Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

Part 1: Hive

Table creations:

```
create table part (
  p_partkey     int,
  p_name        varchar(22),
  p_mfgr        varchar(6),
  p_category    varchar(7),
  p_brand1      varchar(9),
  p_color       varchar(11),
  p_type        varchar(25),
  p_size        int,
  p_container   varchar(10)
)
ROW FORMAT DELIMITED FIELDS
TERMINATED BY '|' STORED AS TEXTFILE;

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

create table supplier (
  s_suppkey     int,
  s_name        varchar(25),
  s_address     varchar(25),
  s_city        varchar(10),
  s_nation      varchar(15),
  s_region      varchar(12),
  s_phone       varchar(15)
)
ROW FORMAT DELIMITED FIELDS
TERMINATED BY '|' STORED AS TEXTFILE;

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

create table customer (
  c_custkey     int,
  c_name        varchar(25),
  c_address     varchar(25),
  c_city        varchar(10),
  c_nation      varchar(15),
  c_region      varchar(12),
  c_phone       varchar(15),
  c_mktsegment  varchar(10)
)
ROW FORMAT DELIMITED FIELDS
TERMINATED BY '|' STORED AS TEXTFILE;

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

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
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;

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;
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

SQL queries:

--Q0.1 Added simple test query
SELECT SUM(lo_revenue)
FROM lineorder;

```
hive> SELECT SUM(lo_revenue)
    > 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_20161026010812_c6244fba-8cc6-4b4c-8e8e-0b6f2cc598f4
Total jobs = 1
Launching Job 1 out of 1
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_1477443484335_0001, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0001/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0001
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2016-10-26 01:08:26,807 Stage-1 map = 0%,  reduce = 0%
2016-10-26 01:08:50,588 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 5.81 sec
2016-10-26 01:08:55,214 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 9.47 sec
2016-10-26 01:08:58,759 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.63 sec
2016-10-26 01:09:03,076 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.81 sec
MapReduce Total cumulative CPU time: 11 seconds 810 msec
Ended Job = job_1477443484335_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 1   Cumulative CPU: 11.81 sec   HDFS Read: 594353315 HDFS Write: 15 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 810 msec
OK
21810219932448
Time taken: 51.174 seconds, Fetched: 1 row(s)
hive>
```

--Q0.2 Added simple test query
SELECT lo_discount, COUNT(lo_extendedprice)
FROM lineorder
GROUP BY lo_discount;

```
hive> SELECT lo_discount, COUNT(lo_extendedprice)
    > FROM lineorder
    > GROUP BY lo_discount;
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_20161026011259_e89d9122-a0a3-4dd4-ad94-b1aab10279ce
Total jobs = 1
Launching Job 1 out of 1
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_1477443484335_0002, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0002/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0002
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3
2016-10-26 01:13:08,211 Stage-1 map = 0%,  reduce = 0%
2016-10-26 01:13:31,266 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 6.01 sec
2016-10-26 01:13:35,772 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 10.11 sec
2016-10-26 01:13:40,761 Stage-1 map = 78%,  reduce = 0%, Cumulative CPU 11.08 sec
2016-10-26 01:13:41,987 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.3 sec
2016-10-26 01:13:52,491 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 12.53 sec
2016-10-26 01:13:53,601 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 13.76 sec
2016-10-26 01:13:54,651 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 14.99 sec
MapReduce Total cumulative CPU time: 14 seconds 990 msec
Ended Job = job_1477443484335_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 14.99 sec   HDFS Read: 594365119 HDFS Write: 100 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 990 msec
OK
0       544886
3       545293
6       544970
9       545309
1       545834
4       545545
7       546192
10      545815
2       546173
5       546395
8       544803
Time taken: 55.729 seconds, Fetched: 11 row(s)
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

--Q0.3 Added simple test query
SELECT lo_quantity, SUM(lo_revenue)
FROM lineorder
WHERE lo_discount > 2
GROUP BY lo_quantity;

```
hive> SELECT lo_quantity, SUM(lo_revenue)
    > FROM lineorder
    > WHERE lo_discount > 2
    > GROUP BY lo_quantity;
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_20161026011418_92463089-7186-4d2f-b35c-36406a801028
Total jobs = 1
Launching Job 1 out of 1
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_1477443484335_0003, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0003/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0003
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3
2016-10-26 01:14:26,450 Stage-1 map = 0%,  reduce = 0%
2016-10-26 01:14:51,679 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 8.63 sec
2016-10-26 01:14:55,055 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 10.88 sec
2016-10-26 01:15:03,907 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.19 sec
2016-10-26 01:15:14,243 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 14.58 sec
2016-10-26 01:15:15,274 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.24 sec
MapReduce Total cumulative CPU time: 17 seconds 240 msec
Ended Job = job_1477443484335_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 17.24 sec   HDFS Read: 594367372 HDFS Write: 783 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 240 msec
OK
3       36741426823
6       73051794402
9       110995371932
12      146618690455
15      183716500772
18      219285225539
21      257940008985
24      293515134089
27      329860583929
30      366267941974
33      403069114917
36      442350607436
39      478088120865
42      515015195570
```

```
Total MapReduce CPU Time Spent: 17 seconds 240 msec
OK
3       36741426823
6       73051794402
9       110995371932
12      146618690455
15      183716500772
18      219285225539
21      257940008985
24      293515134089
27      329860583929
30      366267941974
33      403069114917
36      442350607436
39      478088120865
42      515015195570
45      549226253763
48      588408783728
1       12270917013
4       48774169472
7       85856252908
10      121796006919
13      158899514794
16      196359425660
19      232654870670
22      268096778479
25      306929186636
28      344003246981
31      379551414675
34      416535036839
37      454316923501
40      487839190553
43      525590086101
46      564234885296
49      599469056351
2       24370723319
5       61026172645
8       98209492651
11      134166410947
14      170856792200
17      208300748323
20      245014170435
23      282074197891
26      316879287165
29      353872128807
32      392471794497
35      431730728563
38      464266209209
41      504472343691
44      539140257365
47      574869730412
50      610179291714
Time taken: 57.981 seconds, Fetched: 50 row(s)
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

--Q1.1 Simplified to remove expression in sum

```
select sum(lo_extendedprice) as revenue
from lineorder, dwdate
where lo_orderdate = d_datekey
  and d_year = 1993
  and lo_discount between 1 and 3
  and lo_quantity < 25;
```

```
hive> select sum(lo_extendedprice) as revenue
    > from lineorder, dwdate
    > where lo_orderdate = d_datekey
    >   and d_year = 1993
    >   and lo_discount between 1 and 3
    >   and lo_quantity < 25;
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_20161026011712_385f1173-270e-4a97-82c8-fa30979062b4
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_20161026011712_385f1173-270e-4a97-82c8-fa30979062b4.log
2016-10-26 01:17:20     Starting to launch local task to process map join;      maximum memory = 518979584
2016-10-26 01:17:22     Dump the side-table for tag: 1 with group count: 365 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-17-12_992_4506385675677195886-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile01--.hashtable
2016-10-26 01:17:22     Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-17-12_992_4506385675677195886-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile01--.hashtable (8328 bytes)
2016-10-26 01:17:22     End of local task; Time Taken: 1.594 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
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_1477443484335_0004, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0004/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0004
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2016-10-26 01:17:30,785 Stage-2 map = 0%,  reduce = 0%
2016-10-26 01:17:55,906 Stage-2 map = 33%,  reduce = 0%, Cumulative CPU 9.01 sec
2016-10-26 01:17:58,862 Stage-2 map = 44%,  reduce = 0%, Cumulative CPU 11.28 sec
2016-10-26 01:18:01,459 Stage-2 map = 56%,  reduce = 0%, Cumulative CPU 12.54 sec
2016-10-26 01:18:07,427 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 14.41 sec
2016-10-26 01:18:10,639 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 15.77 sec
MapReduce Total cumulative CPU time: 15 seconds 770 msec
Ended Job = job_1477443484335_0004
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3  Reduce: 1   Cumulative CPU: 15.77 sec   HDFS Read: 594368484 HDFS Write: 13 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 770 msec
OK
222914249488
Time taken: 59.747 seconds, Fetched: 1 row(s)
hive>
```

--Q1.2 Simplified to remove expression in sum

```
select sum(lo_extendedprice) as revenue
from lineorder, dwdate
where lo_orderdate = d_datekey
  and d_yearmonth = 'Jan1994'
  and lo_discount between 4 and 6
  and lo_quantity between 26 and 35;
```

```
hive> select sum(lo_extendedprice) as revenue
    > from lineorder, dwdate
    > where lo_orderdate = d_datekey
    >   and d_yearmonth = 'Jan1994'
    >   and lo_discount between 4 and 6
    >   and lo_quantity between 26 and 35;
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_20161026012113_e1e711bf-ae00-4b11-8e56-ca4ca2b447f6
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_20161026012113_e1e711bf-ae00-4b11-8e56-ca4ca2b447f6.log
2016-10-26 01:21:20     Starting to launch local task to process map join;      maximum memory = 518979584
2016-10-26 01:21:22     Dump the side-table for tag: 1 with group count: 31 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-21-13_846_5065180237672833160-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile11--.hashtable
2016-10-26 01:21:22     Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-21-13_846_5065180237672833160-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile11--.hashtable (945 bytes)
2016-10-26 01:21:22     End of local task; Time Taken: 1.713 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
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_1477443484335_0005, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0005/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0005
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2016-10-26 01:21:30,296 Stage-2 map = 0%,  reduce = 0%
2016-10-26 01:21:55,375 Stage-2 map = 33%,  reduce = 0%, Cumulative CPU 8.8 sec
2016-10-26 01:21:58,745 Stage-2 map = 56%,  reduce = 0%, Cumulative CPU 11.08 sec
2016-10-26 01:22:05,771 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 13.61 sec
2016-10-26 01:22:09,989 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 14.93 sec
MapReduce Total cumulative CPU time: 14 seconds 930 msec
Ended Job = job_1477443484335_0005
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3  Reduce: 1   Cumulative CPU: 14.93 sec   HDFS Read: 594368485 HDFS Write: 12 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 930 msec
OK
19713249040
Time taken: 57.216 seconds, Fetched: 1 row(s)
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

