016-11-22 05:36:50,916 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
lineorder: {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_comitdate: int,lo_shipmode: chararray}
2016-11-22 05:36:51,085 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
customer: {c_custkey: int,c_name: chararray,c_address: chararray,c_city: chararray,c_nation: chararray,c_region: chararray,c_phone: chararray,c_mktsegment: chararray}
2016-11-22 05:36:51,218 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
supplier: {s_suppkey: int,s_name: chararray,s_address: chararray,s_city: chararray,s_nation: chararray,s_region: chararray,s_phone: chararray}
2016-11-22 05:36:51,589 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,ORDER_BY,FILTER
2016-11-22 05:36:51,609 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 05:36:51,612 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-22 05:36:51,640 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOpt
imizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-22 05:36:51,801 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-22 05:36:51,839 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-249
2016-11-22 05:36:51,843 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 05:36:51,843 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 05:36:51,844 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 05:36:51,851 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 6
2016-11-22 05:36:51,851 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 6
2016-11-22 05:36:51,874 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 05:36:51,915 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
```

```
2016-11-22 05:44:07,456 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 83% complete
2016-11-22 05:44:07,457 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479756709677_0073]
2016-11-22 05:44:12,965 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Ooops! Some job has failed! Specify -stop_on_failure if you want Pig to stop immediate
ly on failure.
2016-11-22 05:44:12,965 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - job job_1479756709677_0073 has failed! Stop running all dependent jobs
2016-11-22 05:44:12,965 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-22 05:44:13,013 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 05:44:13,019 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=FAILED. Redirecting to job history server
2016-11-22 05:44:13,060 [main] ERROR org.apache.pig.tools.pigstats.PigStats - ERROR 0: org.apache.pig.backend.executionengine.ExecException: ERROR 2997: Unable to recreate exception from backed error: Err
or: GC overhead limit exceeded
2016-11-22 05:44:13,060 [main] ERROR org.apache.pig.tools.pigstats.mapreduce.MRPigStatsUtil - 1 map reduce job(s) failed!
2016-11-22 05:44:13,062 [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        2016-11-22 05:36:52     2016-11-22 05:44:13     HASH_JOIN,GROUP_BY,ORDER_BY,FILTER

Some jobs have failed! Stop running all dependent jobs

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTimMedianReducetime Alias   Feature Outputs
job_1479756709677_0069  19      3       97      8       79      91      160     80      121     122     customer,filtereddata,join1,lineorder        HASH_JOIN
job_1479756709677_0070  8       1       54      5       40      52      75      75      75      75      join2,supplier  HASH_JOIN
job_1479756709677_0071  3       1       14      5       10      13      20      20      20      20      dwdate,join3    HASH_JOIN
job_1479756709677_0072  3       1       18      10      15      18      17      17      17      17      groupeddata,query31     GROUP_BY

Failed Jobs:
JobId   Alias   Feature Message Outputs
job_1479756709677_0073  orderby_result  SAMPLER Message: Job failed!

Input(s):
Successfully read 23996604 records from: "/user/ec2-user/scale4/lineorder.tbl"
Successfully read 120000 records from: "/user/ec2-user/scale4/customer.tbl"
Successfully read 40000 records from: "/user/ec2-user/scale4/supplier.tbl"
Successfully read 2556 records from: "/user/ec2-user/scale4/dwdate.tbl"

Output(s):
```

Output with sampled data:
Pig script for Query 2A: Time taken: 3 minutes, 29 seconds and 661 ms.

```
[ec2-user@ip-172-31-46-250 ~]$ $PIG_HOME/bin/pig -f Part2A_pig.pig
16/11/22 22:32:32 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/22 22:32:32 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/11/22 22:32:32 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-11-22 22:32:32,523 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-11-22 22:32:32,524 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig_1479853952522.log
2016-11-22 22:32:33,145 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-11-22 22:32:33,259 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-11-22 22:32:33,259 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 22:32:33,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.31.46.250/
2016-11-22 22:32:34,145 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
dwdate: {d_datekey: int,d_date: chararray,d_dayofweek: chararray,d_month: chararray,d_year: int,d_yearmonthnum: int,d_yearmonth: 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}
2016-11-22 22:32:34,371 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
lineorder: {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_extendedpri
ce: int,lo_ordertotalprice: int,lo_discount: int,lo_revenue: int,lo_supplycost: int,lo_tax: int,lo_comitdate: int,lo_shipmode: chararray}
2016-11-22 22:32:34,515 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
customer: {c_custkey: int,c_name: chararray,c_address: chararray,c_city: chararray,c_nation: chararray,c_region: chararray,c_phone: chararray,c_mktsegment: chararray}
2016-11-22 22:32:34,614 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
supplier: {s_suppkey: int,s_name: chararray,s_address: chararray,s_city: chararray,s_nation: chararray,s_region: chararray,s_phone: chararray}
2016-11-22 22:32:34,844 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 22:32:34,875 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2016-11-22 22:32:34,936 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,ORDER_BY,FILTER
2016-11-22 22:32:34,960 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 22:32:34,963 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-22 22:32:34,989 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, Limi
tOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-22 22:32:35,161 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-22 22:32:35,210 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2016-11-22 22:32:35,233 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-265
2016-11-22 22:32:35,237 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 22:32:35,237 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 22:32:35,237 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 22:32:35,244 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 6
2016-11-22 22:32:35,244 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 6
2016-11-22 22:32:35,265 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 22:32:35,294 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
2016-11-22 22:36:01,306 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,354 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,358 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,382 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,386 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,408 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-22 22:36:01,436 [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        2016-11-22 22:32:35     2016-11-22 22:36:01     HASH_JOIN,GROUP_BY,ORDER_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1479756709677_0123  19      3       67      8       57      65      68      51      59      59      customer,customer_sample,join1,lineorder,lo_sample      HASH_JOIN
job_1479756709677_0124  2       1       10      5       8       8       7       7       7       7       join2,supplier,supplier_sample  HASH_JOIN
job_1479756709677_0125  2       1       5       5       5       5       3       3       3       3       dwdate,dwdate_sample,join3      HASH_JOIN
job_1479756709677_0126  1       1       3       3       3       3       3       3       3       3       groupeddata,query31     GROUP_BY,COMBINER
job_1479756709677_0127  1       1       2       2       2       2       2       2       2       2       ordered_result  SAMPLER
job_1479756709677_0128  1       1       3       3       3       3       3       3       3       3       ordered_result  ORDER_BY        /user/ec2-user/Part2A_pig,

Input(s):
Successfully read 120000 records from: "/user/ec2-user/scale4/customer.tbl"
Successfully read 23996604 records from: "/user/ec2-user/scale4/lineorder.tbl"
Successfully read 40000 records from: "/user/ec2-user/scale4/supplier.tbl"
Successfully read 2556 records from: "/user/ec2-user/scale4/dwdate.tbl"

Output(s):
Successfully stored 150 records (4409 bytes) in: "/user/ec2-user/Part2A_pig"

Counters:
Total records written : 150
Total bytes written : 4409
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
2016-11-22 22:36:01,437 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,446 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,537 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,542 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,572 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,575 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,600 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,603 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,643 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,646 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,669 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,673 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,697 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,701 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,735 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,739 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,762 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,765 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,788 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,791 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,822 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,826 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,862 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,867 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,894 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,902 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,937 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,940 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,961 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,965 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:01,988 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:01,991 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:02,021 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:02,024 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:02,047 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 22:36:02,051 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 22:36:02,071 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-11-22 22:36:02,088 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 29 seconds and 661 milliseconds (209661 ms)
```

Query 3.1 --> Output file

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/Part2A_pig
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-22 22:35 /user/ec2-user/Part2A_pig/_SUCCESS
-rw-r--r--   2 ec2-user supergroup       4409 2016-11-22 22:35 /user/ec2-user/Part2A_pig/part-r-00000
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/Part2A_pig/part-r-00000
VIETNAM|CHINA|1992|105608246
JAPAN|VIETNAM|1992|120881861
INDONESIA|VIETNAM|1992|133048231
VIETNAM|INDONESIA|1992|133761139
VIETNAM|JAPAN|1992|143266545
INDONESIA|CHINA|1992|143828638
VIETNAM|INDIA|1992|145105012
CHINA|INDIA|1992|159681580
CHINA|JAPAN|1992|161113349
JAPAN|INDIA|1992|176147354
INDONESIA|INDONESIA|1992|178260026
JAPAN|JAPAN|1992|180198398
INDIA|CHINA|1992|182939792
INDIA|JAPAN|1992|186912348
INDONESIA|INDIA|1992|187061043
CHINA|CHINA|1992|187555080
CHINA|VIETNAM|1992|194334094
VIETNAM|VIETNAM|1992|195104383
INDIA|INDONESIA|1992|211734353
INDONESIA|JAPAN|1992|211781338
JAPAN|CHINA|1992|219703475
JAPAN|INDONESIA|1992|223218819
INDIA|INDIA|1992|224868170
INDIA|VIETNAM|1992|228850814
CHINA|INDONESIA|1992|230434148
INDIA|VIETNAM|1993|104970329
INDIA|INDIA|1993|142674174
CHINA|VIETNAM|1993|144523632
INDONESIA|INDIA|1993|158360915
VIETNAM|INDONESIA|1993|161588198
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

2B. Run SSBM Query 2.1 using what you have created in 1-D (instead of lineorder and dwdate tables). You would need to rewrite the query accordingly (change the FROM clause to use your new table instead of lineorder and dwdate).

Method 1: Hive

```
--Q2.1
select sum(lo_revenue), d_year, p_brand1
from prejoin1, part, supplier
where 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;
```

Query 2.1 with prejoin1 in Hive:

