Pig – Lineorder data

SELECT lo_discount, COUNT(lo_extendedprice)
FROM lineorder
GROUP BY lo_discount;

[ec2-user@ip-172-31-75-50 pig-0.15.0]$ cat lo_discount_count.pig
a1 = 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:long, lo_revenue:int, lo_supplycost:int, lo_tax:int, lo_commitdate:int, lo_shipmode:chararray);
a2 = GROUP a1 BY lo_discount;
a5 = foreach a2 generate group as lo_discount, COUNT(a1.lo_extendedprice) as cnt;
result = FOREACH a5 GENERATE lo_discount, cnt;
dump result;

```
[ec2-user@ip-172-31-75-50 pig-0.15.0]$ cat lo_discount_count.pig
a1 = LOAD '/user/ec2-user/lineorder.tbl' using PigStorage('|') AS (lo_orderkey:i
nt, lo_linenumber:int, lo_custkey:int, lo_partkey:int, lo_suppkey:int, lo_orderd
ate:int, lo_orderpriority:chararray, lo_shippriority:chararray, lo_quantity:int,
 lo_extendedprice:int, lo_ordertotalprice:int, lo_discount:long, lo_revenue:int,
 lo_supplycost:int, lo_tax:int, lo_commitdate:int, lo_shipmode:chararray);
a2 = GROUP a1 BY lo_discount;
a5 = foreach a2 generate group as lo_discount, COUNT(a1.lo_extendedprice) as cnt
;
result = FOREACH a5 GENERATE lo_discount, cnt;
dump result;
```

Output:

```
2020-10-24 18:26:21,383 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-10-24 18:26:21,383 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2020-10-24 18:26:21,395 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2020-10-24 18:26:21,395 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(0,544886)
(1,545834)
(2,546173)
(3,545293)
(4,545545)
(5,546395)
(6,544970)
(7,546192)
(8,544803)
(9,545309)
(10,545815)
2020-10-24 18:26:21,454 [main] INFO  org.apache.pig.Main - Pig script completed
in 1 minute, 3 seconds and 938 milliseconds (63938 ms)
```

The script completes in 1 minute 3 seconds.

SELECT lo_quantity, SUM(lo_revenue)
FROM lineorder

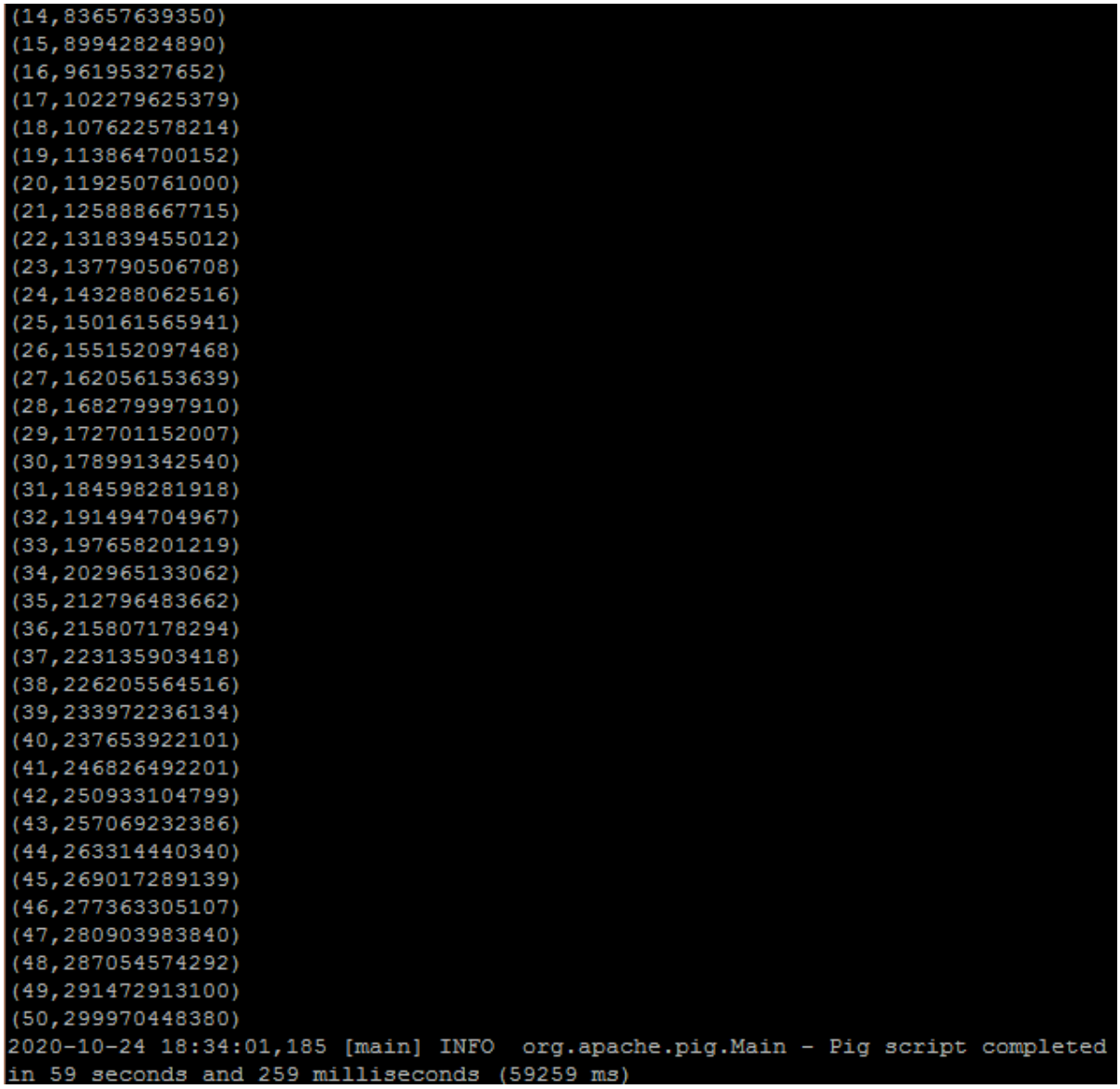
WHERE lo_discount > 6
GROUP BY lo_quantity;