--Q1.3 Simplified to remove expression in sum

```
select sum(lo_extendedprice) as revenue
from lineorder, dwdate
where lo_orderdate = d_datekey
  and d_weeknuminyear = 6 and d_year = 1994
  and lo_discount between 5 and 7
  and lo_quantity between 36 and 40;
```

```
hive> select sum(lo_extendedprice) as revenue
    > from lineorder, dwdate
    > where lo_orderdate = d_datekey
    >   and d_weeknuminyear = 6 and d_year = 1994
    >   and lo_discount between 5 and 7
    >   and lo_quantity between 36 and 40;
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_20161026012254_fd0c356f-dfd4-4597-b027-8d53e38fcf96
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_20161026012254_fd0c356f-dfd4-4597-b027-8d53e38fcf96.log
2016-10-26 01:23:01     Starting to launch local task to process map join;      maximum memory = 518979584
2016-10-26 01:23:02     Dump the side-table for tag: 1 with group count: 7 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-22-54_478_
4659840215796937903-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile21--.hashtable
2016-10-26 01:23:02     Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-22-54_478_4659840215796937903-1/-local-10005/HashTab
le-Stage-2/MapJoin-mapfile21--.hashtable (414 bytes)
2016-10-26 01:23:02     End of local task; Time Taken: 1.59 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
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_1477443484335_0006, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0006/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0006
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2016-10-26 01:23:11,586 Stage-2 map = 0%,  reduce = 0%
2016-10-26 01:23:35,456 Stage-2 map = 33%,  reduce = 0%, Cumulative CPU 8.68 sec
2016-10-26 01:23:38,781 Stage-2 map = 56%,  reduce = 0%, Cumulative CPU 10.82 sec
2016-10-26 01:23:43,515 Stage-2 map = 78%,  reduce = 0%, Cumulative CPU 12.18 sec
2016-10-26 01:23:44,616 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 12.5 sec
2016-10-26 01:23:47,739 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 13.79 sec
MapReduce Total cumulative CPU time: 13 seconds 790 msec
Ended Job = job_1477443484335_0006
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3  Reduce: 1   Cumulative CPU: 13.79 sec   HDFS Read: 594368485 HDFS Write: 11 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 790 msec
OK
2832973571
Time taken: 54.322 seconds, Fetched: 1 row(s)
hive>
```

--Q2.1 No simpifications

```
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;
```