```
hive> select sum(lo_revenue), d_year, p_brand1
    > from prejoin1, part, supplier
    > where 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, spar
k) or using Hive 1.X releases.
Query ID = ec2-user_20161120021221_5d517a81-c494-4173-b0d0-6e6fe1320057
Total jobs = 3
Stage-1 is selected by condition resolver.
Launching Job 1 out of 3
Number of reduce tasks not specified. Estimated from input data size: 3
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_1479510364858_0025, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0025/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0025
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 3
2016-11-20 02:12:27,722 Stage-1 map = 0%,  reduce = 0%
2016-11-20 02:12:34,994 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 4.17 sec
2016-11-20 02:12:45,450 Stage-1 map = 29%,  reduce = 3%, Cumulative CPU 32.37 sec
2016-11-20 02:12:47,530 Stage-1 map = 29%,  reduce = 6%, Cumulative CPU 34.43 sec
2016-11-20 02:12:48,562 Stage-1 map = 29%,  reduce = 8%, Cumulative CPU 40.66 sec
2016-11-20 02:12:51,664 Stage-1 map = 33%,  reduce = 8%, Cumulative CPU 47.37 sec
2016-11-20 02:12:52,705 Stage-1 map = 50%,  reduce = 8%, Cumulative CPU 49.45 sec
2016-11-20 02:12:54,820 Stage-1 map = 58%,  reduce = 8%, Cumulative CPU 56.89 sec
2016-11-20 02:12:57,916 Stage-1 map = 63%,  reduce = 8%, Cumulative CPU 64.02 sec
2016-11-20 02:13:00,005 Stage-1 map = 63%,  reduce = 11%, Cumulative CPU 65.56 sec
2016-11-20 02:13:01,064 Stage-1 map = 63%,  reduce = 17%, Cumulative CPU 68.75 sec
2016-11-20 02:13:03,138 Stage-1 map = 67%,  reduce = 17%, Cumulative CPU 74.75 sec
2016-11-20 02:13:06,260 Stage-1 map = 71%,  reduce = 17%, Cumulative CPU 79.84 sec
2016-11-20 02:13:09,392 Stage-1 map = 75%,  reduce = 17%, Cumulative CPU 86.11 sec
2016-11-20 02:13:15,584 Stage-1 map = 77%,  reduce = 17%, Cumulative CPU 95.8 sec
2016-11-20 02:13:18,690 Stage-1 map = 83%,  reduce = 17%, Cumulative CPU 101.16 sec
2016-11-20 02:13:21,781 Stage-1 map = 86%,  reduce = 17%, Cumulative CPU 106.36 sec
2016-11-20 02:13:23,847 Stage-1 map = 88%,  reduce = 17%, Cumulative CPU 108.75 sec
2016-11-20 02:13:24,891 Stage-1 map = 88%,  reduce = 19%, Cumulative CPU 110.83 sec
2016-11-20 02:13:26,954 Stage-1 map = 88%,  reduce = 22%, Cumulative CPU 111.09 sec
2016-11-20 02:13:27,989 Stage-1 map = 88%,  reduce = 25%, Cumulative CPU 114.48 sec
2016-11-20 02:13:31,077 Stage-1 map = 92%,  reduce = 25%, Cumulative CPU 118.34 sec
2016-11-20 02:13:36,229 Stage-1 map = 94%,  reduce = 25%, Cumulative CPU 123.85 sec
2016-11-20 02:13:39,317 Stage-1 map = 96%,  reduce = 25%, Cumulative CPU 126.37 sec
2016-11-20 02:13:42,401 Stage-1 map = 98%,  reduce = 25%, Cumulative CPU 129.35 sec
2016-11-20 02:13:44,461 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 131.51 sec
2016-11-20 02:13:46,525 Stage-1 map = 100%,  reduce = 45%, Cumulative CPU 134.03 sec
2016-11-20 02:13:47,581 Stage-1 map = 100%,  reduce = 59%, Cumulative CPU 136.06 sec
2016-11-20 02:13:49,660 Stage-1 map = 100%,  reduce = 71%, Cumulative CPU 143.96 sec
2016-11-20 02:13:50,693 Stage-1 map = 100%,  reduce = 74%, Cumulative CPU 147.91 sec
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1479510364858_0027, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0027/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0027
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2016-11-20 02:14:33,476 Stage-4 map = 0%,  reduce = 0%
2016-11-20 02:14:38,656 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 0.84 sec
2016-11-20 02:14:44,906 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 2.04 sec
MapReduce Total cumulative CPU time: 2 seconds 40 msec
Ended Job = job_1479510364858_0027
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4  Reduce: 3   Cumulative CPU: 171.13 sec   HDFS Read: 658367856 HDFS Write: 35389772 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 6.27 sec   HDFS Read: 35402456 HDFS Write: 9913 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 2.04 sec   HDFS Read: 15713 HDFS Write: 7217 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 59 seconds 440 msec
OK
2609439709      1992    MFGR#121
2581493128      1992    MFGR#1210
2608759988      1992    MFGR#1211
2683200706      1992    MFGR#1212
2589059999      1992    MFGR#1213
2593490528      1992    MFGR#1214
2711027508      1992    MFGR#1215
2627907374      1992    MFGR#1216
2989130307      1992    MFGR#1217
2438600810      1992    MFGR#1218
2422823149      1992    MFGR#1219
2848135064      1992    MFGR#122
2310959466      1992    MFGR#1220
2751773999      1992    MFGR#1221
2564382720      1992    MFGR#1222
2460467242      1992    MFGR#1223
2920530585      1992    MFGR#1224
2374827655      1992    MFGR#1225
2741597560      1992    MFGR#1226
2504750491      1992    MFGR#1227
2543768920      1992    MFGR#1228
2423149055      1992    MFGR#1229
2948244703      1992    MFGR#123
2370313985      1992    MFGR#1230
2403734535      1992    MFGR#1231
2975640877      1992    MFGR#1232
2584833966      1992    MFGR#1233
2621327620      1992    MFGR#1234
2891776202      1992    MFGR#1235
2719789027      1992    MFGR#1236
2967771857      1992    MFGR#1237
2213940956      1992    MFGR#1238
2671115151      1992    MFGR#1239
2811616293      1992    MFGR#124
2665412497      1992    MFGR#1240
2297811827      1992    MFGR#125
2438946265      1992    MFGR#126
2725411158      1992    MFGR#127
```

TIME TAKEN: 144.214 seconds

```
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: 144.214 seconds, Fetched: 280 row(s)
hive>
```

Method 2: PIG

```
prejoin1 = LOAD '/user/ec2-user/prejoin1' USING PigStorage('|')
AS (lo_partkey:int, lo_suppkey:int, d_year:int, lo_revenue:int);

DESCRIBE prejoin1;

part = LOAD '/user/ec2-user/scale4/part.tbl' USING PigStorage('|')
AS (p_partkey:int, p_name:chararray, p_mfgr: chararray, p_category: chararray, p_brand1: chararray, p_color: chararray, p_type:
chararray, p_size: int, p_container: chararray);

DESCRIBE part;
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
supplier= LOAD '/user/ec2-user/scale4/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);

DESCRIBE supplier;

join1 = JOIN prejoin1 by lo_partkey, part by p_partkey;
join2 = JOIN join1 by lo_suppkey, supplier by s_suppkey;
filtereddata = FILTER join2 BY p_category == 'MFGR#12' AND s_region == 'AMERICA';
groupeddata = GROUP filtereddata BY (d_year, p_brand1);
query21 = FOREACH groupeddata GENERATE SUM(filtereddata.lo_revenue) as revenue, group.$0 , group.$1;
ordered_result = ORDER query21 BY $0,$1;
STORE ordered_result INTO '/user/ec2-user/Part2B_pig' USING PigStorage('|');
```

Pig script for query 2.1 with prejoin1 --> Time taken: 3 minutes, 48 seconds and 376 ms.

```
[ec2-user@ip-172-31-46-250 ~]$ $PIG_HOME/bin/pig -f Part2B_pig.pig
16/11/22 23:37:39 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/22 23:37:39 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/11/22 23:37:39 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-11-22 23:37:39,367 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-11-22 23:37:39,367 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig_1479857859364.log
2016-11-22 23:37:40,011 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-11-22 23:37:40,120 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-11-22 23:37:40,120 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:40,120 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.31.46.250/
2016-11-22 23:37:41,016 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
prejoin1: {lo_partkey: int,lo_suppkey: int,d_year: int,lo_revenue: int}
2016-11-22 23:37:41,130 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
part: {p_partkey: int,p_name: chararray,p_mfgr: chararray,p_category: chararray,p_brand1: chararray,p_color: chararray,p_type: chararray,p_size: int,p_container: chararray}
2016-11-22 23:37:41,323 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
supplier: {s_suppkey: int,s_name: chararray,s_address: chararray,s_city: chararray,s_nation: chararray,s_region: chararray,s_phone: chararray}
2016-11-22 23:37:41,564 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:41,588 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2016-11-22 23:37:41,639 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,ORDER_BY,FILTER
2016-11-22 23:37:41,660 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:41,662 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-22 23:37:41,690 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitO
ptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-22 23:37:41,800 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-22 23:37:41,832 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2016-11-22 23:37:41,846 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-131
2016-11-22 23:37:41,850 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 23:37:41,850 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 23:37:41,857 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 5
2016-11-22 23:37:41,857 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 5
2016-11-22 23:37:41,879 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:41,909 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032


2016-11-22 23:41:27,036 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,091 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,095 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,118 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,122 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,144 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-22 23:41:27,188 [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        2016-11-22 23:37:42     2016-11-22 23:41:27     HASH_JOIN,GROUP_BY,ORDER_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1479756709677_0136  7       1       83      5       61      68      110     110     110     110     filtereddata,join1,part,prejoin1        HASH_JOIN
job_1479756709677_0137  2       1       13      5       9       9       9       9       9       9       join2,supplier  HASH_JOIN
job_1479756709677_0138  1       1       5       5       5       5       3       3       3       3       groupeddata,query21     GROUP_BY,COMBINER
job_1479756709677_0139  1       1       2       2       2       2       3       3       3       3       ordered_result  SAMPLER
job_1479756709677_0140  1       1       3       3       3       3       3       3       3       3       ordered_result  ORDER_BY        /user/ec2-user/Part2B_pig,

Input(s):
Successfully read 23996604 records from: "/user/ec2-user/prejoin1"
Successfully read 600000 records from: "/user/ec2-user/scale4/part.tbl"
Successfully read 40000 records from: "/user/ec2-user/scale4/supplier.tbl"

Output(s):
Successfully stored 280 records (7217 bytes) in: "/user/ec2-user/Part2B_pig"

Counters:
Total records written : 280
Total bytes written : 7217
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
2016-11-22 23:41:27,189 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,192 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,239 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,242 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,265 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,268 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,293 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,296 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,320 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,324 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,346 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,349 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,374 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,377 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,403 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,406 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,428 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,432 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,455 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,458 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,479 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,483 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,504 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,507 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,547 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,559 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,580 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,584 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,606 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,610 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,632 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-11-22 23:41:27,648 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 48 seconds and 376 milliseconds (228376 ms)
```

Query 2.1 with prejoin1 -- output file:

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/Part2B_pig
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-22 23:41 /user/ec2-user/Part2B_pig/_SUCCESS
-rw-r--r--   2 ec2-user supergroup       7217 2016-11-22 23:41 /user/ec2-user/Part2B_pig/part-r-00000
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/Part2B_pig/part-r-00000
1207004714|1998|MFGR#129
1285428693|1998|MFGR#1238
1350601347|1998|MFGR#1229
1358086103|1998|MFGR#1222
1383876006|1998|MFGR#1213
1406153892|1998|MFGR#1219
1409115372|1998|MFGR#1223
1441898473|1998|MFGR#1230
1445039464|1998|MFGR#1231
1451906941|1998|MFGR#1225
1457307800|1998|MFGR#1220
1459545128|1998|MFGR#1239
1466946762|1998|MFGR#126
1470503353|1998|MFGR#123
1477715730|1998|MFGR#125
1508262019|1998|MFGR#1210
1509229650|1998|MFGR#1212
1522581193|1998|MFGR#1218
1525737275|1998|MFGR#124
1530901152|1998|MFGR#1221
1532309319|1998|MFGR#1234
1532687418|1998|MFGR#1237
1538644707|1998|MFGR#128
1538979218|1998|MFGR#1233
1539587552|1998|MFGR#1214
1565657860|1998|MFGR#1227
1568652498|1998|MFGR#1224
1577658136|1998|MFGR#1236
1584019774|1998|MFGR#1217
1587370161|1998|MFGR#1240
```

Part 3: Clustering

3A. Using Mahout as you have in a previous assignment (by default it takes input from testdata in /home/ec2-user).

I tried working with Mahout here using the files created in 1-C – which resulted in an Out of memory error.

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/testdata/
Found 10 items
-rw-r--r--   2 ec2-user supergroup  225891923 2016-11-22 00:06 /user/ec2-user/testdata/000000_0
-rw-r--r--   2 ec2-user supergroup  225715492 2016-11-22 00:06 /user/ec2-user/testdata/000001_0
-rw-r--r--   2 ec2-user supergroup  225887867 2016-11-22 00:06 /user/ec2-user/testdata/000002_0
-rw-r--r--   2 ec2-user supergroup  226074967 2016-11-22 00:06 /user/ec2-user/testdata/000003_0
-rw-r--r--   2 ec2-user supergroup  226205624 2016-11-22 00:06 /user/ec2-user/testdata/000004_0
-rw-r--r--   2 ec2-user supergroup  226307887 2016-11-22 00:06 /user/ec2-user/testdata/000005_0
-rw-r--r--   2 ec2-user supergroup  226303263 2016-11-22 00:06 /user/ec2-user/testdata/000006_0
-rw-r--r--   2 ec2-user supergroup  226304298 2016-11-22 00:07 /user/ec2-user/testdata/000007_0
-rw-r--r--   2 ec2-user supergroup  226297078 2016-11-22 00:07 /user/ec2-user/testdata/000008_0
-rw-r--r--   2 ec2-user supergroup    1548313 2016-11-22 00:07 /user/ec2-user/testdata/000009_0
[ec2-user@ip-172-31-46-250 ~]$ time mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job
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
16/11/22 00:08:33 WARN MahoutDriver: No org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.props found on classpath, will use command-line arguments only
16/11/22 00:08:33 INFO Job: Running with default arguments
16/11/22 00:08:34 INFO HadoopUtil: Deleting output
16/11/22 00:08:34 INFO Job: Preparing Input
16/11/22 00:08:34 INFO RMProxy: Connecting to ResourceManager at /172.31.46.250:8032
16/11/22 00:08:35 WARN JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with To
olRunner to remedy this.
16/11/22 00:08:35 INFO FileInputFormat: Total input paths to process : 10
16/11/22 00:08:35 INFO JobSubmitter: number of splits:19
16/11/22 00:08:36 INFO JobSubmitter: Submitting tokens for job: job_1479756709677_0015
16/11/22 00:08:36 INFO YarnClientImpl: Submitted application application_1479756709677_0015
16/11/22 00:08:36 INFO Job: The url to track the job: http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0015/
16/11/22 00:08:36 INFO Job: Running job: job_1479756709677_0015
16/11/22 00:08:42 INFO Job: Job job_1479756709677_0015 running in uber mode : false
16/11/22 00:08:42 INFO Job:  map 0% reduce 0%
16/11/22 00:08:56 INFO Job:  map 1% reduce 0%
16/11/22 00:08:57 INFO Job:  map 7% reduce 0%
16/11/22 00:08:59 INFO Job:  map 11% reduce 0%
16/11/22 00:09:02 INFO Job:  map 14% reduce 0%
16/11/22 00:09:05 INFO Job:  map 16% reduce 0%
16/11/22 00:09:07 INFO Job:  map 17% reduce 0%
16/11/22 00:09:09 INFO Job:  map 20% reduce 0%
16/11/22 00:09:10 INFO Job:  map 21% reduce 0%
16/11/22 00:09:12 INFO Job:  map 25% reduce 0%
16/11/22 00:09:13 INFO Job:  map 27% reduce 0%
16/11/22 00:09:15 INFO Job:  map 31% reduce 0%
16/11/22 00:09:16 INFO Job:  map 34% reduce 0%
16/11/22 00:09:18 INFO Job:  map 37% reduce 0%
16/11/22 00:09:19 INFO Job:  map 40% reduce 0%
16/11/22 00:09:21 INFO Job:  map 44% reduce 0%
16/11/22 00:09:22 INFO Job:  map 47% reduce 0%
16/11/22 00:09:24 INFO Job:  map 51% reduce 0%
16/11/22 00:09:25 INFO Job:  map 53% reduce 0%
16/11/22 00:09:27 INFO Job:  map 57% reduce 0%
16/11/22 00:09:28 INFO Job:  map 60% reduce 0%
16/11/22 00:09:30 INFO Job:  map 62% reduce 0%
16/11/22 00:09:31 INFO Job:  map 65% reduce 0%
16/11/22 00:09:33 INFO Job:  map 67% reduce 0%
16/11/22 00:09:34 INFO Job:  map 70% reduce 0%
16/11/22 00:09:36 INFO Job:  map 72% reduce 0%
```