[ec2-user@ip-172-31-75-50 pig-0.15.0]$ cat lo_revenue_sum.pig
LineOrder1 = 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:long, lo_revenue:int, lo_supplycost:int, lo_tax:int, lo_commitdate:int, lo_shipmode:chararray);
LineRevenue = FILTER LineOrder1 BY lo_discount > 6;
a2 = GROUP LineRevenue BY lo_quantity;
a5 = foreach a2 generate group as lo_quantity, SUM(LineRevenue.lo_revenue) as total_revenue;
sum_result = FOREACH a5 GENERATE lo_quantity, total_revenue;
dump sum_result;

```
[ec2-user@ip-172-31-75-50 pig-0.15.0]$ cat lo_revenue_sum.pig
LineOrder1 = LOAD '/user/ec2-user/lineorder.tbl' using PigStorage('|') AS (lo_or
derkey:int, lo_linenumber:int, lo_custkey:int, lo_partkey:int, lo_suppkey:int, l
o_orderdate:int, lo_orderpriority:chararray, lo_shippriority:chararray, lo_quant
ity:int, lo_extendedprice:int, lo_ordertotalprice:int, lo_discount:long, lo_reve
nue:int, lo_supplycost:int, lo_tax:int, lo_commitdate:int, lo_shipmode:chararray
);
LineRevenue = FILTER LineOrder1 BY lo_discount > 6;
a2 = GROUP LineRevenue BY lo_quantity;
a5 = foreach a2 generate group as lo_quantity, SUM(LineRevenue.lo_revenue) as to
tal_revenue;
sum_result = FOREACH a5 GENERATE lo_quantity, total_revenue;
dump sum_result;
```

Output:

```
2020-10-24 18:34:01,099 [main] INFO  org.apache.hadoop.mapreduce.lib.input.
nputFormat - Total input paths to process : 1
2020-10-24 18:34:01,099 [main] INFO  org.apache.pig.backend.hadoop.executio
ne.util.MapRedUtil - Total input paths to process : 1
(1,6019209421)
(2,11948248734)
(3,17957817737)
(4,23946021248)
(5,29746174935)
(6,35828498014)
(7,42217327722)
(8,48393723531)
(9,54148548879)
(10,59590403858)
(11,65707781195)
(12,71570674331)
(13,77435829473)
```



The script completes in ~59 seconds.