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

Q2.1 Output –

```
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'
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
  > 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_20161026012444_4eadd706-19ce-45c5-9421-e612340a90a3
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_20161026012444_4eadd706-19ce-45c5-9421-e612340a90a3.log
2016-10-26 01:24:52 Starting to launch local task to process map join;	maximum memory = 518979584
2016-10-26 01:24:53 Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-24-44_478_5887451017733581525-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile61--.hashtable
2016-10-26 01:24:53 Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-24-44_478_5887451017733581525-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile61--.hashtable (67039 bytes)
2016-10-26 01:24:53 End of local task; Time Taken: 1.7 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_1477443484335_0007, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0007/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0007
Hadoop job information for Stage-13: number of mappers: 3; number of reducers: 0
2016-10-26 01:25:02,595 Stage-13 map = 0%,  reduce = 0%
2016-10-26 01:25:33,769 Stage-13 map = 33%,  reduce = 0%, Cumulative CPU 14.93 sec
2016-10-26 01:25:42,698 Stage-13 map = 67%,  reduce = 0%, Cumulative CPU 20.32 sec
2016-10-26 01:25:59,763 Stage-13 map = 83%,  reduce = 0%, Cumulative CPU 27.62 sec
2016-10-26 01:26:00,790 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 28.03 sec
MapReduce Total cumulative CPU time: 28 seconds 30 msec
Ended Job = job_1477443484335_0007
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_20161026012444_4eadd706-19ce-45c5-9421-e612340a90a3.log
2016-10-26 01:26:08 Starting to launch local task to process map join;	maximum memory = 518979584
2016-10-26 01:26:11 Dump the side-table for tag: 1 with group count: 7883 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-24-44_478_5887451017733581525-1/-local-10010/HashTable-Stage-10/MapJoin-mapfile41--.hashtable
2016-10-26 01:26:11 Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-24-44_478_5887451017733581525-1/-local-10010/HashTable-Stage-10/MapJoin-mapfile41--.hashtable (249337 bytes)
2016-10-26 01:26:11 End of local task; Time Taken: 2.689 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 6
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1477443484335_0008, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0008/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0008
Hadoop job information for Stage-10: number of mappers: 1; number of reducers: 0
2016-10-26 01:26:19,976 Stage-10 map = 0%,  reduce = 0%
2016-10-26 01:26:34,941 Stage-10 map = 45%,  reduce = 0%, Cumulative CPU 8.56 sec
2016-10-26 01:26:38,076 Stage-10 map = 100%,  reduce = 0%, Cumulative CPU 11.72 sec
MapReduce Total cumulative CPU time: 11 seconds 720 msec
Ended Job = job_1477443484335_0008
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_20161026012444_4eadd706-19ce-45c5-9421-e612340a90a3.log
2016-10-26 01:26:45 Starting to launch local task to process map join;	maximum memory = 518979584
2016-10-26 01:26:47 Dump the side-table for tag: 1 with group count: 378 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-24-44_478_5887451017733581525-1/-local-10008/HashTable-Stage-4/MapJoin-mapfile31--.hashtable
2016-10-26 01:26:47 Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-24-44_478_5887451017733581525-1/-local-10008/HashTable-Stage-4/MapJoin-mapfile31--.hashtable (7792 bytes)
2016-10-26 01:26:47 End of local task; Time Taken: 1.631 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 4 out of 6
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1477443484335_0009, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0009/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0009
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2016-10-26 01:26:55,761 Stage-4 map = 0%,  reduce = 0%
2016-10-26 01:27:04,413 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.74 sec
2016-10-26 01:27:12,833 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 3.85 sec
MapReduce Total cumulative CPU time: 3 seconds 850 msec
Ended Job = job_1477443484335_0009
Launching Job 5 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_1477443484335_0010, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0010/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0010
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2016-10-26 01:27:25,896 Stage-5 map = 0%,  reduce = 0%
2016-10-26 01:27:33,312 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 0.92 sec
2016-10-26 01:27:40,730 Stage-5 map = 100%,  reduce = 100%, Cumulative CPU 2.06 sec
MapReduce Total cumulative CPU time: 2 seconds 60 msec
Ended Job = job_1477443484335_0010
MapReduce Jobs Launched:
Stage-Stage-13: Map: 3   Cumulative CPU: 28.03 sec   HDFS Read: 594359098 HDFS Write: 184733291 SUCCESS
Stage-Stage-10: Map: 1   Cumulative CPU: 11.72 sec   HDFS Read: 184739523 HDFS Write: 8750735 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 3.85 sec   HDFS Read: 8762962 HDFS Write: 9913 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 1   Cumulative CPU: 2.06 sec   HDFS Read: 15689 HDFS Write: 6937 SUCCESS
Total MapReduce CPU Time Spent: 45 seconds 660 msec
OK
567838207       1992    MFGR#121
610663790       1992    MFGR#1210
550769662       1992    MFGR#1211
649205856       1992    MFGR#1212
624031241       1992    MFGR#1213
670488468       1992    MFGR#1214
633152470       1992    MFGR#1215
674846781       1992    MFGR#1216
675093435       1992    MFGR#1217
600202070       1992    MFGR#1218
538043594       1992    MFGR#1219
655326672       1992    MFGR#122
540262882       1992    MFGR#1220
556120633       1992    MFGR#1221
590762777       1992    MFGR#1222
535448651       1992    MFGR#1223
703752611       1992    MFGR#1224
570832868       1992    MFGR#1225
614061593       1992    MFGR#1226
581759388       1992    MFGR#1227
644642592       1992    MFGR#1228
640858430       1992    MFGR#1229
789755835       1992    MFGR#123
468535087       1992    MFGR#1230
592436656       1992    MFGR#1231
664275152       1992    MFGR#1232
613885100       1992    MFGR#1233
667399281       1992    MFGR#1234
640290070       1992    MFGR#1235
501892561       1992    MFGR#1236
591481503       1992    MFGR#1237
477423770       1992    MFGR#1238
638259374       1992    MFGR#1239
572354196       1992    MFGR#124
740479248       1992    MFGR#1240
478777095       1992    MFGR#125
592174616       1992    MFGR#126
706151632       1992    MFGR#127
542306646       1992    MFGR#128
581987352       1992    MFGR#129
823087702       1993    MFGR#121
648160706       1993    MFGR#1210
634743898       1993    MFGR#1211
785639283       1993    MFGR#1212
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

638255029	1993	MFGR#1213
616837237	1993	MFGR#1214
634687975	1993	MFGR#1215
638353900	1993	MFGR#1216
663372951	1993	MFGR#1217
683985855	1993	MFGR#1218
646950033	1993	MFGR#1219
622532984	1993	MFGR#122
530830127	1993	MFGR#1220
543346337	1993	MFGR#1221
756921203	1993	MFGR#1222
533544350	1993	MFGR#1223
915916085	1993	MFGR#1224
473007381	1993	MFGR#1225
739036124	1993	MFGR#1226
592178887	1993	MFGR#1227
583507058	1993	MFGR#1228
617453491	1993	MFGR#1229
637863868	1993	MFGR#123
625534310	1993	MFGR#1230
580327635	1993	MFGR#1231
697373098	1993	MFGR#1232
515571416	1993	MFGR#1233
651935758	1993	MFGR#1234
575779480	1993	MFGR#1235
591878667	1993	MFGR#1236
609618576	1993	MFGR#1237
444614010	1993	MFGR#1238
595256327	1993	MFGR#1239
660586237	1993	MFGR#124
788730059	1993	MFGR#1240
616224539	1993	MFGR#125
617126754	1993	MFGR#126
654438324	1993	MFGR#127
731657001	1993	MFGR#128
548048395	1993	MFGR#129
564405648	1994	MFGR#121
645404849	1994	MFGR#1210
631620635	1994	MFGR#1211
568332348	1994	MFGR#1212
678785857	1994	MFGR#1213
534002330	1994	MFGR#1214
654400242	1994	MFGR#1215
558646341	1994	MFGR#1216
687845641	1994	MFGR#1217
546674347	1994	MFGR#1218
567272942	1994	MFGR#1219
659884062	1994	MFGR#122
562582172	1994	MFGR#1220
598618997	1994	MFGR#1221
601016441	1994	MFGR#1222
555134404	1994	MFGR#1223
737422302	1994	MFGR#1224
570745955	1994	MFGR#1225
746302245	1994	MFGR#1226
651707481	1994	MFGR#1227
573693547	1994	MFGR#1228
647918373	1994	MFGR#1229
580449592	1994	MFGR#123
493270412	1994	MFGR#1230
603546148	1994	MFGR#1231
719865331	1994	MFGR#1232
638982238	1994	MFGR#1233
743247677	1994	MFGR#1234
598680959	1994	MFGR#1235
615726097	1994	MFGR#1236
542569815	1994	MFGR#1237
573510781	1994	MFGR#1238
579855853	1994	MFGR#1239
684573322	1994	MFGR#124
873735737	1994	MFGR#1240
560488304	1994	MFGR#125
657036514	1994	MFGR#126
622571183	1994	MFGR#127
586845664	1994	MFGR#128
534541525	1994	MFGR#129
706469511	1995	MFGR#121
602892803	1995	MFGR#1210

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

645166092	1995	MFGR#1211
613289283	1995	MFGR#1212
599586479	1995	MFGR#1213
562570804	1995	MFGR#1214
672528755	1995	MFGR#1215
669000972	1995	MFGR#1216
725362449	1995	MFGR#1217
657026635	1995	MFGR#1218
519659003	1995	MFGR#1219
724727741	1995	MFGR#122
517956131	1995	MFGR#1220
635741351	1995	MFGR#1221
564368410	1995	MFGR#1222
600665149	1995	MFGR#1223
762700351	1995	MFGR#1224
671669586	1995	MFGR#1225
572568748	1995	MFGR#1226
530361300	1995	MFGR#1227
633357085	1995	MFGR#1228
547960244	1995	MFGR#1229
660711077	1995	MFGR#123
602735858	1995	MFGR#1230
499852146	1995	MFGR#1231
715300753	1995	MFGR#1232
557149571	1995	MFGR#1233
710023059	1995	MFGR#1234
622425239	1995	MFGR#1235
634565501	1995	MFGR#1236
572847270	1995	MFGR#1237
549318912	1995	MFGR#1238
593851712	1995	MFGR#1239
585421815	1995	MFGR#124
707207888	1995	MFGR#1240
538246872	1995	MFGR#125
605799021	1995	MFGR#126
665978112	1995	MFGR#127
646960956	1995	MFGR#128
508749401	1995	MFGR#129
523879145	1996	MFGR#121
643645053	1996	MFGR#1210
595065339	1996	MFGR#1211
674626440	1996	MFGR#1212
496297087	1996	MFGR#1213
583249505	1996	MFGR#1214
702184857	1996	MFGR#1215
601809334	1996	MFGR#1216
704898387	1996	MFGR#1217
528843086	1996	MFGR#1218
586246330	1996	MFGR#1219
712110492	1996	MFGR#122
518444215	1996	MFGR#1220
499319414	1996	MFGR#1221
679469356	1996	MFGR#1222
628762754	1996	MFGR#1223
724844856	1996	MFGR#1224
660620587	1996	MFGR#1225
667674729	1996	MFGR#1226
483838085	1996	MFGR#1227
609855391	1996	MFGR#1228
658959557	1996	MFGR#1229
566217852	1996	MFGR#123
528879998	1996	MFGR#1230
589481194	1996	MFGR#1231
702805896	1996	MFGR#1232
663679947	1996	MFGR#1233
571149450	1996	MFGR#1234
478648074	1996	MFGR#1235
568249365	1996	MFGR#1236
592616167	1996	MFGR#1237
466676148	1996	MFGR#1238
670693719	1996	MFGR#1239
560667719	1996	MFGR#124
821167950	1996	MFGR#1240
476864333	1996	MFGR#125
558030884	1996	MFGR#126
635873891	1996	MFGR#127
551010618	1996	MFGR#128
560570630	1996	MFGR#129

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

587013207	1997	MFGR#121
616287892	1997	MFGR#1210
548588761	1997	MFGR#1211
589593892	1997	MFGR#1212
424306670	1997	MFGR#1213
511971910	1997	MFGR#1214
631772246	1997	MFGR#1215
692135140	1997	MFGR#1216
777994957	1997	MFGR#1217
707053720	1997	MFGR#1218
561169527	1997	MFGR#1219
664916245	1997	MFGR#122
594466157	1997	MFGR#1220
588848171	1997	MFGR#1221
528988960	1997	MFGR#1222
537098211	1997	MFGR#1223
674763166	1997	MFGR#1224
450402292	1997	MFGR#1225
701360722	1997	MFGR#1226
506011570	1997	MFGR#1227
585578737	1997	MFGR#1228
622744016	1997	MFGR#1229
646503168	1997	MFGR#123
571800941	1997	MFGR#1230
502601790	1997	MFGR#1231
677924656	1997	MFGR#1232
534455976	1997	MFGR#1233
714934715	1997	MFGR#1234
767151420	1997	MFGR#1235
618877179	1997	MFGR#1236
639638057	1997	MFGR#1237
401953419	1997	MFGR#1238
610756714	1997	MFGR#1239
543248087	1997	MFGR#124
675132692	1997	MFGR#1240
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

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
361416497       1998    MFGR#128
318769573       1998    MFGR#129
Time taken: 177.319 seconds, Fetched: 280 row(s)
hive>
```

```
--Q2.2  No simpifications
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_brand1 between 'MFGR#2221'
  and 'MFGR#2228'
  and s_region = 'ASIA'
group by d_year, p_brand1
order by d_year, p_brand1;
```

```
747356383       1995    MFGR#2226
802502540       1995    MFGR#2227
895936786       1995    MFGR#2228
700010008       1996    MFGR#2221
777310085       1996    MFGR#2222
656095314       1996    MFGR#2223
656859917       1996    MFGR#2224
765820896       1996    MFGR#2225
808177734       1996    MFGR#2226
729563303       1996    MFGR#2227
819665874       1996    MFGR#2228
727342382       1997    MFGR#2221
664533779       1997    MFGR#2222
748288392       1997    MFGR#2223
630422081       1997    MFGR#2224
757391203       1997    MFGR#2225
747889257       1997    MFGR#2226
728857899       1997    MFGR#2227
775312985       1997    MFGR#2228
335304504       1998    MFGR#2221
409347137       1998    MFGR#2222
459109577       1998    MFGR#2223
413318072       1998    MFGR#2224
410402095       1998    MFGR#2225
453515044       1998    MFGR#2226
390506405       1998    MFGR#2227
397939103       1998    MFGR#2228
Time taken: 176.808 seconds, Fetched: 56 row(s)
hive>
```

```
Q2.2 – output
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_brand1 between 'MFGR#2221'
    >   and 'MFGR#2228'
    >   and s_region = 'ASIA'
    > 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_20161026013301_f360e00b-cfbb-41ad-96e2-eb878b0cd1d7
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_20161026013301_f360e00b-cfbb-41ad-96e2-eb878b0cd1d7.log
2016-10-26 01:33:08 Starting to launch local task to process map join;      maximum memory = 518979584
2016-10-26 01:33:10 Dump the side-table for tag: 1 with group count: 2556 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-33-01_622_2668058604798613698-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile101--.hashtable
2016-10-26 01:33:10 Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-33-01_622_2668058604798613698-1/-local-10014/HashTable-Stage-13/MapJoin-mapfile101--.hashtable (67039 bytes)
2016-10-26 01:33:10 End of local task; Time Taken: 1.681 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_1477443484335_0011, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0011/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0011
Hadoop job information for Stage-13: number of mappers: 3; number of reducers: 0
2016-10-26 01:33:18,871 Stage-13 map = 0%,  reduce = 0%
2016-10-26 01:33:49,776 Stage-13 map = 33%,  reduce = 0%, Cumulative CPU 14.8 sec
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 01:33:58,614 Stage-13 map = 67%,  reduce = 0%, Cumulative CPU 19.95 sec
2016-10-26 01:34:17,770 Stage-13 map = 100%,  reduce = 0%, Cumulative CPU 28.16 sec
MapReduce Total cumulative CPU time: 28 seconds 160 msec
Ended Job = job_1477443484335_0011
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_20161026013301_f360e00b-cfbb-41ad-96e2-eb878b0cd1d7.log
2016-10-26 01:34:25 Starting to launch local task to process map join;	maximum memory = 518979584
2016-10-26 01:34:28 Dump the side-table for tag: 1 with group count: 1584 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-33-01_622_2668058604798613698-1/-local-10010/HashTable-Stage-10/MapJoin-mapfile81--.hashtable
2016-10-26 01:34:28 Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-33-01_622_2668058604798613698-1/-local-10010/HashTable-Stage-10/MapJoin-mapfile81--.hashtable (50647 bytes)
2016-10-26 01:34:28 End of local task; Time Taken: 2.601 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 6
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1477443484335_0012, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0012/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0012
Hadoop job information for Stage-10: number of mappers: 1; number of reducers: 0
2016-10-26 01:34:36,614 Stage-10 map = 0%,  reduce = 0%
2016-10-26 01:34:51,542 Stage-10 map = 45%,  reduce = 0%, Cumulative CPU 8.56 sec
2016-10-26 01:34:53,644 Stage-10 map = 100%,  reduce = 0%, Cumulative CPU 10.81 sec
MapReduce Total cumulative CPU time: 10 seconds 810 msec
Ended Job = job_1477443484335_0012
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_20161026013301_f360e00b-cfbb-41ad-96e2-eb878b0cd1d7.log
2016-10-26 01:35:01 Starting to launch local task to process map join;	maximum memory = 518979584
2016-10-26 01:35:02 Dump the side-table for tag: 1 with group count: 449 into file: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-33-01_622_2668058604798613698-1/-local-10008/HashTable-Stage-4/MapJoin-mapfile71--.hashtable
2016-10-26 01:35:02 Uploaded 1 File to: file:/tmp/ec2-user/04556699-e4a7-4295-baa0-ab2c70d8d653/hive_2016-10-26_01-33-01_622_2668058604798613698-1/-local-10008/HashTable-Stage-4/MapJoin-mapfile71--.hashtable (9186 bytes)
2016-10-26 01:35:02 End of local task; Time Taken: 1.666 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 4 out of 6
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1477443484335_0013, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0013/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0013
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2016-10-26 01:35:11,619 Stage-4 map = 0%,  reduce = 0%
2016-10-26 01:35:20,114 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.38 sec
2016-10-26 01:35:28,538 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 3.49 sec
MapReduce Total cumulative CPU time: 3 seconds 490 msec
Ended Job = job_1477443484335_0013
Launching Job 5 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_1477443484335_0014, Tracking URL = http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477443484335_0014/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1477443484335_0014
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2016-10-26 01:35:41,530 Stage-5 map = 0%,  reduce = 0%
2016-10-26 01:35:48,967 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 0.85 sec
2016-10-26 01:35:56,334 Stage-5 map = 100%,  reduce = 100%, Cumulative CPU 1.96 sec
MapReduce Total cumulative CPU time: 1 seconds 960 msec
Ended Job = job_1477443484335_0014
MapReduce Jobs Launched:
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
Stage-Stage-13: Map: 3   Cumulative CPU: 28.16 sec   HDFS Read: 594359101 HDFS Write: 184733291 SUCCESS
Stage-Stage-10: Map: 1   Cumulative CPU: 10.81 sec   HDFS Read: 184739541 HDFS Write: 1768617 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 3.49 sec   HDFS Read: 1780818 HDFS Write: 2076 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 1   Cumulative CPU: 1.96 sec   HDFS Read: 7852 HDFS Write: 1400 SUCCESS
Total MapReduce CPU Time Spent: 44 seconds 420 msec
OK
709524929       1992    MFGR#2221
783846394       1992    MFGR#2222
765052002       1992    MFGR#2223
651488962       1992    MFGR#2224
646528589       1992    MFGR#2225
709650548       1992    MFGR#2226
745556316       1992    MFGR#2227
756901875       1992    MFGR#2228
766521103       1993    MFGR#2221
691475597       1993    MFGR#2222
758220752       1993    MFGR#2223
669662707       1993    MFGR#2224
773854228       1993    MFGR#2225
737087518       1993    MFGR#2226
781967766       1993    MFGR#2227
680880216       1993    MFGR#2228
685777518       1994    MFGR#2221
666524807       1994    MFGR#2222
733993590       1994    MFGR#2223
707869040       1994    MFGR#2224
721251967       1994    MFGR#2225
822495919       1994    MFGR#2226
720837128       1994    MFGR#2227
826225350       1994    MFGR#2228
775437074       1995    MFGR#2221
761354792       1995    MFGR#2222
637832575       1995    MFGR#2223
589765707       1995    MFGR#2224
708290039       1995    MFGR#2225
747356383       1995    MFGR#2226
802502540       1995    MFGR#2227
895936786       1995    MFGR#2228
700010008       1996    MFGR#2221
777310085       1996    MFGR#2222
656095314       1996    MFGR#2223
656859917       1996    MFGR#2224
765820896       1996    MFGR#2225
808177734       1996    MFGR#2226
729563303       1996    MFGR#2227
819665874       1996    MFGR#2228
727342382       1997    MFGR#2221
664533779       1997    MFGR#2222
748288392       1997    MFGR#2223
630422081       1997    MFGR#2224
757391203       1997    MFGR#2225
747889257       1997    MFGR#2226
728857899       1997    MFGR#2227
775312985       1997    MFGR#2228
335304504       1998    MFGR#2221
409347137       1998    MFGR#2222
459109577       1998    MFGR#2223
413318072       1998    MFGR#2224
410402095       1998    MFGR#2225
453515044       1998    MFGR#2226
390506405       1998    MFGR#2227
397939103       1998    MFGR#2228
Time taken: 176.808 seconds, Fetched: 56 row(s)
hive>
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

Part 2 – Pig

Table creation:

```
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;

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;

--Q0.1 Added simple test query
SELECT SUM(lo_revenue)
FROM lineorder;

query01script.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;
group_lineorder = GROUP lineorder ALL;
query01= FOREACH group_lineorder GENERATE SUM(lineorder.lo_revenue);
DUMP query01;
```

```
2016-10-26 18:19:32,377 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 52% complete
2016-10-26 18:19:32,378 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:37,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 69% complete
2016-10-26 18:19:37,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:47,411 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:50,423 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:50,438 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:19:51,811 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:51,817 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:19:51,930 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:51,935 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:19:52,014 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-10-26 18:19:52,016 [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        2016-10-26 18:16:08     2016-10-26 18:19:52     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1477503894873_0002  5       1       202     72      148     137     128     128     128     128     group_lineorder,lineorder,query01       GROUP_BY,COMBINER       hdfs://localhost/tmp/temp-11726
51976/tmp-1193399122,

Input(s):
Successfully read 6001215 records (594331240 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 1 records (13 bytes) in: "hdfs://localhost/tmp/temp-1172651976/tmp-1193399122"

Counters:
Total records written : 1
Total bytes written : 13
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 10
Total records proactively spilled: 8996650

Job DAG:
job_1477503894873_0002


2016-10-26 18:19:52,023 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:52,032 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:19:52,105 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:52,110 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:19:52,180 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:52,190 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:19:52,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-10-26 18:19:52,294 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:19:52,295 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 18:19:52,329 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 18:19:52,329 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(21810219932448)
2016-10-26 18:19:52,477 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 48 seconds and 682 milliseconds (228682 ms)
[ec2-user@ip-172-31-13-121 pig-0.15.0]$
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

OUTPUT--->

```
[ec2-user@ip-172-31-13-121 pig-0.15.0]$ bin/pig -f query01script.pig
16/10/26 18:16:03 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/10/26 18:16:03 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/10/26 18:16:03 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-10-26 18:16:03,911 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-10-26 18:16:03,911 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig-0.15.0/pig_1477505763909.log
2016-10-26 18:16:05,092 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-10-26 18:16:05,509 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-10-26 18:16:05,513 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:16:05,514 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost/
2016-10-26 18:16:06,738 [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}
2016-10-26 18:16:07,439 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2016-10-26 18:16:07,545 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:16:07,556 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 18:16:07,648 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer,
PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-10-26 18:16:07,786 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-10-26 18:16:07,815 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2016-10-26 18:16:07,849 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2016-10-26 18:16:07,849 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2016-10-26 18:16:07,897 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:16:07,946 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:16:08,213 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-10-26 18:16:08,218 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use
mapreduce.reduce.markreset.buffer.percent
2016-10-26 18:16:08,219 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is
not set, set to default 0.3
2016-10-26 18:16:08,219 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use
mapreduce.output.fileoutputformat.compress
2016-10-26 18:16:08,222 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of
required reducers.
2016-10-26 18:16:08,222 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2016-10-26 18:16:08,222 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2016-10-26 18:16:08,222 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-10-26 18:16:08,501 [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 through /tmp/temp-1172651976/tmp-285213670/pig-0.15.0-core-h2.jar
2016-10-26 18:16:08,530 [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-1172651976/tmp-617762308/automaton-1.11-8.jar
2016-10-26 18:16:08,555 [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 DistributedCache through /tmp/temp-1172651976/tmp-1388990332/antlr-runtime-3.4.jar
2016-10-26 18:16:08,586 [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 through /tmp/temp-1172651976/tmp-951600102/joda-time-2.5.jar
2016-10-26 18:16:08,653 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2016-10-26 18:16:08,666 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-10-26 18:16:08,673 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2016-10-26 18:16:08,673 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2016-10-26 18:16:08,772 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2016-10-26 18:16:08,772 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use
mapreduce.jobtracker.http.address
2016-10-26 18:16:08,786 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:16:08,800 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:16:08,869 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or
Job#setJar(String).
2016-10-26 18:16:08,956 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 18:16:08,957 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 18:16:08,997 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 5
2016-10-26 18:16:09,077 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:5
2016-10-26 18:16:09,508 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1477503894873_0002
2016-10-26 18:16:09,686 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2016-10-26 18:16:09,942 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1477503894873_0002
2016-10-26 18:16:10,049 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-13-121.us-west-
2.compute.internal:8088/proxy/application_1477503894873_0002/
2016-10-26 18:16:10,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1477503894873_0002
2016-10-26 18:16:10,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases
group_lineorder,lineorder,query01
2016-10-26 18:16:10,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M:
lineorder[1,11],lineorder[-1,-1],query01[6,9],group_lineorder[5,18] C: query01[6,9],group_lineorder[5,18] R: query01[6,9]
2016-10-26 18:16:10,144 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2016-10-26 18:16:10,145 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:17:07,474 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 4% complete
2016-10-26 18:17:07,474 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:17:12,484 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 9% complete
2016-10-26 18:17:12,485 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:17:19,497 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 13% complete
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 18:17:19,498 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:17:24,505 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 17% complete
2016-10-26 18:17:24,506 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:17:34,572 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 24% complete
2016-10-26 18:17:34,572 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:17:44,594 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 29% complete
2016-10-26 18:17:44,594 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:17:59,616 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 33% complete
2016-10-26 18:17:59,616 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:18:07,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 39% complete
2016-10-26 18:18:07,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:18:52,309 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 44% complete
2016-10-26 18:18:52,315 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:09,340 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 48% complete
2016-10-26 18:19:09,340 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:32,377 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 52% complete
2016-10-26 18:19:32,378 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:37,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 59% complete
2016-10-26 18:19:37,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:47,411 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0002]
2016-10-26 18:19:50,423 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:50,438 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 18:19:51,811 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:51,817 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 18:19:51,930 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:51,935 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 18:19:52,014 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-10-26 18:19:52,016 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAtFeatures
2.6.4   0.15.0  ec2-user        2016-10-26 18:16:08     2016-10-26 18:19:52     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime
        AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1477503894873_0002  5       1       202     72      148     137     128     128     128     128     group_lineorder,lineorder,query01
        GROUP_BY,COMBINER       hdfs://localhost/tmp/temp-1172651976/tmp-1193399122,

Input(s):
Successfully read 6001215 records (594331240 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 1 records (13 bytes) in: "hdfs://localhost/tmp/temp-1172651976/tmp-1193399122"

Counters:
Total records written : 1
Total bytes written : 13
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 10
Total records proactively spilled: 8996650

Job DAG:
job_1477503894873_0002


2016-10-26 18:19:52,023 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:52,032 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 18:19:52,105 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:52,110 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 18:19:52,180 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:19:52,190 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 18:19:52,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-10-26 18:19:52,294 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:19:52,295 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 18:19:52,329 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 18:19:52,329 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(21810219932448)
2016-10-26 18:19:52,477 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 48 seconds and 682 milliseconds (228682 ms)
[ec2-user@ip-172-31-13-121 pig-0.15.0]$
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

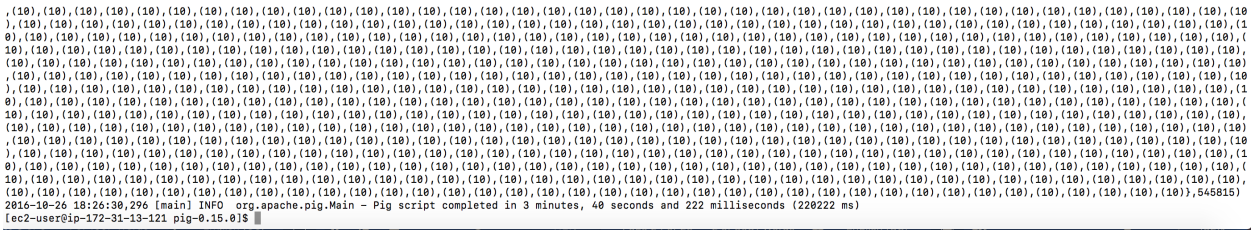
--Q0.2 Added simple test query
SELECT lo_discount, COUNT(lo_extendedprice)
FROM lineorder
GROUP BY lo_discount;

query02script.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;

group_lo_discount= GROUP lineorder BY lo_discount;
query02 = FOREACH group_lo_discount GENERATE lineorder.lo_discount, COUNT(lineorder.lo_extendedprice);
DUMP query02;
```