```
                Total megabyte-milliseconds taken by all map tasks=1834656768
        Map-Reduce Framework
                Map input records=23996604
                Map output records=23996604
                Input split bytes=4588
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=15902
                CPU time spent (ms)=310200
                Physical memory (bytes) snapshot=6869221376
                Virtual memory (bytes) snapshot=36785410048
                Total committed heap usage (bytes)=35981885544
        File Input Format Counters
                Bytes Read=3347912463
        File Output Format Counters
                Bytes Written=4466837696
16/11/22 00:26:05 INFO Job: Dumping out clusters from clusters: output/clusters-*-final and clusteredPoints: output/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$5.hasNext(Iterators.java:543)
        at com.google.common.collect.ForwardingIterator.hasNext(ForwardingIterator.java:43)
        at org.apache.mahout.utils.clustering.ClusterDumper.readPoints(ClusterDumper.java:311)
        at org.apache.mahout.utils.clustering.ClusterDumper.init(ClusterDumper.java:262)
        at org.apache.mahout.utils.clustering.ClusterDumper.<init>(ClusterDumper.java:92)
        at org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.run(Job.java:141)
        at org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.main(Job.java:60)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.ProgramDriver$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)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)

real    18m1.300s
user    2m1.580s
sys     0m4.108s
[ec2-user@ip-172-31-46-250 ~]$
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Worked with a reasonably small sample that resulted in the same error again.

```
        File Input Format Counters
                Bytes Read=1865369688
        File Output Format Counters
                Bytes Written=2489454743
16/11/22 00:45:42 INFO Job: Dumping out clusters from clusters: output/clusters-*-final and clusteredPoints: output/clusteredPoints
Exception in thread "main" java.lang.OutOfMemoryError: GC overhead limit exceeded
        at org.apache.hadoop.hdfs.DFSInputStream.readWithStrategy(DFSInputStream.java:831)
        at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:889)
        at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:696)
        at java.io.DataInputStream.readInt(DataInputStream.java:390)
        at org.apache.hadoop.io.SequenceFile$Reader.readRecordLength(SequenceFile.java:2398)
        at org.apache.hadoop.io.SequenceFile$Reader.next(SequenceFile.java:2429)
        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$5.hasNext(Iterators.java:543)
        at com.google.common.collect.ForwardingIterator.hasNext(ForwardingIterator.java:43)
        at org.apache.mahout.utils.clustering.ClusterDumper.readPoints(ClusterDumper.java:311)
        at org.apache.mahout.utils.clustering.ClusterDumper.init(ClusterDumper.java:262)
        at org.apache.mahout.utils.clustering.ClusterDumper.<init>(ClusterDumper.java:92)
        at org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.run(Job.java:141)
        at org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.main(Job.java:60)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.ProgramDriver$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    11m39.583s
user    1m38.628s
sys     0m3.784s
[ec2-user@ip-172-31-46-250 ~]$
```

KMeans clustering with a smaller file –
time mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cp /user/ec2-user/lineorder_numeric/000009_0 /user/ec2-user/testdata/
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/testdata/
Found 1 items
-rw-r--r--   2 ec2-user supergroup    1548313 2016-11-22 17:26 /user/ec2-user/testdata/000009_0
[ec2-user@ip-172-31-46-250 ~]$ time mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job
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
16/11/22 17:28:32 WARN MahoutDriver: No org.apache.mahout.clustering.syntheticcontrol.kmeans.Job.props found on classpath, will use command-line arguments only
16/11/22 17:28:32 INFO Job: Running with default arguments
16/11/22 17:28:33 INFO HadoopUtil: Deleting output
16/11/22 17:28:33 INFO Job: Preparing Input
16/11/22 17:28:33 INFO RMProxy: Connecting to ResourceManager at /172.31.46.250:8032
16/11/22 17:28:33 WARN JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
16/11/22 17:28:35 INFO FileInputFormat: Total input paths to process : 1
16/11/22 17:28:35 INFO JobSubmitter: number of splits:1
16/11/22 17:28:35 INFO JobSubmitter: Submitting tokens for job: job_1479756709677_0101
16/11/22 17:28:35 INFO YarnClientImpl: Submitted application application_1479756709677_0101
16/11/22 17:28:35 INFO Job: The url to track the job: http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0101/
16/11/22 17:28:35 INFO Job: Running job: job_1479756709677_0101
16/11/22 17:28:40 INFO Job: Job job_1479756709677_0101 running in uber mode : false
16/11/22 17:28:40 INFO Job:  map 0% reduce 0%
16/11/22 17:28:46 INFO Job:  map 100% reduce 0%
16/11/22 17:28:46 INFO Job: Job job_1479756709677_0101 completed successfully
16/11/22 17:28:47 INFO Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=106550
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1548430
                HDFS: Number of bytes written=2530523
                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
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3589
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3589
                Total vcore-milliseconds taken by all map tasks=3589
                Total megabyte-milliseconds taken by all map tasks=3675136
        Map-Reduce Framework
                Map input records=18145
                Map output records=18145
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Runtime of the command for the smaller file – 210238 ms or 3.50 minutes

```
        1.0 : [distance=3064557.160007217]: [2.3999904E7,2.0,100930.0,253974.0,14296.0,1.996022E7,21.0,4048716.0,2.7238365E7,10.0,3643844.0,115677.0,8.0,1.9960519E7]
        1.0 : [distance=2175825.1890942436]: [2.3999904E7,3.0,100930.0,133945.0,21922.0,1.996022E7,6.0,1187364.0,2.7238365E7,9.0,1080501.0,118736.0,2.0,1.9960513E7]
        1.0 : [distance=2918796.620683257]: [2.3999904E7,4.0,100930.0,10517.0,34989.0,1.996022E7,27.0,3854277.0,2.7238365E7,7.0,3584477.0,85650.0,3.0,1.9960322E7]
        1.0 : [distance=2447050.55909762253]: [2.3999904E7,6.0,100930.0,85792.0,1818.0,1.996022E7,5.0,888895.0,2.7238365E7,10.0,800005.0,106667.0,8.0,1.9960507E7]
        1.0 : [distance=2530625.0850828337]: [2.3999907E7,1.0,88108.0,130535.0,16997.0,1.9940822E7,4.0,626212.0,2.4048636E7,7.0,582377.0,93931.0,1.0,1.994112E7]
        1.0 : [distance=2150844.757453336]: [{"0":2.3999907E7},{"1":2.0},{"2":88108.0},{"3":301671.0},{"4":18701.0},{"5":1.9940822E7},{"6":19.0},{"7":3178054.0},{"8":2.4048636E7},{"10":3178054.0
"11":100359.0},{"12":5.0},{"13":1.9940921E7}]
        1.0 : [distance=1473372.0437689277]: [2.3999907E7,4.0,88108.0,130164.0,4991.0,1.9940822E7,17.0,2030072.0,2.4048636E7,9.0,1847365.0,71649.0,1.0,1.9941113E7]
        1.0 : [distance=1679216.4951881825]: [2.3999907E7,5.0,88108.0,57675.0,30653.0,1.9940822E7,17.0,2775539.0,2.4048636E7,10.0,2497985.0,97960.0,6.0,1.9941116E7]
        1.0 : [distance=1818550.9121847162]: [{"0":2.399991E7},{"1":1.0},{"2":80446.0},{"3":242210.0},{"4":28979.0},{"5":1.9951219E7},{"6":28.0},{"7":3226160.0},{"8":2.4732649E7},{"10":3226160.0
"11":69132.0},{"13":1.9960229E7}]
        1.0 : [distance=2255328.696411789]: [{"0":2.399991E7},{"1":5.0},{"2":80446.0},{"3":308893.0},{"4":23422.0},{"5":1.9951219E7},{"6":3.0},{"7":570564.0},{"8":2.4732649E7},{"10":570564.0},{"
:114112.0},{"12":4.0},{"13":1.9960315E7}]
        1.0 : [distance=2539809.900068507]: [2.399991E7,6.0,80446.0,353806.0,6068.0,1.9951219E7,2.0,371958.0,2.4732649E7,7.0,345920.0,111587.0,5.0,1.9960202E7]
        1.0 : [distance=2427556.59708296]: [2.399994E7,1.0,59602.0,259002.0,30454.0,1.9950723E7,37.0,3555663.0,2.4043776E7,7.0,3306766.0,57659.0,1.0,1.9950919E7]
        1.0 : [distance=1867374.1623127232]: [{"0":2.399994E7},{"1":3.0},{"2":59602.0},{"3":18771.0},{"4":15171.0},{"5":1.9950723E7},{"6":17.0},{"7":2872609.0},{"8":2.4043776E7},{"10":2872609.0}
11":101386.0},{"12":7.0},{"13":1.9951013E7}]
        1.0 : [distance=3265840.1528564696]: [{"0":2.3999941E7},{"1":2.0},{"2":66772.0},{"3":262168.0},{"4":38619.0},{"5":1.9980217E7},{"6":34.0},{"7":3842510.0},{"8":2.7871821E7},{"9":10.0},{"1
3458259.0},{"11":67809.0},{"13":1.998042 6E7}]
        1.0 : [distance=2367347.19981037]: [2.3999941E7,4.0,66772.0,248214.0,30839.0,1.9980217E7,19.0,2208180.0,2.7871821E7,6.0,2075689.0,69732.0,1.0,1.9980517E7]
        1.0 : [distance=2468943.642084606]: [{"0":2.3999941E7},{"1":5.0},{"2":66772.0},{"3":131196.0},{"4":24051.0},{"5":1.9980217E7},{"6":21.0},{"7":2577099.0},{"8":2.7871821E7},{"9":1.0},{"10'
51328.0},{"11":73631.0},{"13":1.998042E7}]
        1.0 : [distance=2432288.6545535666]: [2.3999941E7,7.0,66772.0,176301.0,29162.0,1.9980217E7,18.0,2479140.0,2.7871821E7,1.0,2454348.0,82638.0,5.0,1.9980331E7]
        1.0 : [distance=2712936.833478343]: [{"0":2.399997E7},{"1":1.0},{"2":69265.0},{"3":213929.0},{"4":18219.0},{"5":1.9950908E7},{"6":5.0},{"7":921455.0},{"8":2.7684925E7},{"10":921455.0},{'
:110574.0},{"12":6.0},{"13":1.9951028E7}]
        1.0 : [distance=3124802.8934372803]: [2.399997E7,2.0,69265.0,288021.0,34917.0,1.9950908E7,38.0,3834238.0,2.7684925E7,10.0,3450814.0,60540.0,2.0,1.995113E7]
        1.0 : [distance=2795763.8039835375]: [2.399997E7,3.0,69265.0,368019.0,34639.0,1.9950908E7,8.0,869600.0,2.7684925E7,9.0,791336.0,65220.0,7.0,1.9951114E7]
        1.0 : [distance=4266320.287766789]: [2.3999973E7,2.0,42319.0,341873.0,38406.0,1.9960611E7,2.0,382972.0,2.1972075E7,1.0,379142.0,114891.0,5.0,1.9960731E7]
        1.0 : [distance=3539804.8421776025]: [2.3999973E7,3.0,42319.0,253720.0,33454.0,1.9960611E7,12.0,2008452.0,2.1972075E7,5.0,1908029.0,100422.0,2.0,1.9960712E7]
        1.0 : [distance=3927346.3459854797]: [2.3999973E7,7.0,42319.0,351054.0,34301.0,1.9960611E7,8.0,884032.0,2.1972075E7,5.0,839830.0,66302.0,3.0,1.9960814E7]
        1.0 : [distance=2487318.7912235941]: [2.3999974E7,5.0,107303.0,368703.0,30741.0,1.9970314E7,2.0,354338.0,2.4997105E7,3.0,343707.0,106301.0,7.0,1.9970531E7]
16/11/22 17:32:02 INFO ClusterDumper: Wrote 6 clusters
16/11/22 17:32:02 INFO MahoutDriver: Program took 210238 ms (Minutes: 3.5039666666666665)

real    3m35.616s
user    0m15.580s
sys     0m2.828s
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/output
Found 15 items
-rw-r--r--   2 ec2-user supergroup        194 2016-11-22 17:31 /user/ec2-user/output/_policy
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:31 /user/ec2-user/output/clusteredPoints
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:28 /user/ec2-user/output/clusters-0
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:29 /user/ec2-user/output/clusters-1
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:31 /user/ec2-user/output/clusters-10-final
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:29 /user/ec2-user/output/clusters-2
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:29 /user/ec2-user/output/clusters-3
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:30 /user/ec2-user/output/clusters-4
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:30 /user/ec2-user/output/clusters-5
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:30 /user/ec2-user/output/clusters-6
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:30 /user/ec2-user/output/clusters-7
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:31 /user/ec2-user/output/clusters-8
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:31 /user/ec2-user/output/clusters-9
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:28 /user/ec2-user/output/data
drwxr-xr-x   - ec2-user supergroup          0 2016-11-22 17:28 /user/ec2-user/output/random-seeds
[ec2-user@ip-172-31-46-250 ~]$
```

