Pig – Join and Aggregate

Command

Lineorder = LOAD '/data/joinLineorderCustomer/lineorder.tbl' using PigStorage('|') AS (lo_orderkey: int, lo_linenumber: int, lo_custkey: int, lo_partkey: int, lo_suppkey: int, lo_orderdate: int, lo_orderpriority: chararray, lo_shippriority: chararray, lo_quantity: int, lo_extendedprice: int, lo_ordertotalprice: int, lo_discount: int, lo_revenue: int, lo_supplycost: int, lo_tax: int, lo_commitdate: int, lo_shipmode: chararray);

Customer = LOAD '/data/joinLineorderCustomer/customer.tbl' using PigStorage('|') AS (c_custkey: int, c_name: chararray, c_address: chararray, c_city: chararray, c_nation: chararray, c_region: chararray, c_phone: chararray, c_mktsegment: chararray);

LineFilt = FILTER Lineorder BY lo_discount == 6;

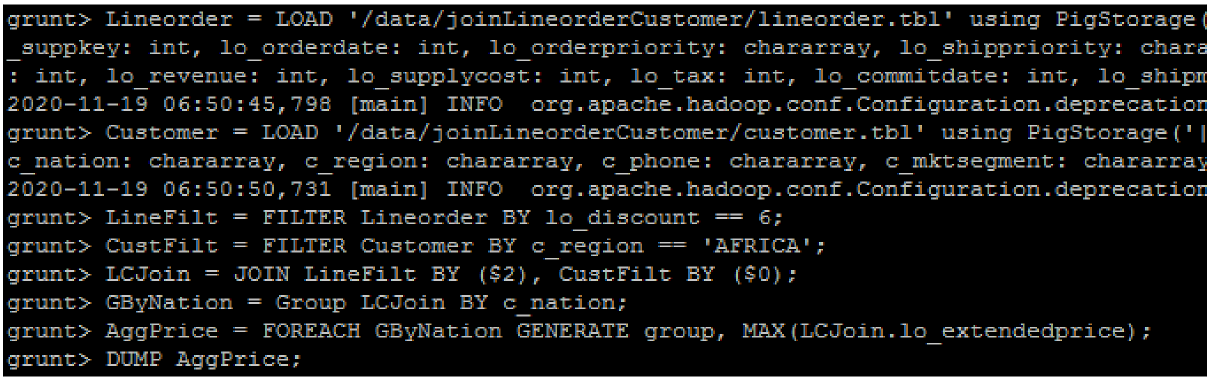
CustFilt = FILTER Customer BY c_region == 'AFRICA';

LCJoin = JOIN LineFilt BY ($2), CustFilt BY ($0);

GByNation = Group LCJoin BY c_nation;

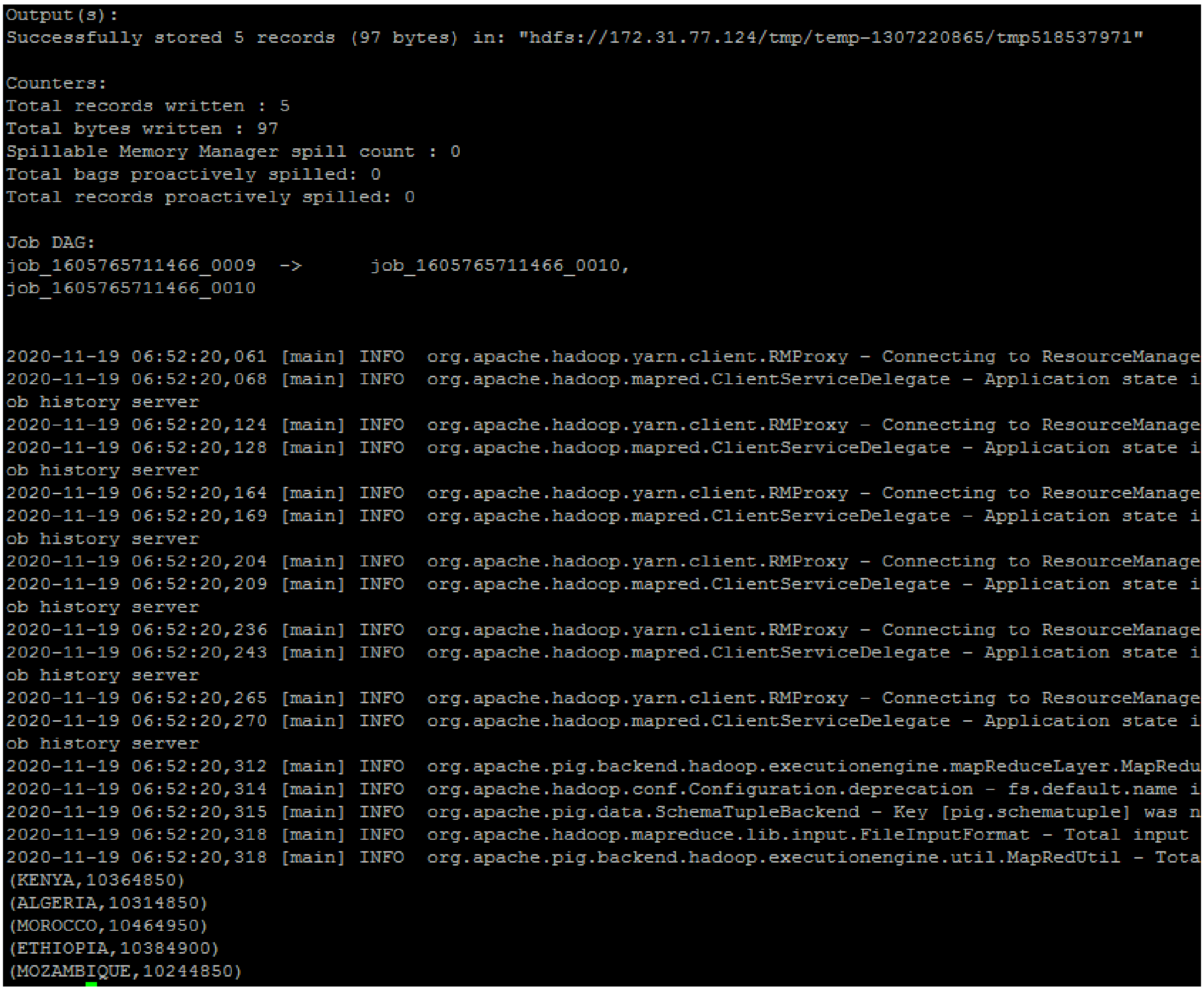
AggPrice = FOREACH GByNation GENERATE group, MAX(LCJoin.lo_extendedprice);