OUTPUT-->

```
[ec2-user@ip-172-31-13-121 pig-0.15.0]$ bin/pig -f query02script.pig
16/10/26 18:22:50 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/10/26 18:22:50 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/10/26 18:22:50 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-10-26 18:22:50,237 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-10-26 18:22:50,238 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig-0.15.0/pig_1477506170233.log
2016-10-26 18:22:51,476 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-10-26 18:22:51,854 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-10-26 18:22:51,857 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:22:51,857 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost/
2016-10-26 18:22:53,134 [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}
2016-10-26 18:22:53,771 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2016-10-26 18:22:53,875 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:22:53,880 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 18:22:53,968 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer,
PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-10-26 18:22:54,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-10-26 18:22:54,190 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2016-10-26 18:22:54,191 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2016-10-26 18:22:54,248 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:22:54,293 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:22:54,607 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-10-26 18:22:54,612 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use
mapreduce.reduce.markreset.buffer.percent
2016-10-26 18:22:54,613 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is
not set, set to default 0.3
2016-10-26 18:22:54,613 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use
mapreduce.output.fileoutputformat.compress
2016-10-26 18:22:54,616 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of
required reducers.
2016-10-26 18:22:54,617 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator:
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2016-10-26 18:22:54,623 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
maxReducers=999 totalInputFileSize=594313001
2016-10-26 18:22:54,623 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2016-10-26 18:22:54,624 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2016-10-26 18:22:54,624 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-10-26 18:22:54,904 [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 through /tmp/temp-664958524/tmp-380955364/pig-0.15.0-core-h2.jar
2016-10-26 18:22:54,944 [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-664958524/tmp-1111109197/automaton-1.11-8.jar
2016-10-26 18:22:54,975 [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 DistributedCache through /tmp/temp-664958524/tmp-1899972575/antlr-runtime-3.4.jar
2016-10-26 18:22:55,007 [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 through /tmp/temp-664958524/tmp110660210/joda-time-2.5.jar
2016-10-26 18:22:55,075 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2016-10-26 18:22:55,080 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-10-26 18:22:55,083 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2016-10-26 18:22:55,083 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2016-10-26 18:22:55,187 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2016-10-26 18:22:55,193 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracker.http.address
2016-10-26 18:22:55,203 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:22:55,215 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:22:55,293 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2016-10-26 18:22:55,383 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 18:22:55,383 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 18:22:55,406 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 5
2016-10-26 18:22:55,461 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:5
2016-10-26 18:22:55,948 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1477503894873_0003
2016-10-26 18:22:56,119 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2016-10-26 18:22:56,215 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1477503894873_0003
2016-10-26 18:22:56,257 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477503894873_0003/
2016-10-26 18:22:56,257 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1477503894873_0003
2016-10-26 18:22:56,257 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases group_lo_discount,lineorder,query02
2016-10-26 18:22:56,257 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: lineorder[1,11],lineorder[-1,-1],group_lo_discount[6,19] C:  R: query02[7,10]
2016-10-26 18:22:56,272 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2016-10-26 18:22:56,272 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:23:48,520 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 5% complete
2016-10-26 18:23:48,520 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:23:55,531 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 9% complete
2016-10-26 18:23:55,532 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:24:05,547 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 15% complete
2016-10-26 18:24:05,547 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:24:13,558 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 23% complete
2016-10-26 18:24:13,558 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:24:33,587 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 28% complete
2016-10-26 18:24:33,587 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:24:48,673 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 36% complete
2016-10-26 18:24:48,673 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:25:06,717 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 41% complete
2016-10-26 18:25:06,717 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:25:10,725 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 45% complete
2016-10-26 18:25:10,725 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:25:15,744 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 53% complete
2016-10-26 18:25:15,744 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:25:20,752 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 60% complete
2016-10-26 18:25:20,752 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:25:23,755 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 83% complete
2016-10-26 18:25:23,756 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:25:35,773 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 87% complete
2016-10-26 18:25:35,773 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:25:48,790 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 91% complete
2016-10-26 18:25:48,791 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:26:00,807 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 95% complete
2016-10-26 18:26:00,807 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:26:11,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0003]
2016-10-26 18:26:16,839 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:26:16,849 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:26:17,404 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:26:17,409 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:26:17,527 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:26:17,532 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:26:17,609 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 18:26:17,612 [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	2016-10-26 18:22:54	2016-10-26 18:26:17	GROUP_BY

Success!

Job Stats (time in seconds):
JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime
	AvgReduceTime	MedianReducetime	Alias	Feature	Outputs
job_1477503894873_0003	5	1	130	66	117	130	116	116	116	116	group_lo_discount,lineorder,query02
	GROUP_BY	hdfs://localhost/tmp/temp-664958524/tmp-40370149,

Input(s):
Successfully read 6001215 records (594331240 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 11 records (16913123 bytes) in: "hdfs://localhost/tmp/temp-664958524/tmp-40370149"

Counters:
Total records written : 11
Total bytes written : 16913123
Spillable Memory Manager spill count : 27
Total bags proactively spilled: 11
Total records proactively spilled: 5088803

Job DAG:
job_1477503894873_0003


2016-10-26 18:26:17,619 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:26:17,628 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:26:17,705 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:26:17,710 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:26:17,802 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 18:26:17,810 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 18:26:17,906 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-10-26 18:26:17,910 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 18:26:17,911 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 18:26:17,937 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 18:26:17,940 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 18:26:30,296 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 40 seconds and 222 milliseconds (220222 ms)
```

```
--Q0.3 Added simple test query
SELECT lo_quantity, SUM(lo_revenue)
FROM lineorder
WHERE lo_discount > 2
GROUP BY lo_quantity;

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;

filtered_lo_discount = FILTER lineorder BY lo_discount > 2;
group_quantity = GROUP filtered_lo_discount BY lo_quantity;
query03 = FOREACH group_quantity GENERATE lineorder.lo_quantity, SUM(filtered_lo_discount.lo_revenue);

DUMP query03;
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 19:08:19,629 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-10-26 19:08:19,646 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:19,656 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=FAILED. Redirecting to job history server
2016-10-26 19:08:19,812 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:19,821 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=FAILED. Redirecting to job history server
2016-10-26 19:08:19,900 [main] ERROR org.apache.pig.tools.pigstats.PigStats - ERROR 0: org.apache.pig.backend.executionengine.ExecException: ERROR 0: Scalar has more than one row in the output. 1st : (1,1
,7381,155190,828,19960102,5-LOW,0,17,2116823,17366547,4,2032150,74711,2,19960212,TRUCK), 2nd :(1,2,7381,67310,163,19960102,5-LOW,0,36,4598316,17366547,9,4184467,76638,6,19960228,MAIL) (common cause: "JOIN
" then "FOREACH ... GENERATE foo.bar" should be "foo::bar" )
2016-10-26 19:08:19,900 [main] ERROR org.apache.pig.tools.pigstats.mapreduce.MRPigStatsUtil - 1 map reduce job(s) failed!
2016-10-26 19:08:19,903 [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        2016-10-26 19:03:56     2016-10-26 19:08:19     GROUP_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   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1477503894873_0004  5       0       122     70      110     119     0       0       0       0       lineorder       MULTI_QUERY,MAP_ONLY

Failed Jobs:
JobId   Alias   Feature Message Outputs
job_1477503894873_0005  filtered_lo_discount,group_quantity,query03     GROUP_BY,COMBINER       Message: Job failed!    hdfs://localhost/tmp/temp-813227653/tmp-877501454,

Input(s):
Successfully read 6001215 records (594331240 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Failed to produce result in "hdfs://localhost/tmp/temp-813227653/tmp-877501454"

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_1477503894873_0004  ->      job_1477503894873_0005,
job_1477503894873_0005


2016-10-26 19:08:19,908 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:19,919 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:08:20,029 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:20,043 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:08:20,110 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:20,119 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:08:20,221 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Some jobs have failed! Stop running all dependent jobs
2016-10-26 19:08:20,240 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1066: Unable to open iterator for alias query03. Backend error : org.apache.pig.backend.executionengine.ExecException: ERROR 0
: Scalar has more than one row in the output. 1st : (1,1,7381,155190,828,19960102,5-LOW,0,17,2116823,17366547,4,2032150,74711,2,19960212,TRUCK), 2nd :(1,2,7381,67310,163,19960102,5-LOW,0,36,4598316,173665
47,9,4184467,76638,6,19960228,MAIL) (common cause: "JOIN" then "FOREACH ... GENERATE foo.bar" should be "foo::bar" )
Details at logfile: /home/ec2-user/pig-0.15.0/pig_1477508632231.log
2016-10-26 19:08:20,282 [main] INFO  org.apache.pig.Main - Pig script completed in 4 minutes, 28 seconds and 144 milliseconds (268144 ms)
[ec2-user@ip-172-31-13-121 pig-0.15.0]$
```