The file clusteranalyze.txt contains the results of the Kmeans run after 10 iterations:

```
[ec2-user@ip-172-31-46-250 ~]$ mahout clusterdump --input output/clusters-10-final --pointsDir output/clusteredPoints --output clusteranalyze.txt
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
16/11/22 17:51:27 INFO AbstractJob: Command line arguments: {--dictionaryType=[text], --distanceMeasure=[org.apache.mahout.common.distance.SquaredEuclideanDistanceMeasure],
 --endPhase=[2147483647], --input=[output/clusters-10-final], --output=[clusteranalyze.txt], --outputFormat=[TEXT], --pointsDir=[output/clusteredPoints], --startPhase=[0],
--tempDir=[temp]}
16/11/22 17:51:29 INFO ClusterDumper: Wrote 6 clusters
16/11/22 17:51:29 INFO MahoutDriver: Program took 2696 ms (Minutes: 0.04493333333333333)
```

```
-rw-rw-r--  1 ec2-user ec2-user   3112010 Nov 22 17:51 clusteranalyze.txt
[ec2-user@ip-172-31-46-250 ~]$ head -n 100 clusteranalyze.txt
{"r":[5224.141,1.358,34837.603,115710.299,11499.154,18997.415,14.376,2270062.596,1870905.661,3.19,2164025.967,17742.877,2.588,19007.331],"c":[2.3991233946E7,2.539,60804.166,196380.826,20125.197,1.
9950034633E7,24.478,3619854.528,1.3028532467E7,5.001,3439860.851,89009.796,4.024,1.99515754 65E7],"n":4028,"identifier":"CL-15359"}
        Weight : [props - optional]:  Point:
        1.0 : [distance=2916909.417619272]: [2.3982209E7,1.0,100858.0,312175.0,3873.0,1.997052E7,33.0,3917628.0,1.0163071E7,1.0,3878451.0,71229.0,7.0,1.9970727E7]
        1.0 : [distance=4438312.2006252]: [2.3982209E7,2.0,100858.0,320906.0,10194.0,1.997052E7,32.0,6166048.0,1.0163071E7,8.0,5672764.0,115613.0,6.0,1.9970806E7]
        1.0 : [distance=3490037.6164534385]: [2.3982212E7,1.0,99019.0,177587.0,22848.0,1.9980724E7,36.0,5992488.0,1.4122844E7,4.0,5752788.0,99874.0,6.0,1.9980918E7]
        1.0 : [distance=3686889.201931284]: [2.3982212E7,2.0,99019.0,301187.0,3531.0,1.9980724E7,9.0,1069353.0,1.4122844E7,5.0,1015885.0,71290.0,6.0,1.9980901E7]
        1.0 : [distance=2915692.4391340557]: [2.3982212E7,3.0,99019.0,49556.0,25134.0,1.9980724E7,11.0,1656105.0,1.4122844E7,4.0,1589860.0,90333.0,8.0,1.9980924E7]
        1.0 : [distance=1193957.0353482847]: [2.3982212E7,4.0,99019.0,76504.0,13928.0,1.9980724E7,22.0,3257100.0,1.4122844E7,3.0,3159387.0,88830.0,3.0,1.9981016E7]
        1.0 : [distance=3042409.2519207536]: [2.3982212E7,5.0,99019.0,191611.0,34724.0,1.9980724E7,9.0,1532349.0,1.4122844E7,1.0,1517025.0,102156.0,7.0,1.998102E7]
        1.0 : [distance=4634989.6111805905]: [2.3982212E7,6.0,99019.0,396111.0,5317.0,1.9980724E7,3.0,362130.0,1.4122844E7,7.0,336780.0,72426.0,5.0,1.9980919E7]
        1.0 : [distance=2618202.7818832137]: [2.3982214E7,1.0,46991.0,341521.0,32742.0,1.9970805E7,18.0,2812518.0,1.5353536E7,9.0,2559391.0,93750.0,3.0,1.9970916E7]
        1.0 : [distance=2384790.1613065456]: [2.3982214E7,2.0,46991.0,328576.0,5971.0,1.9970805E7,25.0,4011400.0,1.5353536E7,6.0,3770716.0,96273.0,2.0,1.9971004E7]
        1.0 : [distance=7787870.87373128]: [{"0":2.3982214E7},{"1":3.0},{"2":46991.0},{"3":364692.0},{"4":20470.0},{"5":1.9970805E7},{"6":50.0},{"7":8783400.0},{"8":1.5353536E7},{"10":8783400.0},{
"11":105400.0},{"12":1.0},{"13":1.9971014E7}]
        1.0 : [distance=4547830.106915714]: [2.3982275E7,1.0,8440.0,109388.0,3914.0,1.9961009E7,49.0,6847162.0,1.2963129E7,3.0,6641747.0,83842.0,2.0,1.9961202E7]
        1.0 : [distance=658628.243776662]: [2.3982275E7,2.0,8440.0,73169.0,6695.0,1.9961009E7,35.0,3997560.0,1.2963129E7,1.0,3957584.0,68529.0,5.0,1.9961124E7]
        1.0 : [distance=1977235.1540566692]: [{"0":2.3982275E7},{"1":3.0},{"2":8440.0},{"3":145444.0},{"4":543.0},{"5":1.9961009E7},{"6":15.0},{"7":2234160.0},{"8":1.2963129E7},{"9":9.0},{"10":203
3085.0},{"11":89366.0},{"13":1.9961207E7}]
        1.0 : [distance=2668001.520884008]: [2.3982278E7,1.0,30289.0,246353.0,32798.0,1.998042E7,30.0,3898020.0,1.5647613E7,1.0,3859039.0,77960.0,6.0,1.9980702E7]
        1.0 : [distance=4266221.891658649]: [2.3982278E7,2.0,30289.0,47987.0,25190.0,1.998042E7,31.0,5998438.0,1.5647613E7,3.0,5818484.0,116098.0,5.0,1.998063E7]
        1.0 : [distance=2815874.498494917]: [2.3982278E7,3.0,30289.0,104501.0,35811.0,1.998042E7,19.0,2860450.0,1.5647613E7,4.0,2746032.0,90330.0,2.0,1.9980701E7]
        1.0 : [distance=2838186.7274432457]: [2.3982278E7,4.0,30289.0,335784.0,30791.0,1.998042E7,16.0,2911632.0,1.5647613E7,10.0,2620468.0,109186.0,1.0,1.9980705E7]
        1.0 : [distance=4407934.905545453]: [{"0":2.3982336E7},{"1":1.0},{"2":34676.0},{"3":193070.0},{"4":33155.0},{"5":1.9970324E7},{"6":4.0},{"7":465228.0},{"8":1.2370586E7},{"9":7.0},{"10":432
662.0},{"11":69784.0},{"13":1.9970529E7}]
        1.0 : [distance=747229.5906538777]: [{"0":2.3982336E7},{"1":2.0},{"2":34676.0},{"3":126475.0},{"4":31190.0},{"5":1.9970324E7},{"6":22.0},{"7":3303234.0},{"8":1.2370586E7},{"10":3303234.0},
{"11":90088.0},{"12":4.0},{"13":1.9970526E7}]
        1.0 : [distance=1791431.233012169]: [2.3982336E7,3.0,34676.0,60521.0,1567.0,1.9970324E7,16.0,2370432.0,1.2370586E7,1.0,2346727.0,88891.0,7.0,1.9970502E7]
        1.0 : [distance=3330046.246661073]: [2.3982336E7,4.0,34676.0,372235.0,6881.0,1.9970324E7,46.0,6013212.0,1.2370586E7,6.0,5652419.0,78433.0,6.0,1.9970428E7]
        1.0 : [distance=4199960.855044604]: [2.3982342E7,1.0,83969.0,180437.0,28956.0,1.9960804E7,5.0,758715.0,1.1551289E7,2.0,743540.0,91045.0,1.0,1.9961031E7]
        1.0 : [distance=2768396.1702685365]: [2.3982342E7,2.0,83969.0,198458.0,9066.0,1.9960804E7,35.0,5447575.0,1.1551289E7,10.0,4902817.0,93387.0,3.0,1.9961001E7]
        1.0 : [distance=1520333.9686267152]: [{"0":2.3982342E7},{"1":3.0},{"2":83969.0},{"3":271463.0},{"4":8737.0},{"5":1.9960804E7},{"6":23.0},{"7":3299235.0},{"8":1.1551289E7},{"10":3299235.0},
{"11":86067.0},{"12":5.0},{"13":1.9960917E7}]
        1.0 : [distance=2261423.2903119884]: [2.3982342E7,4.0,83969.0,69863.0,39697.0,1.9960804E7,13.0,2382718.0,1.1551289E7,5.0,2263582.0,109971.0,1.0,1.9960928E7]
        1.0 : [distance=8486005.465749359]: [2.398237E7,2.0,109714.0,379987.0,19749.0,1.9940706E7,45.0,9301365.0,9486391.0,7.0,8650269.0,124018.0,2.0,1.9940919E7]
        1.0 : [distance=3532356.354912539]: [2.3982371E7,1.0,26348.0,23572.0,35481.0,1.9940518E7,9.0,1346013.0,1.4642548E7,5.0,1278712.0,89734.0,6.0,1.9940618E7]
        1.0 : [distance=5070210.631569067]: [2.3982371E7,2.0,26348.0,168311.0,32733.0,1.9940518E7,1.0,137931.0,1.4642548E7,8.0,126896.0,82758.0,2.0,1.9940628E7]
        1.0 : [distance=1905286.5998879275]: [2.3982371E7,3.0,26348.0,396330.0,16303.0,1.9940518E7,31.0,4421592.0,1.4642548E7,9.0,4023648.0,85579.0,6.0,1.9940709E7]
        1.0 : [distance=4832435.594241842]: [2.3982371E7,4.0,26348.0,318052.0,209.0,1.9940518E7,3.0,321012.0,1.4642548E7,6.0,301751.0,64202.0,5.0,1.9940627E7]
        1.0 : [distance=1620399.0591706007]: [{"0":2.3982371E7},{"1":5.0},{"2":26348.0},{"3":293383.0},{"4":19623.0},{"5":1.9940518E7},{"6":27.0},{"7":3716199.0},{"8":1.4642548E7},{"9":8.0},{"10":
3418903.0},{"11":82582.0},{"13":1.9940706E7}]
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Part 4: Performance

4A. Query 2.1 in Hive with

a. **Scale1: single node and cluster of at least 4 nodes**

Query 2.1 -- Single node: Time taken 188.355

