**CSC 555 Assignment 3**

**Yiyang Yang**

**Part 1**

a).

Mapper: Keys are the lo\_orderdate and d\_datekey, values are the sum of lo\_extendedprice, lo\_discount and d\_yearmonth.

Reducer: Show the sum of the lo\_extendedprice, for each lo\_orderdate same as d\_datekey, and then sort them when d\_yearmonth is equal to Jan1994, and lo\_discount is between 4 and 6.

b).

Mapper: Key is the variable d\_month, value is the count of the variable d\_sellingseason.

Reducer: Group all the d\_sellingseason into different d\_month, and then order them according to the correponding d\_sellingseason.

**Part 2**

a).

2 \* 75 + 2 \* 2500 / 60 = 150 + 83.3 = 233.3 mins

b).

There are 75 blocks divided into 30 blocks for 2 rounds, and 15 blocks left for the third round. 2 \* 3 = 6 mins. Time for reducer 2500 \* 2 / 30 / 60 = 2.8 mins. Time is 6 + 2.8 = 8.8 mins

c).

75 blocks are divided into 50 blocks for 1 round, and 25 blocks left for second round. 2 \* 2 = 4 mins. Time for reducer 2500 \* 2 / 50 / 60 = 1.7 mins. Time is 4 + 1.7 = 5.7 mins

d).

75 blocks are divided into 100 block for 1 round. 2 \* 1 = 2 mins. Time for reducer 2500 \* 2 / 75 /60 = 1.1 mins. Time is 2 + 1.1 = 3.1 mins

e).

I don’t think combiner will affect the runtime of this job, since combiner is used on reducer, the runtime of reducer on this problem will not be affected by the combiner that much.

**Part 3**

a).

i).

I think HDFS will reform the clusters with three factors.

ii).

I think MapReduce engine will store the data in the memory and then restart the engine when node failure.

b).

I think local file system is the place Mapper store the output before sent to Reducer.

**Part 4**

a). After compressed A4;B3;A5;B4;A4, compression ratio is 14/20 = 0.7.

b). After compressed ABCD5, compression ratio is 5/20 = 0.25.

c). Compressed ratio is 1.

**Part 5**

‘Swap.py’

#!/usr/bin/python

import sys

for line in sys.stdin:

line = line.strip().split('\t')

name = line[1]

first, last = name.split(‘\t’)

print '\t'.join([last, first])

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 ‘|’ STROED AS TEXTFILE;

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

ADD FILE /home/ec2-user/swap.py;

SELECT TRANSFROM(p\_name) USING ‘python swap.py’ AS (p\_name) FROM part;

CREATE TABLE PartX(

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,

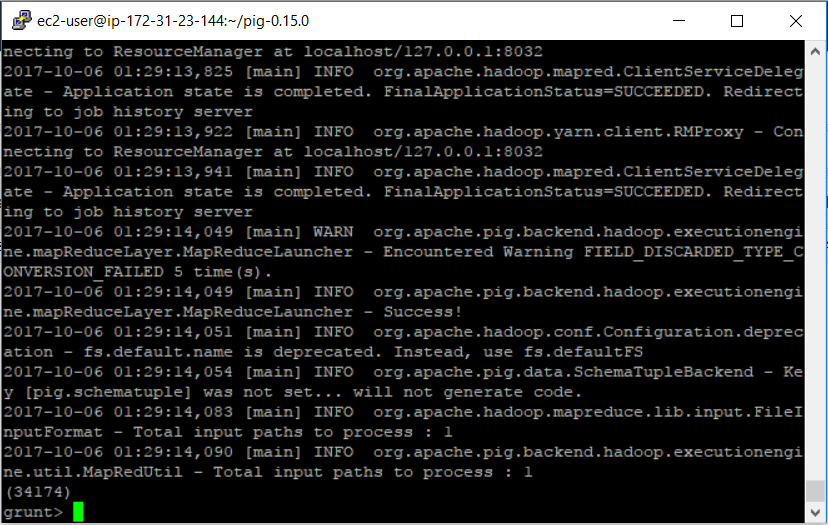
p\_container varchar(10)

)

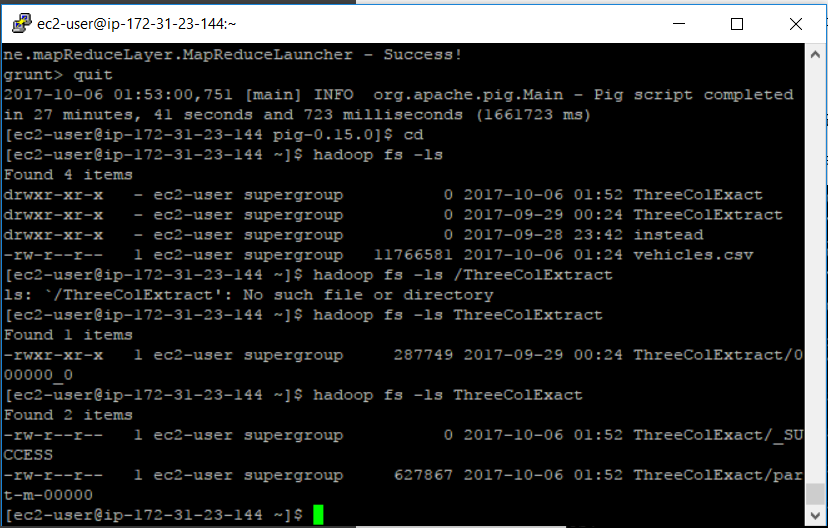
ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘\t’ STROED AS TEXTFILE;

INSERT OVERWRITE TABLE PartX SELECT TRANSFORM(p\_name) USING ‘python swap.py’ AS (p\_name) FROM part;

**Part 6**



There are 34174 rows in the table.



The size of the new file is 627867 bytes.