OUTPUT-->
[ec2-user@ip-172-31-13-121 pig-0.15.0]$ bin/pig -f query03script.pig
16/10/26 19:03:52 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/10/26 19:03:52 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/10/26 19:03:52 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-10-26 19:03:52,232 [main] INFO org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-10-26 19:03:52,232 [main] INFO org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig-0.15.0/pig_1477508632231.log
2016-10-26 19:03:53,383 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-10-26 19:03:53,789 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-10-26 19:03:53,790 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 19:03:53,790 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost/
2016-10-26 19:03:54,990 [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}
2016-10-26 19:03:55,707 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,FILTER
2016-10-26 19:03:55,813 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 19:03:55,831 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 19:03:55,905 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer,
PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-10-26 19:03:56,024 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-10-26 19:03:56,053 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 19:03:56,056 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 19:03:56,107 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 5
2016-10-26 19:03:56,107 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - number of input files: 5
2016-10-26 19:03:56,107 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - number of input files: 0
2016-10-26 19:03:56,112 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2016-10-26 19:03:56,190 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2016-10-26 19:03:56,190 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 1 map-only splittees.
2016-10-26 19:03:56,190 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 1 out of total 3 MR operators.
2016-10-26 19:03:56,190 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 2
2016-10-26 19:03:56,236 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 19:03:56,289 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:03:56,615 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-10-26 19:03:56,621 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use
mapreduce.reduce.markreset.buffer.percent
2016-10-26 19:03:56,626 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is
not set, set to default 0.3
2016-10-26 19:03:56,626 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use
mapreduce.output.fileoutputformat.compress
2016-10-26 19:03:56,627 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-10-26 19:03:56,859 [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 through /tmp/temp-813227653/tmp-1977460989/pig-0.15.0-core-h2.jar
2016-10-26 19:03:56,878 [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-813227653/tmp1673987345/automaton-1.11-8.jar
2016-10-26 19:03:56,903 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
0.15.0/lib/antlr-runtime-3.4.jar to DistributedCache through /tmp/temp-813227653/tmp447594768/antlr-runtime-3.4.jar
2016-10-26 19:03:56,934 [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 through /tmp/temp-813227653/tmp-1901580675/joda-time-2.5.jar
2016-10-26 19:03:56,993 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up multi store job
2016-10-26 19:03:57,005 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-10-26 19:03:57,005 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2016-10-26 19:03:57,006 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2016-10-26 19:03:57,091 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2016-10-26 19:03:57,096 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracker.http.address
2016-10-26 19:03:57,099 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:03:57,111 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 19:03:57,179 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2016-10-26 19:03:57,263 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 19:03:57,264 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 19:03:57,267 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 5
2016-10-26 19:03:57,329 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:5
2016-10-26 19:03:57,787 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1477503894873_0004
2016-10-26 19:03:57,966 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2016-10-26 19:03:58,056 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1477503894873_0004
2016-10-26 19:03:58,095 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477503894873_0004/
2016-10-26 19:03:58,095 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1477503894873_0004
2016-10-26 19:03:58,095 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases lineorder
2016-10-26 19:03:58,095 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: lineorder[1,11],lineorder[-1,-1] C:  R:
2016-10-26 19:03:58,106 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2016-10-26 19:03:58,106 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0004]
2016-10-26 19:04:52,907 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 4% complete
2016-10-26 19:04:52,907 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0004]
2016-10-26 19:05:05,429 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 8% complete
2016-10-26 19:05:05,429 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0004]
2016-10-26 19:05:17,453 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 13% complete
2016-10-26 19:05:17,453 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0004]
2016-10-26 19:05:35,479 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 17% complete
2016-10-26 19:05:35,479 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0004]
2016-10-26 19:05:55,509 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 21% complete
2016-10-26 19:05:55,510 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0004]
2016-10-26 19:06:13,534 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2016-10-26 19:06:13,544 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:06:13,555 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:06:13,853 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:06:13,861 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:06:13,916 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2016-10-26 19:06:13,917 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:06:13,921 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:06:13,983 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-10-26 19:06:13,988 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-10-26 19:06:13,989 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2016-10-26 19:06:13,990 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2016-10-26 19:06:14,013 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=478908948
2016-10-26 19:06:14,014 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2016-10-26 19:06:14,014 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-10-26 19:06:14,067 [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 through /tmp/temp-813227653/tmp1408707129/pig-0.15.0-core-h2.jar
2016-10-26 19:06:14,084 [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-813227653/tmp-1268605596/automaton-1.11-8.jar
2016-10-26 19:06:14,099 [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 DistributedCache through /tmp/temp-813227653/tmp26003363/antlr-runtime-3.4.jar
2016-10-26 19:06:14,123 [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 through /tmp/temp-813227653/tmp-1718634979/joda-time-2.5.jar
2016-10-26 19:06:14,139 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2016-10-26 19:06:14,140 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-10-26 19:06:14,140 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2016-10-26 19:06:14,140 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2016-10-26 19:06:14,230 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2016-10-26 19:06:14,238 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:06:14,252 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 19:06:14,290 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 19:06:14,348 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 5
2016-10-26 19:06:14,349 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 5
2016-10-26 19:06:14,350 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 5
2016-10-26 19:06:14,387 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:5
2016-10-26 19:06:14,443 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1477503894873_0005
2016-10-26 19:06:14,447 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2016-10-26 19:06:14,491 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1477503894873_0005
2016-10-26 19:06:14,494 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-13-121.us-west-
2.compute.internal:8088/proxy/application_1477503894873_0005/
2016-10-26 19:06:14,738 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1477503894873_0005
2016-10-26 19:06:14,738 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases
filtered_lo_discount,group_quantity,query03
2016-10-26 19:06:14,738 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M:
filtered_lo_discount[6,23],query03[8,9],group_quantity[7,16] C: query03[8,9],group_quantity[7,16] R: query03[8,9]
2016-10-26 19:07:07,009 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 55% complete
2016-10-26 19:07:07,009 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0005]
2016-10-26 19:07:14,027 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 60% complete
2016-10-26 19:07:14,028 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0005]
2016-10-26 19:07:18,534 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 64% complete
2016-10-26 19:07:18,535 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0005]
2016-10-26 19:07:43,570 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 70% complete
2016-10-26 19:07:43,570 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0005]
2016-10-26 19:07:49,578 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 79% complete
2016-10-26 19:07:49,578 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0005]
2016-10-26 19:08:09,609 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0005]
2016-10-26 19:08:19,629 [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 immediately on failure.
2016-10-26 19:08:19,629 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - job job_1477503894873_0005 has failed! Stop
running all dependent jobs
2016-10-26 19:08:19,629 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-10-26 19:08:19,646 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:19,656 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=FAILED. Redirecting to job
history server
2016-10-26 19:08:19,812 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:19,821 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=FAILED. Redirecting to job
history server
2016-10-26 19:08:19,900 [main] ERROR org.apache.pig.tools.pigstats.PigStats - ERROR 0: org.apache.pig.backend.executionengine.ExecException: ERROR 0: Scalar has more
than one row in the output. 1st : (1,1,7381,155190,828,19960102,5-LOW,0,17,2116823,17366547,4,2032150,74711,2,19960212,TRUCK), 2nd :(1,2,7381,67310,163,19960102,5-
LOW,0,36,4598316,17366547,9,4184467,76638,6,19960228,MAIL) (common cause: "JOIN" then "FOREACH ... GENERATE foo.bar" should be "foo::bar" )
2016-10-26 19:08:19,900 [main] ERROR org.apache.pig.tools.pigstats.mapreduce.MRPigStatsUtil - 1 map reduce job(s) failed!
2016-10-26 19:08:19,903 [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	2016-10-26 19:03:56	2016-10-26 19:08:19	GROUP_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
	AvgReduceTime	MedianReducetime	Alias	Feature	Outputs
job_1477503894873_0004	5	0	122	70	110	119	0	0	0	0	lineorder	MULTI_QUERY,MAP_ONLY

Failed Jobs:
JobId	Alias	Feature	Message	Outputs
job_1477503894873_0005	filtered_lo_discount,group_quantity,query03	GROUP_BY,COMBINER	Message: Job failed!
	hdfs://localhost/tmp/temp-813227653/tmp-877501454,

Input(s):
Successfully read 6001215 records (594331240 bytes) from: "/user/ec2-user/lineorder.tbl"

Output(s):
Failed to produce result in "hdfs://localhost/tmp/temp-813227653/tmp-877501454"

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_1477503894873_0004	->	job_1477503894873_0005,
job_1477503894873_0005


2016-10-26 19:08:19,908 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:19,919 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 19:08:20,029 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:20,043 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:08:20,110 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 19:08:20,119 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 19:08:20,221 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Some jobs have failed! Stop running all dependent jobs
2016-10-26 19:08:20,240 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1066: Unable to open iterator for alias query03. Backend error : org.apache.pig.backend.executionengine.ExecException: ERROR 0: Scalar has more than one row in the output. 1st : (1,1,7381,155190,828,19960102,5-LOW,0,17,2116823,17366547,4,2032150,74711,2,19960212,TRUCK), 2nd :(1,2,7381,67310,163,19960102,5-LOW,0,36,4598316,17366547,9,4184467,76638,6,19960228,MAIL) (common cause: "JOIN" then "FOREACH ... GENERATE foo.bar" should be "foo::bar" )
Details at logfile: /home/ec2-user/pig-0.15.0/pig_1477508632231.log
2016-10-26 19:08:20,282 [main] INFO  org.apache.pig.Main - Pig script completed in 4 minutes, 28 seconds and 144 milliseconds (268144 ms)
[ec2-user@ip-172-31-13-121 pig-0.15.0]$
```

```
--Q1.1  Simplified to remove expression in sum
select sum(lo_extendedprice) as revenue
from lineorder, dwdate
where lo_orderdate = d_datekey
  and d_year = 1993
  and lo_discount between 1 and 3
  and lo_quantity < 25;
```

```
query11script.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;

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;

group_year = FILTER dwdate BY d_year==1993;
group_discount = FILTER lineorder BY lo_discount==1 OR lo_discount==2 OR lo_discount==3;
group_quantity = FILTER group_discount BY lo_quantity < 25;
join_date = JOIN group_quantity BY lo_orderdate, group_year BY d_datekey;
query11 = FOREACH (GROUP join_date ALL) GENERATE SUM(join_date.lo_extendedprice);

DUMP query11;
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 20:37:08,117 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,227 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,235 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,311 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-10-26 20:37:08,334 [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        2016-10-26 20:34:05     2016-10-26 20:37:08     HASH_JOIN,GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1477503894873_0008  6       1       130     38      96      111     94      94      94      94      dwdate,group_discount,group_year,join_date,lineorder    HASH_JOIN
job_1477503894873_0009  1       1       7       7       7       7       5       5       5       5       1-13,query11    GROUP_BY,COMBINER       hdfs://localhost/tmp/temp-1205792689/tmp1305542459,

Input(s):
Successfully read 2556 records from: "/user/ec2-user/dwdate.tbl"
Successfully read 6001215 records from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 1 records (13 bytes) in: "hdfs://localhost/tmp/temp-1205792689/tmp1305542459"

Counters:
Total records written : 1
Total bytes written : 13
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1477503894873_0008  ->      job_1477503894873_0009,
job_1477503894873_0009