```
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 release
s.
Query ID = ec2-user_20161122063917_2498e2a4-c765-48db-b907-561b55453c17
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_20161122063917_2498e2a4-c765-48db-b907-561b55453c17.log
2016-11-22 06:39:27     Starting to launch local task to process map join;      maximum memory = 518979584
2016-11-22 06:39:29     Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/2856b2a7-311e-4095-9e88-2912645c47cd/hive_2016-11-22_06-39-17_876_195687
8932793697491-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile31--.hashtable
2016-11-22 06:39:29     Uploaded 1 File to: file:/tmp/ec2-user/2856b2a7-311e-4095-9e88-2912645c47cd/hive_2016-11-22_06-39-17_876_1956878932793697491-1/-local-10014/HashTable-Stage-
13/MapJoin-mapfile31--.hashtable (67039 bytes)
2016-11-22 06:39:29     End of local task; Time Taken: 2.039 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_1479792040574_0001, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1479792040574_0001/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479792040574_0001
Hadoop job information for Stage-13: number of mappers: 3; number of reducers: 0
2016-11-22 06:39:41,688 Stage-13 map = 0%,  reduce = 0%
2016-11-22 06:40:14,733 Stage-13 map = 33%,  reduce = 0%, Cumulative CPU 14.43 sec
2016-11-22 06:40:24,887 Stage-13 map = 67%,  reduce = 0%, Cumulative CPU 20.63 sec
2016-11-22 06:40:42,842 Stage-13 map = 83%,  reduce = 0%, Cumulative CPU 27.37 sec
2016-11-22 06:40:43,076 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 27.94 sec
MapReduce Total cumulative CPU time: 27 seconds 940 msec
Ended Job = job_1479792040574_0001
Stage-15 is selected by condition resolver.
Stage-16 is filtered out by condition resolver.
Stage-2 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_20161122063917_2498e2a4-c765-48db-b907-561b55453c17.log
2016-11-22 06:40:51     Starting to launch local task to process map join;      maximum memory = 518979584
2016-11-22 06:40:54     Dump the side-table for tag: 1 with group count: 7883 into file: file:/tmp/ec2-user/2856b2a7-311e-4095-9e88-2912645c47cd/hive_2016-11-22_06-39-17_876_195687
8932793697491-1/-local-10010/HashTable-Stage-10/MapJoin-mapfile11--.hashtable
2016-11-22 06:40:54     Uploaded 1 File to: file:/tmp/ec2-user/2856b2a7-311e-4095-9e88-2912645c47cd/hive_2016-11-22_06-39-17_876_1956878932793697491-1/-local-10010/HashTable-Stage-
10/MapJoin-mapfile11--.hashtable (249337 bytes)
2016-11-22 06:40:54     End of local task; Time Taken: 2.796 sec.
```

```
344889822        1998     MFGR#1223
396906691        1998     MFGR#1224
290208878        1998     MFGR#1225
419415707        1998     MFGR#1226
358466340        1998     MFGR#1227
251549955        1998     MFGR#1228
383138860        1998     MFGR#1229
296330561        1998     MFGR#123
437181243        1998     MFGR#1230
398944492        1998     MFGR#1231
424062455        1998     MFGR#1232
406967188        1998     MFGR#1233
428867240        1998     MFGR#1234
352277781        1998     MFGR#1235
361827086        1998     MFGR#1236
341618569        1998     MFGR#1237
244739231        1998     MFGR#1238
414151803        1998     MFGR#1239
330082371        1998     MFGR#124
415312453        1998     MFGR#1240
360289624        1998     MFGR#125
341657580        1998     MFGR#126
377507061        1998     MFGR#127
361416497        1998     MFGR#128
318769573        1998     MFGR#129
Time taken: 188.355 seconds, Fetched: 280 row(s)
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Query 2.1 – Cluster of 4 nodes: Time taken 106.149 seconds

```
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_20161122065931_736e362e-84e3-47a9-8f11-412f9d4322dd
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_20161122065931_736e362e-84e3-47a9-8f11-412f9d4322dd.log
2016-11-22 06:59:38     Starting to launch local task to process map join;      maximum memory = 477102080
2016-11-22 06:59:39     Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/81a6a05a-87a0-4aac-9dd5-ce7b3788d12c/hive_2016-11-22_06-59-31_60
7_8385144612754686981-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile31--.hashtable
2016-11-22 06:59:39     Uploaded 1 File to: file:/tmp/ec2-user/81a6a05a-87a0-4aac-9dd5-ce7b3788d12c/hive_2016-11-22_06-59-31_607_8385144612754686981-1/-local-10014/HashTabl
e-Stage-13/MapJoin-mapfile31--.hashtable (67039 bytes)
2016-11-22 06:59:39     End of local task; Time Taken: 1.193 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_1479756709677_0093, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0093/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479756709677_0093
Hadoop job information for Stage-13: number of mappers: 3; number of reducers: 0
2016-11-22 06:59:45,454 Stage-13 map = 0%,  reduce = 0%
2016-11-22 06:59:56,204 Stage-13 map = 17%,  reduce = 0%, Cumulative CPU 7.28 sec
2016-11-22 06:59:58,328 Stage-13 map = 50%,  reduce = 0%, Cumulative CPU 19.59 sec
2016-11-22 07:00:00,404 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 26.5 sec
MapReduce Total cumulative CPU time: 26 seconds 500 msec
```

```
312440027        1998     MFGR#122
348123961        1998     MFGR#1220
339845398        1998     MFGR#1221
355416161        1998     MFGR#1222
344889822        1998     MFGR#1223
396906691        1998     MFGR#1224
290208878        1998     MFGR#1225
419415707        1998     MFGR#1226
358466340        1998     MFGR#1227
251549955        1998     MFGR#1228
383138860        1998     MFGR#1229
296330561        1998     MFGR#123
437181243        1998     MFGR#1230
398944492        1998     MFGR#1231
424062455        1998     MFGR#1232
406967188        1998     MFGR#1233
428867240        1998     MFGR#1234
352277781        1998     MFGR#1235
361827086        1998     MFGR#1236
341618569        1998     MFGR#1237
244739231        1998     MFGR#1238
414151803        1998     MFGR#1239
330082371        1998     MFGR#124
415312453        1998     MFGR#1240
360289624        1998     MFGR#125
341657580        1998     MFGR#126
377507061        1998     MFGR#127
361416497        1998     MFGR#128
318769573        1998     MFGR#129
Time taken: 106.149 seconds, Fetched: 280 row(s)
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

b. Scale4: single node and a cluster of at least 4 nodes

Query 2.1 –Single node:

Disk space available before execution -

```
[ec2-user@ip-172-31-13-121 ~]$ ls -ltr
total 16
drwxr-xr-x 10 ec2-user ec2-user 4096 Oct 14 17:43 hadoop-2.6.4
drwxr-xr-x 16 ec2-user ec2-user 4096 Nov  3 19:19 pig-0.15.0
drwxrwxr-x  9 ec2-user ec2-user 4096 Nov 22 18:13 apache-hive-2.0.1-bin
drwxrwxr-x  2 ec2-user ec2-user 4096 Nov 22 18:18 scale4
[ec2-user@ip-172-31-13-121 ~]$ df -h
Filesystem      Size  Used Avail Use% Mounted on
devtmpfs        992M   56K  992M   1% /dev
tmpfs          1002M     0 1002M   0% /dev/shm
/dev/xvda1      7.8G  4.0G  3.7G  53% /
[ec2-user@ip-172-31-13-121 ~]$
```

Map reduce task gets stuck in between and there is no progress-

```
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_20161122182410_0ee0f003-ff98-450b-8a56-bac954046092
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_20161122182410_0ee0f003-ff98-450b-8a56-bac954046092.log
2016-11-22 18:24:19     Starting to launch local task to process map join;      maximum memory = 518979584
2016-11-22 18:24:21     Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/b1c84851-65e5-4778-85a2-26ed780c529e/hive_2016-11-22_18-24-10_470_116179467020372
7503-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile31--.hashtable
2016-11-22 18:24:21     Uploaded 1 File to: file:/tmp/ec2-user/b1c84851-65e5-4778-85a2-26ed780c529e/hive_2016-11-22_18-24-10_470_1161794670203727503-1/-local-10014/HashTable-Stage-13/MapJoi
n-mapfile31--.hashtable (67039 bytes)
2016-11-22 18:24:21     End of local task; Time Taken: 1.704 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_1479838664462_0001, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1479838664462_0001/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479838664462_0001
Hadoop job information for Stage-13: number of mappers: 10; number of reducers: 0
2016-11-22 18:24:33,168 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:25:33,838 Stage-13 map = 0%,  reduce = 0%, Cumulative CPU 26.36 sec
2016-11-22 18:25:55,769 Stage-13 map = 25%,  reduce = 0%, Cumulative CPU 47.26 sec
2016-11-22 18:25:56,985 Stage-13 map = 30%,  reduce = 0%, Cumulative CPU 47.71 sec
2016-11-22 18:26:22,401 Stage-13 map = 40%,  reduce = 0%, Cumulative CPU 70.16 sec
2016-11-22 18:26:23,613 Stage-13 map = 60%,  reduce = 0%, Cumulative CPU 71.75 sec
2016-11-22 18:27:01,498 Stage-13 map = 70%,  reduce = 0%, Cumulative CPU 84.45 sec
2016-11-22 18:27:35,466 Stage-13 map = 85%,  reduce = 0%, Cumulative CPU 98.3 sec
2016-11-22 18:27:49,485 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:28:50,145 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:29:50,860 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:30:51,764 Stage-13 map = 0%,  reduce = 0%, Cumulative CPU 23.49 sec
2016-11-22 18:31:14,587 Stage-13 map = 10%,  reduce = 0%, Cumulative CPU 45.8 sec
2016-11-22 18:31:18,098 Stage-13 map = 30%,  reduce = 0%, Cumulative CPU 48.55 sec
2016-11-22 18:31:42,919 Stage-13 map = 45%,  reduce = 0%, Cumulative CPU 70.94 sec
2016-11-22 18:31:44,018 Stage-13 map = 60%,  reduce = 0%, Cumulative CPU 71.45 sec
2016-11-22 18:31:49,750 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:32:50,328 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:33:50,833 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:34:51,327 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:35:51,860 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:36:52,302 Stage-13 map = 0%,  reduce = 0%
2016-11-22 18:37:52,773 Stage-13 map = 0%,  reduce = 0%
```

Tried restarting Hadoop several times and cleared enough space, but still doesn't seem to work with scale4 data.

```
[ec2-user@ip-172-31-13-121 apache-hive-2.0.1-bin]$ df -h
Filesystem      Size  Used Avail Use% Mounted on
devtmpfs        992M   56K  992M   1% /dev
tmpfs          1002M     0 1002M   0% /dev/shm
/dev/xvda1      7.8G  6.4G  1.4G  83% /
[ec2-user@ip-172-31-13-121 apache-hive-2.0.1-bin]$
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Query 2.1 – Cluster of 4 nodes: Time taken: 200.506 seconds

```
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 relea
ses.
Query ID = ec2-user_20161122073456_a4a2126a-62f0-43fd-894b-bf011aac3c7e
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_20161122073456_a4a2126a-62f0-43fd-894b-bf011aac3c7e.log
2016-11-22 07:35:03     Starting to launch local task to process map join;      maximum memory = 477102080
2016-11-22 07:35:04     Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/ca33226e-6f56-4ca8-b358-f00f3085f9fe/hive_2016-11-22_07-34-56_566_7849
71192315559175-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile31--.hashtable
2016-11-22 07:35:04     Uploaded 1 File to: file:/tmp/ec2-user/ca33226e-6f56-4ca8-b358-f00f3085f9fe/hive_2016-11-22_07-34-56_566_7849711923155591 75-1/-local-10014/HashTable-Stage
-13/MapJoin-mapfile31--.hashtable (67039 bytes)
2016-11-22 07:35:04     End of local task; Time Taken: 1.205 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_1479756709677_0097, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0097/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479756709677_0097
Hadoop job information for Stage-13: number of mappers: 10; number of reducers: 0
2016-11-22 07:35:10,815 Stage-13 map = 0%,  reduce = 0%
2016-11-22 07:35:22,250 Stage-13 map = 10%,  reduce = 0%, Cumulative CPU 2.78 sec
2016-11-22 07:35:25,647 Stage-13 map = 15%,  reduce = 0%, Cumulative CPU 22.93 sec
2016-11-22 07:35:28,064 Stage-13 map = 20%,  reduce = 0%, Cumulative CPU 33.45 sec
2016-11-22 07:35:30,384 Stage-13 map = 25%,  reduce = 0%, Cumulative CPU 39.42 sec
2016-11-22 07:35:31,523 Stage-13 map = 35%,  reduce = 0%, Cumulative CPU 48.39 sec
2016-11-22 07:35:37,096 Stage-13 map = 50%,  reduce = 0%, Cumulative CPU 67.97 sec
2016-11-22 07:35:43,682 Stage-13 map = 65%,  reduce = 0%, Cumulative CPU 82.54 sec
2016-11-22 07:35:44,769 Stage-13 map = 70%,  reduce = 0%, Cumulative CPU 83.63 sec
2016-11-22 07:35:45,886 Stage-13 map = 75%,  reduce = 0%, Cumulative CPU 84.53 sec
2016-11-22 07:35:52,404 Stage-13 map = 85%,  reduce = 0%, Cumulative CPU 98.71 sec
2016-11-22 07:35:54,486 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 102.24 sec
MapReduce Total cumulative CPU time: 1 minutes 42 seconds 240 msec
```

```
1509229650       1998     MFGR#1212
1383876006       1998     MFGR#1213
1539587552       1998     MFGR#1214
1757204380       1998     MFGR#1215
1652822919       1998     MFGR#1216
1584019774       1998     MFGR#1217
1522581193       1998     MFGR#1218
1406153892       1998     MFGR#1219
1608848579       1998     MFGR#122
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: 200.506 seconds, Fetched: 280 row(s)
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

c. Scale14: cluster of at least 4 nodes

Scale14 data: 4 node cluster: Query 2.1 – Time Taken: 404.999 seconds.