DUMP AggPrice;



```
grunt> Lineorder = LOAD '/data/joinLineorderCustomer/lineorder.tbl' using PigStorage(
_suppkey: int, lo_orderdate: int, lo_orderpriority: chararray, lo_shippriority: chara
: int, lo_revenue: int, lo_supplycost: int, lo_tax: int, lo_commitdate: int, lo_shipm
2020-11-19 06:50:45,798 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
grunt> Customer = LOAD '/data/joinLineorderCustomer/customer.tbl' using PigStorage('|
c_nation: chararray, c_region: chararray, c_phone: chararray, c_mktsegment: chararray
2020-11-19 06:50:50,731 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
grunt> LineFilt = FILTER Lineorder BY lo_discount == 6;
grunt> CustFilt = FILTER Customer BY c_region == 'AFRICA';
grunt> LCJoin = JOIN LineFilt BY ($2), CustFilt BY ($0);
grunt> GByNation = Group LCJoin BY c_nation;
grunt> AggPrice = FOREACH GByNation GENERATE group, MAX(LCJoin.lo_extendedprice);
grunt> DUMP AggPrice;
```

Ouput:

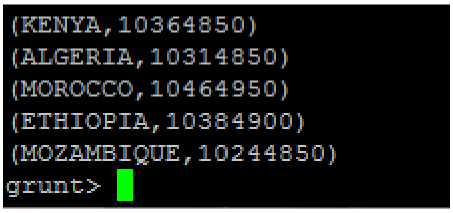


```
Output(s):
Successfully stored 5 records (97 bytes) in: "hdfs://172.31.77.124/tmp/temp-1307220865/tmp518537971"

Counters:
Total records written : 5
Total bytes written : 97
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1605765711466_0009  ->      job_1605765711466_0010,
job_1605765711466_0010


2020-11-19 06:52:20,061 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManage
2020-11-19 06:52:20,068 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state i
ob history server
2020-11-19 06:52:20,124 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManage
2020-11-19 06:52:20,128 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state i
ob history server
2020-11-19 06:52:20,164 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManage
2020-11-19 06:52:20,169 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state i
ob history server
2020-11-19 06:52:20,204 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManage
2020-11-19 06:52:20,209 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state i
ob history server
2020-11-19 06:52:20,236 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManage
2020-11-19 06:52:20,243 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state i
ob history server
2020-11-19 06:52:20,265 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManage
2020-11-19 06:52:20,270 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state i
ob history server
2020-11-19 06:52:20,312 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
2020-11-19 06:52:20,314 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name i
2020-11-19 06:52:20,315 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was n
2020-11-19 06:52:20,318 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input
2020-11-19 06:52:20,318 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Tota
(KENYA,10364850)
(ALGERIA,10314850)
(MOROCCO,10464950)
(ETHIOPIA,10384900)
(MOZAMBIQUE,10244850)
```



```
(KENYA,10364850)
(ALGERIA,10314850)
(MOROCCO,10464950)
(ETHIOPIA,10384900)
(MOZAMBIQUE,10244850)
grunt>
```