2016-10-26 20:37:08,338 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,350 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,475 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,488 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,577 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,589 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,698 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,715 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,813 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,830 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,906 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,915 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,981 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-10-26 20:37:08,984 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 20:37:08,990 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 20:37:08,996 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 20:37:08,996 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(22291249488)
2016-10-26 20:37:09,118 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 8 seconds and 639 milliseconds (188639 ms)
[ec2-user@ip-172-31-13-121 pig-0.15.0]$
```

OUTPUT ---
[ec2-user@ip-172-31-13-121 pig-0.15.0]$ bin/pig -f query11script.pig
16/10/26 20:34:00 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/10/26 20:34:00 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/10/26 20:34:00 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-10-26 20:34:00,604 [main] INFO org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35
2016-10-26 20:34:00,604 [main] INFO org.apache.pig.Main - Logging error messages to: /home/ec2-user/pig-0.15.0/pig_1477514040602.log
2016-10-26 20:34:01,759 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/ec2-user/.pigbootup not found
2016-10-26 20:34:02,169 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-10-26 20:34:02,177 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 20:34:02,178 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost/
2016-10-26 20:34:03,473 [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}
2016-10-26 20:34:03,865 [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}
2016-10-26 20:34:04,545 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,FILTER
2016-10-26 20:34:04,603 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 20:34:04,608 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 20:34:04,647 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2016-10-26 20:34:04,783 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-10-26 20:34:04,802 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2016-10-26 20:34:04,822 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2016-10-26 20:34:04,833 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2016-10-26 20:34:04,833 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 2
2016-10-26 20:34:04,854 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 20:34:04,958 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:34:05,204 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-10-26 20:34:05,219 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2016-10-26 20:34:05,220 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-10-26 20:34:05,220 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2016-10-26 20:34:05,229 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2016-10-26 20:34:05,230 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2016-10-26 20:34:05,238 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=594542966
2016-10-26 20:34:05,238 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2016-10-26 20:34:05,238 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2016-10-26 20:34:05,239 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-10-26 20:34:05,503 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/ec2-user/pig-0.15.0/pig-

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

0.15.0-core-h2.jar to DistributedCache through /tmp/temp-1205792689/tmp-1910681143/pig-0.15.0-core-h2.jar
2016-10-26 20:34:05,522 [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-1205792689/tmp-1853464396/automaton-1.11-8.jar
2016-10-26 20:34:05,547 [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 DistributedCache through /tmp/temp-1205792689/tmp1687947110/antlr-runtime-3.4.jar
2016-10-26 20:34:05,580 [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 through /tmp/temp-1205792689/tmp97805630/joda-time-2.5.jar
2016-10-26 20:34:05,648 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2016-10-26 20:34:05,660 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-10-26 20:34:05,660 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2016-10-26 20:34:05,660 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2016-10-26 20:34:05,774 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2016-10-26 20:34:05,775 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracker.http.address
2016-10-26 20:34:05,778 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:34:05,796 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 20:34:05,868 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See Job or Job#setJar(String).
2016-10-26 20:34:05,945 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 20:34:05,946 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 20:34:05,985 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2016-10-26 20:34:05,994 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 20:34:05,994 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 20:34:05,998 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 5
2016-10-26 20:34:06,062 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:6
2016-10-26 20:34:06,495 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1477503894873_0008
2016-10-26 20:34:06,699 [JobControl] INFO org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2016-10-26 20:34:06,777 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1477503894873_0008
2016-10-26 20:34:06,817 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477503894873_0008/
2016-10-26 20:34:06,818 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1477503894873_0008
2016-10-26 20:34:06,818 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases dwdate,group_discount,group_year,join_date,lineorder
2016-10-26 20:34:06,818 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: dwdate[5,8],dwdate[-1,-1],group_year[9,13],join_date[12,12],lineorder[1,11],lineorder[-1,-1],group_discount[10,17],join_date[12,12] C: R:
2016-10-26 20:34:06,831 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2016-10-26 20:34:06,832 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:34:53,875 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 4% complete
2016-10-26 20:34:53,876 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:35:21,921 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 8% complete
2016-10-26 20:35:21,921 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:35:31,940 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 13% complete
2016-10-26 20:35:31,941 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:35:53,982 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 17% complete
2016-10-26 20:35:53,982 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:36:09,013 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 25% complete
2016-10-26 20:36:09,013 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:36:24,034 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 29% complete
2016-10-26 20:36:24,034 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:36:29,041 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 41% complete
2016-10-26 20:36:29,041 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:36:32,045 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2016-10-26 20:36:32,045 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0008]
2016-10-26 20:36:37,066 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:36:37,081 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:36:37,355 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:36:37,361 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:36:37,442 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:36:37,452 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:36:37,528 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2016-10-26 20:36:37,529 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-10-26 20:36:37,530 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2016-10-26 20:36:37,530 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2016-10-26 20:36:37,530 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2016-10-26 20:36:37,589 [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 through /tmp/temp-1205792689/tmp-81457318/pig-0.15.0-core-h2.jar
2016-10-26 20:36:37,617 [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-1205792689/tmp-1931305721/automaton-1.11-8.jar
2016-10-26 20:36:37,642 [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 DistributedCache through /tmp/temp-1205792689/tmp-1309718419/antlr-runtime-3.4.jar
2016-10-26 20:36:37,663 [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 through /tmp/temp-1205792689/tmp602071614/joda-time-2.5.jar
2016-10-26 20:36:37,675 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 20:36:37,676 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-10-26 20:36:37,676 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2016-10-26 20:36:37,676 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2016-10-26 20:36:37,736 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2016-10-26 20:36:37,739 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:36:37,754 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 20:36:37,769 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or
Job#setJar(String).
2016-10-26 20:36:37,813 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 20:36:37,813 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2016-10-26 20:36:37,813 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2016-10-26 20:36:37,846 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2016-10-26 20:36:37,930 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1477503894873_0009
2016-10-26 20:36:37,935 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2016-10-26 20:36:37,986 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1477503894873_0009
2016-10-26 20:36:37,991 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-172-31-13-121.us-west-
2.compute.internal:8088/proxy/application_1477503894873_0009/
2016-10-26 20:36:38,238 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1477503894873_0009
2016-10-26 20:36:38,238 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases 1-13,query11
2016-10-26 20:36:38,238 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: query11[13,10],1-
13[13,19] C: query11[13,10],1-13[13,19] R: query11[13,10]
2016-10-26 20:36:55,486 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 75% complete
2016-10-26 20:36:55,486 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0009]
2016-10-26 20:37:05,501 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1477503894873_0009]
2016-10-26 20:37:08,011 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,020 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 20:37:08,110 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,117 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 20:37:08,227 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,235 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 20:37:08,311 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-10-26 20:37:08,334 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion       PigVersionUserId      StartedAt  FinishedAtFeatures
2.6.4       0.15.0      ec2-user    2016-10-26 20:34:05 2016-10-26 20:37:08 HASH_JOIN,GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId       Maps        Reduces     MaxMapTime          MinMapTime          AvgMapTime          MedianMapTime       MaxReduceTime       MinReduceTime
            AvgReduceTime       MedianReducetime    Alias       Feature     Outputs
job_1477503894873_0008  6           1           130         38          96          111         94          94          94          94
            dwdate,group_discount,group_year,join_date,lineorder        HASH_JOIN
job_1477503894873_0009  1           1           7           7           7           7           5           5           5           5           1-13,query11GROUP_BY,COMBINER
            hdfs://localhost/tmp/temp-1205792689/tmp1305542459,

Input(s):
Successfully read 2556 records from: "/user/ec2-user/dwdate.tbl"
Successfully read 6001215 records from: "/user/ec2-user/lineorder.tbl"

Output(s):
Successfully stored 1 records (13 bytes) in: "hdfs://localhost/tmp/temp-1205792689/tmp1305542459"

Counters:
Total records written : 1
Total bytes written : 13
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1477503894873_0008      ->      job_1477503894873_0009,
job_1477503894873_0009


2016-10-26 20:37:08,338 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,350 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 20:37:08,475 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,488 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 20:37:08,577 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,589 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job
history server
2016-10-26 20:37:08,698 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
2016-10-26 20:37:08,715 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,813 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,830 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,906 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2016-10-26 20:37:08,915 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-10-26 20:37:08,981 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-10-26 20:37:08,984 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-10-26 20:37:08,990 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2016-10-26 20:37:08,996 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-10-26 20:37:08,996 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(222914249488)
2016-10-26 20:37:09,118 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 8 seconds and 639 milliseconds (188639 ms)
[ec2-user@ip-172-31-13-121 pig-0.15.0]$
```

PART 3 – Hadoop Streaming

```
--Q0.3 Added simple test query
SELECT lo_quantity, SUM(lo_revenue)
FROM lineorder
WHERE lo_discount > 2
GROUP BY lo_quantity;
######CODE############
mymapper.py ----->
#!/usr/bin/python
import sys, datetime

for line in sys.stdin:
   line = line.strip()
   vals = line.split("|")
   lo_discount = int(vals[11])
   if(lo_discount > 2):
      lo_quantity = int(vals[8])
      lo_revenue = int(vals[12])
      print "%s\t%s\t" % (lo_quantity, lo_revenue)

myreducer.py ----->
#!/usr/bin/python
import sys

curr_id = None
curr_count = 0
key_id = None

for line in sys.stdin:
   line = line.strip()
   val = line.split("\t")
   #key is lo_quantity, value is lo_revenue
   key_id = int(val[0])
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
    value = int(val[1])

    if curr_id == key_id:
        curr_count += value
    else:
        if curr_id:
            print '%d\t%d'%(curr_id,curr_count)
        curr_id = key_id
        curr_count=value

# output the last key
if curr_id == key_id:
    print '%d\t%d' %(curr_id,curr_count)
#########CODE ENDS HERE############
```