```
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_20161122201202_ddf1b1cb-9be4-43b7-a4fe-602e3b7352b8
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_20161122201202_ddf1b1cb-9be4-43b7-a4fe-602e3b7352b8.log
2016-11-22 20:12:08     Starting to launch local task to process map join;      maximum memory = 477102080
2016-11-22 20:12:10     Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/f7ab701e-e319-4890-8b99-d06e9764cf03/hive_2016-11-22_20-12-02_104_284266114270845257
2-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile31--.hashtable
2016-11-22 20:12:10     Uploaded 1 File to: file:/tmp/ec2-user/f7ab701e-e319-4890-8b99-d06e9764cf03/hive_2016-11-22_20-12-02_104_284266114270845257 2-1/-local-10014/HashTable-Stage-13/MapJoin-m
apfile31--.hashtable (67039 bytes)
2016-11-22 20:12:10     End of local task; Time Taken: 1.269 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_1479756709677_0113, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0113/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479756709677_0113
Hadoop job information for Stage-13: number of mappers: 33; number of reducers: 0
2016-11-22 20:12:16,630 Stage-13 map = 0%,  reduce = 0%
2016-11-22 20:12:40,863 Stage-13 map = 2%,  reduce = 0%, Cumulative CPU 18.46 sec
2016-11-22 20:12:51,210 Stage-13 map = 3%,  reduce = 0%, Cumulative CPU 81.85 sec
2016-11-22 20:12:58,489 Stage-13 map = 5%,  reduce = 0%, Cumulative CPU 123.06 sec
2016-11-22 20:12:59,657 Stage-13 map = 6%,  reduce = 0%, Cumulative CPU 131.91 sec
2016-11-22 20:13:00,852 Stage-13 map = 8%,  reduce = 0%, Cumulative CPU 133.62 sec
2016-11-22 20:13:02,043 Stage-13 map = 9%,  reduce = 0%, Cumulative CPU 142.02 sec
2016-11-22 20:13:05,576 Stage-13 map = 14%,  reduce = 0%, Cumulative CPU 161.62 sec
2016-11-22 20:13:06,642 Stage-13 map = 15%,  reduce = 0%, Cumulative CPU 164.33 sec
2016-11-22 20:13:09,065 Stage-13 map = 20%,  reduce = 0%, Cumulative CPU 177.63 sec
2016-11-22 20:13:10,281 Stage-13 map = 26%,  reduce = 0%, Cumulative CPU 183.66 sec
2016-11-22 20:13:11,722 Stage-13 map = 32%,  reduce = 0%, Cumulative CPU 189.76 sec
2016-11-22 20:13:12,883 Stage-13 map = 33%,  reduce = 0%, Cumulative CPU 193.22 sec
```

```
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2016-11-22 20:18:08,300 Stage-4 map = 0%,  reduce = 0%
2016-11-22 20:18:19,710 Stage-4 map = 42%,  reduce = 0%, Cumulative CPU 7.32 sec
2016-11-22 20:18:20,740 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 8.9 sec
2016-11-22 20:18:26,924 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 10.02 sec
MapReduce Total cumulative CPU time: 10 seconds 20 msec
Ended Job = job_1479756709677_0115
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_1479756709677_0116, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0116/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479756709677_0116
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2016-11-22 20:18:34,632 Stage-5 map = 0%,  reduce = 0%
2016-11-22 20:18:39,792 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 0.87 sec
2016-11-22 20:18:46,037 Stage-5 map = 100%,  reduce = 100%, Cumulative CPU 2.11 sec
MapReduce Total cumulative CPU time: 2 seconds 110 msec
Ended Job = job_1479756709677_0116
MapReduce Jobs Launched:
Stage-Stage-13: Map: 33   Cumulative CPU: 367.66 sec   HDFS Read: 8627782542 HDFS Write: 2619104422 SUCCESS
Stage-Stage-2: Map: 10  Reduce: 11   Cumulative CPU: 702.15 sec   HDFS Read: 2688123632 HDFS Write: 124444623 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 10.02 sec   HDFS Read: 124459465 HDFS Write: 10213 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 1   Cumulative CPU: 2.11 sec   HDFS Read: 16013 HDFS Write: 7237 SUCCESS
Total MapReduce CPU Time Spent: 18 minutes 1 seconds 940 msec
OK
9027430868      1992    MFGR#121
9739714581      1992    MFGR#1210
9682495641      1992    MFGR#1211
9171347479      1992    MFGR#1212
9642593033      1992    MFGR#1213
9226335112      1992    MFGR#1214
9505581761      1992    MFGR#1215
9694284779      1992    MFGR#1216
9848923154      1992    MFGR#1217
9531440771      1992    MFGR#1218
9110751316      1992    MFGR#1219
9720306079      1992    MFGR#122
9253136288      1992    MFGR#1220
9670135815      1992    MFGR#1221
9661366467      1992    MFGR#1222
8986616693      1992    MFGR#1223
9748267899      1992    MFGR#1224
9235598918      1992    MFGR#1225
9897511056      1992    MFGR#1226
9109273227      1992    MFGR#1227
9582960045      1992    MFGR#1228
8907338800      1992    MFGR#1229
9213574115      1992    MFGR#123
8678070422      1992    MFGR#1230
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
9588839388      1997    MFGR#1234
9557073621      1997    MFGR#1235
9765483526      1997    MFGR#1236
10013113417     1997    MFGR#1237
8574429084      1997    MFGR#1238
9277894958      1997    MFGR#1239
9146658661      1997    MFGR#124
10169232494     1997    MFGR#1240
9127172958      1997    MFGR#125
8955811927      1997    MFGR#126
9438467866      1997    MFGR#127
9819102249      1997    MFGR#128
8173997248      1997    MFGR#129
5312699990      1998    MFGR#121
5369622794      1998    MFGR#1210
5554018966      1998    MFGR#1211
5410120237      1998    MFGR#1212
5529705284      1998    MFGR#1213
5444396951      1998    MFGR#1214
5653901634      1998    MFGR#1215
5610498490      1998    MFGR#1216
5734037472      1998    MFGR#1217
5455174567      1998    MFGR#1218
5275443856      1998    MFGR#1219
5605010134      1998    MFGR#122
5575812381      1998    MFGR#1220
5801272541      1998    MFGR#1221
5416917212      1998    MFGR#1222
5246997151      1998    MFGR#1223
5512149734      1998    MFGR#1224
5816170896      1998    MFGR#1225
5725113932      1998    MFGR#1226
5197193565      1998    MFGR#1227
5829253768      1998    MFGR#1228
5213215808      1998    MFGR#1229
5469618301      1998    MFGR#123
5352446219      1998    MFGR#1230
5460899614      1998    MFGR#1231
6146536174      1998    MFGR#1232
5451677914      1998    MFGR#1233
5393254485      1998    MFGR#1234
5685166992      1998    MFGR#1235
5674147366      1998    MFGR#1236
5714075486      1998    MFGR#1237
4895622893      1998    MFGR#1238
5445083090      1998    MFGR#1239
5434441542      1998    MFGR#124
5927941083      1998    MFGR#1240
5138521454      1998    MFGR#125
5092624139      1998    MFGR#126
5800124570      1998    MFGR#127
5494514128      1998    MFGR#128
4845028184      1998    MFGR#129
Time taken: 404.999 seconds, Fetched: 280 row(s)
hive>
```

4B. Both of your solutions for 2-B.

a. Scale4: single node and a cluster of at least 4 nodes

Scale4: Single node HIVE query 2B

Disk space available before execution:

```
[ec2-user@ip-172-31-13-121 ~]$ ls -ltr
total 12
drwxr-xr-x 10 ec2-user ec2-user 4096 Oct 14 17:43 hadoop-2.6.4
drwxrwxr-x  2 ec2-user ec2-user 4096 Nov 22 18:18 scale4
drwxrwxr-x 10 ec2-user ec2-user 4096 Nov 22 18:21 apache-hive-2.0.1-bin
[ec2-user@ip-172-31-13-121 ~]$ df -h
Filesystem      Size  Used Avail Use% Mounted on
devtmpfs        992M   56K  992M   1% /dev
tmpfs          1002M     0 1002M   0% /dev/shm
/dev/xvda1      7.8G  3.8G  3.9G  50% /
[ec2-user@ip-172-31-13-121 ~]$
```

Created table: prejoin1 --> preparing to run the Pig script

```
hive> CREATE TABLE prejoin1 AS
    > SELECT lo_partkey,lo_suppkey,d_year,lo_revenue
    > FROM lineorder,dwdate WHERE lo_orderdate=d_datekey;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution e
ngine (i.e. tez, spark) or using Hive 1.X releases.
Query ID = ec2-user_20161122190307_99f67402-a221-44ff-97ba-6d3cb67b76c7
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/StaticLoggerB
inder.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/Stati
cLoggerBinder.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_20161122190307_99f67402-a221-44ff-97ba-6d3cb67b76c7.log
2016-11-22 19:03:15     Starting to launch local task to process map join;      maximum memory = 518979584
2016-11-22 19:03:17     Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/c49a5ce8-d8b7-4a60-904d
-f3aa64fd650f/hive_2016-11-22_19-03-07_040_8007330942086178483-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile01--.hashtable
2016-11-22 19:03:17     Uploaded 1 File to: file:/tmp/ec2-user/c49a5ce8-d8b7-4a60-904d-f3aa64fd650f/hive_2016-11-22_19-03-07_040_80
07330942086178483-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile01--.hashtable (67039 bytes)
2016-11-22 19:03:17     End of local task; Time Taken: 1.751 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_1479841001015_0001, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_147
9841001015_0001/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479841001015_0001
Hadoop job information for Stage-4: number of mappers: 10; number of reducers: 0
2016-11-22 19:03:29,612 Stage-4 map = 0%,  reduce = 0%
2016-11-22 19:04:29,931 Stage-4 map = 0%,  reduce = 0%, Cumulative CPU 25.84 sec
2016-11-22 19:04:49,370 Stage-4 map = 30%,  reduce = 0%, Cumulative CPU 44.88 sec
2016-11-22 19:05:12,383 Stage-4 map = 35%,  reduce = 0%, Cumulative CPU 64.39 sec
2016-11-22 19:05:33,484 Stage-4 map = 60%,  reduce = 0%, Cumulative CPU 66.28 sec
2016-11-22 19:05:51,486 Stage-4 map = 70%,  reduce = 0%, Cumulative CPU 77.0 sec
2016-11-22 19:06:19,271 Stage-4 map = 75%,  reduce = 0%, Cumulative CPU 90.42 sec
2016-11-22 19:06:20,527 Stage-4 map = 85%,  reduce = 0%, Cumulative CPU 91.55 sec
2016-11-22 19:06:47,118 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 102.26 sec
MapReduce Total cumulative CPU time: 1 minutes 42 seconds 260 msec
Ended Job = job_1479841001015_0001
Moving data to: hdfs://localhost/user/hive/warehouse/prejoin1
MapReduce Jobs Launched:
Stage-Stage-4: Map: 10   Cumulative CPU: 102.26 sec   HDFS Read: 2417934674 HDFS Write: 607282429 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 42 seconds 260 msec
OK
Time taken: 222.455 seconds
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Tried to run Query 2B several times with scale4 data but each time the process got killed –

```
hive> select sum(lo_revenue), d_year, p_brand1
    > from prejoin1, part, supplier
    > where 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_20161122195300_3249b990-494c-4ad7-8daa-be07b4aac21a
Total jobs = 3
Stage-1 is selected by condition resolver.
Launching Job 1 out of 3
Number of reduce tasks not specified. Estimated from input data size: 3
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_1479843621488_0002, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1479843621488_0002/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479843621488_0002
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 3
2016-11-22 19:53:09,262 Stage-1 map = 0%,  reduce = 0%
2016-11-22 19:53:46,025 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 14.58 sec
Killed
[ec2-user@ip-172-31-13-121 apache-hive-2.0.1-bin]$
```