Hadoop streaming --->

hadoop jar hadoop-streaming-2.6.4.jar -input lineorder.tbl -output /data/part3 -mapper mymapper.py -reducer myreducer.py -file myreducer.py -file mymapper.py

```
[ec2-user@ip-172-31-13-121 ~]$ hadoop jar hadoop-streaming-2.6.4.jar -input lineorder.tbl -output /data/part3 -mapper mymapper.py -reducer myreducer.py -file myreducer.py -file m
ymapper.py
16/10/26 23:35:50 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [myreducer.py, mymapper.py, /tmp/hadoop-unjar1643721615283798024/] [] /tmp/streamjob7604183718040177240.jar tmpDir=null
16/10/26 23:35:52 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
16/10/26 23:35:53 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
16/10/26 23:35:53 INFO mapred.FileInputFormat: Total input paths to process : 1
16/10/26 23:35:53 INFO mapreduce.JobSubmitter: number of splits:5
16/10/26 23:35:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1477503894873_0012
16/10/26 23:35:54 INFO impl.YarnClientImpl: Submitted application application_1477503894873_0012
16/10/26 23:35:54 INFO mapreduce.Job: The url to track the job: http://ip-172-31-13-121.us-west-2.compute.internal:8088/proxy/application_1477503894873_0012/
16/10/26 23:35:54 INFO mapreduce.Job: Running job: job_1477503894873_0012
16/10/26 23:36:01 INFO mapreduce.Job: Job job_1477503894873_0012 running in uber mode : false
16/10/26 23:36:01 INFO mapreduce.Job:  map 0% reduce 0%
16/10/26 23:36:32 INFO mapreduce.Job:  map 4% reduce 0%
16/10/26 23:36:37 INFO mapreduce.Job:  map 13% reduce 0%
16/10/26 23:36:40 INFO mapreduce.Job:  map 21% reduce 0%
16/10/26 23:36:43 INFO mapreduce.Job:  map 27% reduce 0%
16/10/26 23:36:46 INFO mapreduce.Job:  map 36% reduce 0%
16/10/26 23:36:49 INFO mapreduce.Job:  map 42% reduce 0%
16/10/26 23:36:51 INFO mapreduce.Job:  map 48% reduce 0%
16/10/26 23:36:52 INFO mapreduce.Job:  map 55% reduce 0%
16/10/26 23:36:56 INFO mapreduce.Job:  map 60% reduce 0%
16/10/26 23:36:59 INFO mapreduce.Job:  map 64% reduce 0%
16/10/26 23:37:02 INFO mapreduce.Job:  map 68% reduce 0%
16/10/26 23:37:05 INFO mapreduce.Job:  map 72% reduce 0%
16/10/26 23:37:08 INFO mapreduce.Job:  map 73% reduce 0%
16/10/26 23:37:16 INFO mapreduce.Job:  map 80% reduce 0%
16/10/26 23:37:17 INFO mapreduce.Job:  map 100% reduce 0%
16/10/26 23:37:22 INFO mapreduce.Job:  map 100% reduce 70%
16/10/26 23:37:25 INFO mapreduce.Job:  map 100% reduce 81%
16/10/26 23:37:28 INFO mapreduce.Job:  map 100% reduce 91%
16/10/26 23:37:31 INFO mapreduce.Job:  map 100% reduce 100%
16/10/26 23:37:32 INFO mapreduce.Job: Job job_1477503894873_0012 completed successfully
16/10/26 23:37:32 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=59703998
                FILE: Number of bytes written=120067923
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=594329865
                HDFS: Number of bytes written=783
                HDFS: Number of read operations=18
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=594329865
                HDFS: Number of bytes written=783
                HDFS: Number of read operations=18
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=6
                Launched reduce tasks=1
                Data-local map tasks=6
                Total time spent by all maps in occupied slots (ms)=360731
                Total time spent by all reduces in occupied slots (ms)=38298
                Total time spent by all map tasks (ms)=360731
                Total time spent by all reduce tasks (ms)=38298
                Total vcore-milliseconds taken by all map tasks=360731
                Total vcore-milliseconds taken by all reduce tasks=38298
                Total megabyte-milliseconds taken by all map tasks=369388544
                Total megabyte-milliseconds taken by all reduce tasks=39217152
        Map-Reduce Framework
                Map input records=6001215
                Map output records=4364322
                Map output bytes=50975348
                Map output materialized bytes=59704022
                Input split bytes=480
                Combine input records=0
                Combine output records=0
                Reduce input groups=50
                Reduce shuffle bytes=59704022
                Reduce input records=4364322
                Reduce output records=50
                Spilled Records=8728644
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=2167
                CPU time spent (ms)=49390
                Physical memory (bytes) snapshot=1184129024
                Virtual memory (bytes) snapshot=5727035392
                Total committed heap usage (bytes)=808890368
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=594329385
        File Output Format Counters
                Bytes Written=783
16/10/26 23:37:32 INFO streaming.StreamJob: Output directory: /data/part3
[ec2-user@ip-172-31-13-121 ~]$
```

```
[ec2-user@ip-172-31-13-121 ~]$ hadoop fs -ls /data/part3
Found 2 items
-rw-r--r--   1 ec2-user supergroup          0 2016-10-26 23:37 /data/part3/_SUCCESS
-rw-r--r--   1 ec2-user supergroup        783 2016-10-26 23:37 /data/part3/part-00000
[ec2-user@ip-172-31-13-121 ~]$
```

```
[ec2-user@ip-172-31-13-121 ~]$ hadoop fs -cat /data/part3/part-00000
1       12270917013
10      121796006919
11      134166410947
12      146618690455
13      158899514794
14      170856792200
15      183716500772
16      196359425660
17      208300748323
18      219285225539
19      232654870670
2       24370723319
20      245014170435
21      257940008985
22      268096778479
23      282074197891
24      293515134089
25      306929186636
26      316879287165
27      329860583929
28      344003246981
29      353872128807
3       36741426823
30      366267941974
31      379551414675
32      392471794497
33      403069114917
34      416535036839
35      431730728563
36      442350607436
37      454316923501
38      464266209209
39      478088120865
4       48774169472
40      487839190553
41      504472343691
42      515015195570
43      525590086101
44      539140257365
45      549226253763
46      564234885296
47      574869730412
48      588408783728
49      599469056351
5       61026172645
50      610179291714
6       73051794402
7       85856252908
8       98209492651
9       110995371932
[ec2-user@ip-172-31-13-121 ~]$
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
Output --->
[ec2-user@ip-172-31-13-121 ~]$ hadoop jar hadoop-streaming-2.6.4.jar -input lineorder.tbl -output /data/part3 -mapper mymapper.py -
reducer myreducer.py -file myreducer.py -file mymapper.py
16/10/26 23:35:50 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [myreducer.py, mymapper.py, /tmp/hadoop-unjar1643721615283798024/] [] /tmp/streamjob7604183718040177240.jar
tmpDir=null
16/10/26 23:35:52 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
16/10/26 23:35:53 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
16/10/26 23:35:53 INFO mapred.FileInputFormat: Total input paths to process : 1
16/10/26 23:35:53 INFO mapreduce.JobSubmitter: number of splits:5
16/10/26 23:35:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1477503894873_0012
16/10/26 23:35:54 INFO impl.YarnClientImpl: Submitted application application_1477503894873_0012
16/10/26 23:35:54 INFO mapreduce.Job: The url to track the job: http://ip-172-31-13-121.us-west-
2.compute.internal:8088/proxy/application_1477503894873_0012/
16/10/26 23:35:54 INFO mapreduce.Job: Running job: job_1477503894873_0012
16/10/26 23:36:01 INFO mapreduce.Job: Job job_1477503894873_0012 running in uber mode : false
16/10/26 23:36:01 INFO mapreduce.Job:  map 0% reduce 0%
16/10/26 23:36:32 INFO mapreduce.Job:  map 4% reduce 0%
16/10/26 23:36:37 INFO mapreduce.Job:  map 13% reduce 0%
16/10/26 23:36:40 INFO mapreduce.Job:  map 21% reduce 0%
16/10/26 23:36:43 INFO mapreduce.Job:  map 27% reduce 0%
16/10/26 23:36:46 INFO mapreduce.Job:  map 36% reduce 0%
16/10/26 23:36:49 INFO mapreduce.Job:  map 42% reduce 0%
16/10/26 23:36:51 INFO mapreduce.Job:  map 48% reduce 0%
16/10/26 23:36:52 INFO mapreduce.Job:  map 55% reduce 0%
16/10/26 23:36:56 INFO mapreduce.Job:  map 60% reduce 0%
16/10/26 23:36:59 INFO mapreduce.Job:  map 64% reduce 0%
16/10/26 23:37:02 INFO mapreduce.Job:  map 68% reduce 0%
16/10/26 23:37:05 INFO mapreduce.Job:  map 72% reduce 0%
16/10/26 23:37:08 INFO mapreduce.Job:  map 73% reduce 0%
16/10/26 23:37:16 INFO mapreduce.Job:  map 80% reduce 0%
16/10/26 23:37:17 INFO mapreduce.Job:  map 100% reduce 0%
16/10/26 23:37:22 INFO mapreduce.Job:  map 100% reduce 70%
16/10/26 23:37:25 INFO mapreduce.Job:  map 100% reduce 81%
16/10/26 23:37:28 INFO mapreduce.Job:  map 100% reduce 91%
16/10/26 23:37:31 INFO mapreduce.Job:  map 100% reduce 100%
16/10/26 23:37:32 INFO mapreduce.Job: Job job_1477503894873_0012 completed successfully
16/10/26 23:37:32 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=59703998
                FILE: Number of bytes written=120067923
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=594329865
                HDFS: Number of bytes written=783
                HDFS: Number of read operations=18
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=6
                Launched reduce tasks=1
                Data-local map tasks=6
                Total time spent by all maps in occupied slots (ms)=360731
                Total time spent by all reduces in occupied slots (ms)=38298
                Total time spent by all map tasks (ms)=360731
                Total time spent by all reduce tasks (ms)=38298
                Total vcore-milliseconds taken by all map tasks=360731
                Total vcore-milliseconds taken by all reduce tasks=38298
                Total megabyte-milliseconds taken by all map tasks=369388544
                Total megabyte-milliseconds taken by all reduce tasks=39217152
```

Brunda Chouthoy
CSC 555: Mining Big Data - Project Phase-1
DePaul ID: 1804455

```
        Map-Reduce Framework
                Map input records=6001215
                Map output records=4364322
                Map output bytes=50975348
                Map output materialized bytes=59704022
                Input split bytes=480
                Combine input records=0
                Combine output records=0
                Reduce input groups=50
                Reduce shuffle bytes=59704022
                Reduce input records=4364322
                Reduce output records=50
                Spilled Records=8728644
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=2167
                CPU time spent (ms)=49390
                Physical memory (bytes) snapshot=1184129024
                Virtual memory (bytes) snapshot=5727035392
                Total committed heap usage (bytes)=808890368
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=594329385
        File Output Format Counters
                Bytes Written=783
16/10/26 23:37:32 INFO streaming.StreamJob: Output directory: /data/part3
[ec2-user@ip-172-31-13-121 ~]$
```