Disk space after execution:

```
[ec2-user@ip-172-31-13-121 ~]$ df -h
Filesystem      Size  Used Avail Use% Mounted on
devtmpfs        992M   56K  992M   1% /dev
tmpfs          1002M     0 1002M   0% /dev/shm
/dev/xvda1      7.8G  7.6G  141M  99% /
[ec2-user@ip-172-31-13-121 ~]$
```

Scale4: single node PIG for query 2B --> Resulted in Container launch exception and failed to create table prejoin1

```
2016-11-23 00:06:46,985 [main] INFO  org.apache.hadoop.ipc.Client - Retryin2016-11-23 00:06:46,985 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: ip-172-31-13-121.us-west-2.com
pute.internal/172.31.13.121:35227. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2016-11-23 00:06:47,985 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: ip-172-31-13-121.us-west-2.compute.internal/172.31.13.121:35227. Already tried 2 time(s); retry policy is
 RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2016-11-23 00:07:10,282 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1479843621488_0004]
2016-11-23 00:07:13,296 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Ooops! Some job has failed! Specify -stop_on_failure if you want Pig to stop immedi
ately on failure.
2016-11-23 00:07:13,296 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - job job_1479843621488_0004 has failed! Stop running all dependent jobs
2016-11-23 00:07:13,296 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-23 00:07:13,297 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-11-23 00:07:13,306 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=FAILED. Redirecting to job history server
2016-11-23 00:07:13,685 [main] ERROR org.apache.pig.tools.pigstats.PigStats - ERROR 0: org.apache.pig.backend.executionengine.ExecException: ERROR 2997: Unable to recreate exception from backed error:
Container launch failed for container_1479843621488_0004_02_000005 : org.apache.hadoop.security.token.SecretManager$InvalidToken: No NMToken sent for ip-172-31-13-121.us-west-2.compute.internal:38011
        at org.apache.hadoop.yarn.client.api.impl.ContainerManagementProtocolProxy$ContainerManagementProtocolProxyData.newProxy(ContainerManagementProtocolProxy.java:254)
        at org.apache.hadoop.yarn.client.api.impl.ContainerManagementProtocolProxy$ContainerManagementProtocolProxyData.<init>(ContainerManagementProtocolProxy.java:244)
        at org.apache.hadoop.yarn.client.api.impl.ContainerManagementProtocolProxy.getProxy(ContainerManagementProtocolProxy.java:129)
        at org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl.getCMProxy(ContainerLauncherImpl.java:403)
        at org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl$Container.launch(ContainerLauncherImpl.java:138)
        at org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl$EventProcessor.run(ContainerLauncherImpl.java:369)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
        at java.lang.Thread.run(Thread.java:745)

2016-11-23 00:07:13,685 [main] ERROR org.apache.pig.tools.pigstats.mapreduce.MRPigStatsUtil - 1 map reduce job(s) failed!
2016-11-23 00:07:13,687 [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        2016-11-23 00:05:05     2016-11-23 00:07:13     HASH_JOIN

Failed!

Failed Jobs:
JobId   Alias   Feature Message Outputs
job_1479843621488_0004  dwdate,join_date,lineorder,prejoin1     HASH_JOIN       Message: Job failed!    /user/ec2-user/prejoin1,

Input(s):
Failed to read data from "/user/ec2-user/scale4/dwdate.tbl"
Failed to read data from "/user/ec2-user/scale4/lineorder.tbl"

Output(s):
Failed to produce result in "/user/ec2-user/prejoin1"

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

Job DAG:
job_1479843621488_0004


2016-11-23 00:07:13,687 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Failed!
grunt>
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
[ec2-user@ip-172-31-13-121 ~]$ df -h
Filesystem      Size  Used Avail Use% Mounted on
devtmpfs        992M   56K  992M   1% /dev
tmpfs          1002M     0 1002M   0% /dev/shm
/dev/xvda1      7.8G  4.0G  3.8G  52% /
[ec2-user@ip-172-31-13-121 ~]$
```

```
[ec2-user@ip-172-31-13-121 ~]$ cat pig_1479859430471.log
Backend error message
---------------------
Container launch failed for container_1479843621488_0004_02_000011 :
org.apache.hadoop.security.token.SecretManager$InvalidToken: No NMToken sent for ip-172-31-13-121.us-west-2.compute.internal:38011
        at org.apache.hadoop.yarn.client.api.impl.ContainerManagementProtocolProxy$ContainerManagementProtocolProxyData.newProxy(ContainerManagementProtocolProxy.java:254)
        at org.apache.hadoop.yarn.client.api.impl.ContainerManagementProtocolProxy$ContainerManagementProtocolProxyData.<init>(ContainerManagementProtocolProxy.java:244)
        at org.apache.hadoop.yarn.client.api.impl.ContainerManagementProtocolProxy.getProxy(ContainerManagementProtocolProxy.java:129)
        at org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl.getCMProxy(ContainerLauncherImpl.java:403)
        at org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl$Container.launch(ContainerLauncherImpl.java:138)
        at org.apache.hadoop.mapreduce.v2.app.launcher.ContainerLauncherImpl$EventProcessor.run(ContainerLauncherImpl.java:369)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
        at java.lang.Thread.run(Thread.java:745)
```

On a single node, none of my queries worked successfully with scale4 data. There were performance and memory issues.

Scale 4: Cluster of 4 nodes- HIVE

Time taken: 68.864 seconds.

```
hive> select sum(lo_revenue), d_year, p_brand1
    > from prejoin1, part, supplier
    > where 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_20161120063554_11e03762-92e3-403f-bdfd-8543ef8e8084
Total jobs = 4
Stage-12 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_20161120063554_11e03762-92e3-403f-bdfd-8543ef8e8084.log
2016-11-20 06:35:59     Starting to launch local task to process map join;      maximum memory = 477102080
2016-11-20 06:36:01     Dump the side-table for tag: 1 with group count: 7883 into file: file:/tmp/ec2-user/c433b20e-db07-4e47-ac7e-a05fbe4954d7/hive_2016-11-20_06-35-54_344_4361450023980948872-1/-local-10009/HashTable-Stage-8/MapJoin-mapfile51--.hashtable
2016-11-20 06:36:01     Uploaded 1 File to: file:/tmp/ec2-user/c433b20e-db07-4e47-ac7e-a05fbe4954d7/hive_2016-11-20_06-35-54_344_4361450023980948872-1/-local-10009/HashTable-Stage-8/MapJoin-mapfile51--.h
ashtable (249337 bytes)
2016-11-20 06:36:01     End of local task; Time Taken: 1.949 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_1479510364858_0031, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479510364858_0031/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479510364858_0031
Hadoop job information for Stage-8: number of mappers: 1; number of reducers: 0
2016-11-20 06:36:06,818 Stage-8 map = 0%,  reduce = 0%
2016-11-20 06:36:17,113 Stage-8 map = 45%,  reduce = 0%, Cumulative CPU 7.36 sec
2016-11-20 06:36:20,214 Stage-8 map = 100%,  reduce = 0%, Cumulative CPU 9.56 sec
MapReduce Total cumulative CPU time: 9 seconds 560 msec
Ended Job = job_1479510364858_0031
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_20161120063554_11e03762-92e3-403f-bdfd-8543ef8e8084.log
2016-11-20 06:36:25     Starting to launch local task to process map join;      maximum memory = 477102080
2016-11-20 06:36:27     Dump the side-table for tag: 1 with group count: 378 into file: file:/tmp/ec2-user/c433b20e-db07-4e47-ac7e-a05fbe4954d7/hive_2016-11-20_06-35-54_344_4361450023980948872-1/-local-1
0007/HashTable-Stage-3/MapJoin-mapfile41--.hashtable
2016-11-20 06:36:27     Uploaded 1 File to: file:/tmp/ec2-user/c433b20e-db07-4e47-ac7e-a05fbe4954d7/hive_2016-11-20_06-35-54_344_4361450023980948872-1/-local-10007/HashTable-Stage-3/MapJoin-mapfile41--.h
ashtable (7792 bytes)
2016-11-20 06:36:27     End of local task; Time Taken: 1.161 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):
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
479099365         1997     MFGR#125
570696568         1997     MFGR#126
583074592         1997     MFGR#127
695133104         1997     MFGR#128
655638776         1997     MFGR#129
344575925         1998     MFGR#121
417152416         1998     MFGR#1210
317068168         1998     MFGR#1211
374341516         1998     MFGR#1212
332740903         1998     MFGR#1213
304873002         1998     MFGR#1214
366101132         1998     MFGR#1215
379133898         1998     MFGR#1216
359508497         1998     MFGR#1217
320623334         1998     MFGR#1218
346182862         1998     MFGR#1219
312440027         1998     MFGR#122
348123961         1998     MFGR#1220
339845398         1998     MFGR#1221
355416161         1998     MFGR#1222
344889822         1998     MFGR#1223
396906691         1998     MFGR#1224
290208878         1998     MFGR#1225
419415707         1998     MFGR#1226
358466340         1998     MFGR#1227
251549955         1998     MFGR#1228
383138860         1998     MFGR#1229
296330561         1998     MFGR#123
437181243         1998     MFGR#1230
398944492         1998     MFGR#1231
424062455         1998     MFGR#1232
406967188         1998     MFGR#1233
428867240         1998     MFGR#1234
352277781         1998     MFGR#1235
361827086         1998     MFGR#1236
341618569         1998     MFGR#1237
244739231         1998     MFGR#1238
414151803         1998     MFGR#1239
330082371         1998     MFGR#124
415312453         1998     MFGR#1240
360289624         1998     MFGR#125
341657580         1998     MFGR#126
377507061         1998     MFGR#127
361416497         1998     MFGR#128
318769573         1998     MFGR#129
Time taken: 68.864 seconds, Fetched: 280 row(s)
hive>
```

Scale 4: cluster of at least 4 nodes PIG

Pig script completed in 3 minutes, 48 seconds and 376 ms.

```
[ec2-user@ip-172-31-46-250 ~]$ $PIG_HOME/bin/pig -f Part2B_pig.pig
16/11/22 23:37:39 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/22 23:37:39 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/11/22 23:37:39 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-11-22 23:37:39,367 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-11-22 23:37:39,367 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig_1479857859364.log
2016-11-22 23:37:40,011 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-11-22 23:37:40,120 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-11-22 23:37:40,120 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:40,120 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.31.46.250/
2016-11-22 23:37:41,016 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
prejoin1: {lo_partkey: int,lo_suppkey: int,d_year: int,lo_revenue: int}
2016-11-22 23:37:41,130 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
part: {p_partkey: int,p_name: chararray,p_mfgr: chararray,p_category: chararray,p_brand1: chararray,p_color: chararray,p_type: chararray,p_size: int,p_container: chararray}
2016-11-22 23:37:41,323 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
supplier: {s_suppkey: int,s_name: chararray,s_address: chararray,s_city: chararray,s_nation: chararray,s_region: chararray,s_phone: chararray}
2016-11-22 23:37:41,564 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:41,588 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2016-11-22 23:37:41,639 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,ORDER_BY,FILTER
2016-11-22 23:37:41,660 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:41,662 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-22 23:37:41,690 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitO
ptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-22 23:37:41,800 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-22 23:37:41,832 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2016-11-22 23:37:41,846 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-131
2016-11-22 23:37:41,850 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 23:37:41,850 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-22 23:37:41,857 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 5
2016-11-22 23:37:41,857 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 5
2016-11-22 23:37:41,879 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 23:37:41,909 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032


2016-11-22 23:41:27,189 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,192 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,239 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,242 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,265 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,268 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,293 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,296 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,320 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,324 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,346 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,349 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,374 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,377 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,403 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,406 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,428 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,432 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,455 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,458 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,479 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,483 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,504 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,507 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,547 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,559 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,580 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,584 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,606 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-22 23:41:27,610 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-22 23:41:27,632 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-11-22 23:41:27,648 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 48 seconds and 376 milliseconds (228376 ms)
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Query 2.1 with prejoin1 -- output file:

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/Part2B_pig
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-22 23:41 /user/ec2-user/Part2B_pig/_SUCCESS
-rw-r--r--   2 ec2-user supergroup       7217 2016-11-22 23:41 /user/ec2-user/Part2B_pig/part-r-00000
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/Part2B_pig/part-r-00000
1207004714|1998|MFGR#129
1285428693|1998|MFGR#1238
1350601347|1998|MFGR#1229
1358086103|1998|MFGR#1222
1383876006|1998|MFGR#1213
1406153892|1998|MFGR#1219
1409115372|1998|MFGR#1223
1441898473|1998|MFGR#1230
1445039464|1998|MFGR#1231
1451906941|1998|MFGR#1225
1457307800|1998|MFGR#1220
1459545128|1998|MFGR#1239
1466946762|1998|MFGR#126
1470503353|1998|MFGR#123
1477715730|1998|MFGR#125
1508262019|1998|MFGR#1210
1509229650|1998|MFGR#1212
1522581193|1998|MFGR#1218
1525737275|1998|MFGR#124
1530901152|1998|MFGR#1221
1532309319|1998|MFGR#1234
1532687418|1998|MFGR#1237
1538644707|1998|MFGR#128
1538979218|1998|MFGR#1233
1539587552|1998|MFGR#1214
1565657860|1998|MFGR#1227
1568652498|1998|MFGR#1224
1577658136|1998|MFGR#1236
1584019774|1998|MFGR#1217
1587370161|1998|MFGR#1240
```

b. Scale14: cluster of at least 4 nodes

Scale14: 4 node cluster: HIVE Query 2B: Time taken 255.912 seconds.

```
hive> select sum(lo_revenue), d_year, p_brand1
    > from prejoin1, part, supplier
    > where 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 rel
eases.
Query ID = ec2-user_20161122203349_cda6d1fa-e45d-4296-b139-bd662ff58b74
Total jobs = 3
Stage-1 is selected by condition resolver.
Launching Job 1 out of 3
Number of reduce tasks not specified. Estimated from input data size: 9
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_1479756709677_0118, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0118/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479756709677_0118
Hadoop job information for Stage-1: number of mappers: 10; number of reducers: 9
2016-11-22 20:33:55,960 Stage-1 map = 0%,  reduce = 0%
2016-11-22 20:34:19,878 Stage-1 map = 7%,  reduce = 0%, Cumulative CPU 41.7 sec
2016-11-22 20:34:22,095 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 52.4 sec
2016-11-22 20:34:35,262 Stage-1 map = 20%,  reduce = 1%, Cumulative CPU 82.59 sec
2016-11-22 20:34:37,430 Stage-1 map = 23%,  reduce = 1%, Cumulative CPU 88.52 sec
2016-11-22 20:34:38,513 Stage-1 map = 23%,  reduce = 4%, Cumulative CPU 89.42 sec
2016-11-22 20:34:40,669 Stage-1 map = 25%,  reduce = 4%, Cumulative CPU 95.9 sec
2016-11-22 20:34:43,914 Stage-1 map = 27%,  reduce = 4%, Cumulative CPU 107.66 sec
2016-11-22 20:34:45,020 Stage-1 map = 27%,  reduce = 5%, Cumulative CPU 107.87 sec
2016-11-22 20:34:46,174 Stage-1 map = 27%,  reduce = 6%, Cumulative CPU 112.05 sec
2016-11-22 20:34:47,299 Stage-1 map = 28%,  reduce = 7%, Cumulative CPU 118.95 sec
2016-11-22 20:34:49,485 Stage-1 map = 30%,  reduce = 7%, Cumulative CPU 124.99 sec
2016-11-22 20:34:50,558 Stage-1 map = 33%,  reduce = 7%, Cumulative CPU 129.45 sec
2016-11-22 20:35:01,383 Stage-1 map = 35%,  reduce = 7%, Cumulative CPU 171.97 sec
2016-11-22 20:35:04,669 Stage-1 map = 42%,  reduce = 7%, Cumulative CPU 184.75 sec
2016-11-22 20:35:07,984 Stage-1 map = 47%,  reduce = 7%, Cumulative CPU 193.47 sec
2016-11-22 20:35:11,300 Stage-1 map = 48%,  reduce = 7%, Cumulative CPU 212.01 sec
2016-11-22 20:35:25,185 Stage-1 map = 50%,  reduce = 7%, Cumulative CPU 256.81 sec
2016-11-22 20:35:26,250 Stage-1 map = 57%,  reduce = 7%, Cumulative CPU 265.94 sec
2016-11-22 20:35:29,505 Stage-1 map = 62%,  reduce = 7%, Cumulative CPU 276.36 sec
2016-11-22 20:35:44,591 Stage-1 map = 66%,  reduce = 7%, Cumulative CPU 328.69 sec
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

```
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2016-11-22 20:37:28,814 Stage-3 map = 0%,  reduce = 0%
2016-11-22 20:37:39,111 Stage-3 map = 45%,  reduce = 0%, Cumulative CPU 7.21 sec
2016-11-22 20:37:40,139 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 8.77 sec
2016-11-22 20:37:46,315 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 9.88 sec
MapReduce Total cumulative CPU time: 9 seconds 880 msec
Ended Job = job_1479756709677_0119
Launching Job 3 out of 3
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_1479756709677_0120, Tracking URL = http://ip-172-31-46-250.us-west-2.compute.internal:8088/proxy/application_1479756709677_0120/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1479756709677_0120
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2016-11-22 20:37:53,078 Stage-4 map = 0%,  reduce = 0%
2016-11-22 20:37:58,226 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 0.85 sec
2016-11-22 20:38:04,401 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 2.0 sec
MapReduce Total cumulative CPU time: 2 seconds 0 msec
Ended Job = job_1479756709677_0120
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Reduce: 9   Cumulative CPU: 642.02 sec   HDFS Read: 2192296743 HDFS Write: 124444451 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 9.88 sec   HDFS Read: 124458693 HDFS Write: 10213 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 2.0 sec   HDFS Read: 16004 HDFS Write: 7237 SUCCESS
Total MapReduce CPU Time Spent: 10 minutes 53 seconds 900 msec
OK
9027430868      1992    MFGR#121
9739714581      1992    MFGR#1210
9682495641      1992    MFGR#1211
9171347479      1992    MFGR#1212
9642593033      1992    MFGR#1213
9226335112      1992    MFGR#1214
9505581761      1992    MFGR#1215
9694284779      1992    MFGR#1216
9848923154      1992    MFGR#1217
9531440771      1992    MFGR#1218
9110751316      1992    MFGR#1219
9720306079      1992    MFGR#122
```

```
5444396951      1998    MFGR#1214
5653901634      1998    MFGR#1215
5610498490      1998    MFGR#1216
5734037472      1998    MFGR#1217
5451174567      1998    MFGR#1218
5275443856      1998    MFGR#1219
5605010134      1998    MFGR#122
5575812381      1998    MFGR#1220
5801272541      1998    MFGR#1221
5416917212      1998    MFGR#1222
5246997151      1998    MFGR#1223
5512149734      1998    MFGR#1224
5816170896      1998    MFGR#1225
5725113932      1998    MFGR#1226
5197193565      1998    MFGR#1227
5829253768      1998    MFGR#1228
5213215808      1998    MFGR#1229
5469618301      1998    MFGR#123
5352446219      1998    MFGR#1230
5460899614      1998    MFGR#1231
6146536174      1998    MFGR#1232
5451677914      1998    MFGR#1233
5393254485      1998    MFGR#1234
5685166992      1998    MFGR#1235
5674147366      1998    MFGR#1236
5714075486      1998    MFGR#1237
4895622893      1998    MFGR#1238
5445083090      1998    MFGR#1239
5434441542      1998    MFGR#124
5927941083      1998    MFGR#1240
5138521454      1998    MFGR#125
5092624139      1998    MFGR#126
5800124570      1998    MFGR#127
5494514128      1998    MFGR#128
4845028184      1998    MFGR#129
Time taken: 255.912 seconds, Fetched: 280 row(s)
```

Brunda Chouthoy
CSC 555: Mining Big Data
Project Phase-2

Scale14: 4 node cluster: PIG Query 2B:

Pig script for scale14 data completed in 6 minutes, 3 seconds and 774 ms

```
[ec2-user@ip-172-31-46-250 ~]$ $PIG_HOME/bin/pig -f Part2B_pig_scale14.pig
16/11/23 01:53:54 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/23 01:53:54 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/11/23 01:53:54 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-11-23 01:53:54,783 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-11-23 01:53:54,783 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig_1479866034778.log
2016-11-23 01:53:55,410 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-11-23 01:53:55,522 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-11-23 01:53:55,522 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-23 01:53:55,522 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.31.46.250/
2016-11-23 01:53:56,403 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
prejoin1: {lo_partkey: int,lo_suppkey: int,d_year: int,lo_revenue: int}
2016-11-23 01:53:56,575 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
part: {p_partkey: int,p_name: chararray,p_mfgr: chararray,p_category: chararray,p_brand1: chararray,p_color: chararray,p_type: chararray,p_size: int,p_container: chararray}
2016-11-23 01:53:56,766 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
supplier: {s_suppkey: int,s_name: chararray,s_address: chararray,s_city: chararray,s_nation: chararray,s_region: chararray,s_phone: chararray}
2016-11-23 01:53:56,957 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-23 01:53:56,981 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2016-11-23 01:53:57,019 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,ORDER_BY,FILTER
2016-11-23 01:53:57,039 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-23 01:53:57,041 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-11-23 01:53:57,069 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, Limit
timizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-11-23 01:53:57,179 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-11-23 01:53:57,213 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2016-11-23 01:53:57,226 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-131
2016-11-23 01:53:57,238 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-23 01:53:57,230 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-11-23 01:53:57,237 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 5
2016-11-23 01:53:57,237 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 5
```

```
2016-11-23 01:59:57,955 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-23 01:59:57,977 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-11-23 01:59:58,007 [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        2016-11-23 01:53:57     2016-11-23 01:59:58     HASH_JOIN,GROUP_BY,ORDER_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1479756709677_0143  19      3       143     20      114     123     188     179     183     181     filtereddata,join1,part,prejoin1        HASH_JOIN
job_1479756709677_0144  4       1       27      11      22      26      27      27      27      27      join2,supplier  HASH_JOIN
job_1479756709677_0145  1       1       8       8       8       8       3       3       3       3       groupeddata,query21     GROUP_BY,COMBINER
job_1479756709677_0146  1       1       3       3       3       3       3       3       3       3       ordered_result  SAMPLER
job_1479756709677_0147  1       1       3       3       3       3       3       3       3       3       ordered_result  ORDER_BY        /user/ec2-user/Part2B_pig,

Input(s):
Successfully read 83988094 records from: "/user/ec2-user/prejoin1"
Successfully read 800000 records from: "/user/ec2-user/scale14/part.tbl"
Successfully read 28000 records from: "/user/ec2-user/scale14/supplier.tbl"

Output(s):
Successfully stored 280 records (7237 bytes) in: "/user/ec2-user/Part2B_pig"

Counters:
Total records written : 280
Total bytes written : 7237
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
2016-11-23 01:59:58,397 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-23 01:59:58,400 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-23 01:59:58,424 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /172.31.46.250:8032
2016-11-23 01:59:58,427 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-11-23 01:59:58,449 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-11-23 01:59:58,466 [main] INFO  org.apache.pig.Main - Pig script completed in 6 minutes, 3 seconds and 774 milliseconds (363774 ms)
[ec2-user@ip-172-31-46-250 ~]$
```

Output file -->

```
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -ls /user/ec2-user/Part2B_pig
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2016-11-23 01:59 /user/ec2-user/Part2B_pig/_SUCCESS
-rw-r--r--   2 ec2-user supergroup       7237 2016-11-23 01:59 /user/ec2-user/Part2B_pig/part-r-00000
[ec2-user@ip-172-31-46-250 ~]$ hadoop fs -cat /user/ec2-user/Part2B_pig/part-r-00000
4845028184|1998|MFGR#129
4895622893|1998|MFGR#1238
5092624139|1998|MFGR#126
5138521454|1998|MFGR#125
5197193565|1998|MFGR#1227
5213215808|1998|MFGR#1229
5246997151|1998|MFGR#1223
5275443856|1998|MFGR#1219
5312699990|1998|MFGR#121
5352446219|1998|MFGR#1230
5369622794|1998|MFGR#1210
5393254485|1998|MFGR#1234
5410120237|1998|MFGR#1212
5416917212|1998|MFGR#1222
5434441542|1998|MFGR#124
5444396951|1998|MFGR#1214
5445083090|1998|MFGR#1239
5451174567|1998|MFGR#1218
5451677914|1998|MFGR#1233
5460899614|1998|MFGR#1231
5469618301|1998|MFGR#123
5494514128|1998|MFGR#128
5512149734|1998|MFGR#1224
5529705284|1998|MFGR#1213
5554018966|1998|MFGR#1211
5575812381|1998|MFGR#1220
5605010134|1998|MFGR#122
5610498490|1998|MFGR#1216
5653901634|1998|MFGR#1215
```