Assignment #3
CSC 555
Zehong Zhuang

1)

a)

Mapper1 receives lineorder through frameworks;
Mapper1 reads blocks of files and create pair-values;
Mapper1 (lo_revenue: lo_orderdate, lo_discout) where lo_discount between 5 and 7;
Mapper1 delivers pairs-values to reducers;
Reducer1 receives pair-values and produce outputs-1;
Mapper2 receives dwdate through frameworks;
Mapper2 reads blocks of files and create pair-values;
Mapper2 produces pair-value (dwadate: d_yearmonth) where d_yearmonth = 'Jan 1994';
Mapper2 delivers pair-values to reducers;
Reducers2 produce outputs-2;
Mapper3 receives output-1 and output-2;
Mapper3 produces pair values where lo_orderdate=d_datekey (lo_orderdate:lo_revenue);
Mapper3 delivers pair-values to reducer3;
Reducer3 receives outputs-3 and create outputs-3;
Mapper4 receives output-3;
Mapper4 receives output-3 and produce pair-values(lo_revenue:?)
Reducer4 performs sum function and outputs sum(lo_revenue)

b)

Mapper1 receives dwdate through frameworks;
Mapper1 reads blocks of files and create pair-values;
Mapper1 produces (d_month: d_sellingseason:);
Reducer1 receives pair-values and perform group by functions;
Reducer1 produces the outputs;
Mapper2 receives output-1 through frameworks;
Mapper2 produces pair-values (d_month: d_sellingseason);
Reducer2 receives pair-values;
Reducer2 perform distinct, count function and product output-2 (d_month, d_sellingseason);

2)

a) 120/1 min+4000/1 second=120M mins+66.7 mins=186.7 mins
b) 120/10 mins + 4000/10 second=12 mins+6.7 mins = 18.7 mins
c) 120/30 mins + 4000/30 second=4 mins + 2.2 mins = 6.2 mins
d) 120/50 mins + 4000/50 seconds=2.4 mins + 1.3 mins = 3.7 mins

e) Combiner would put the pair values with the same key together such as {key:1,2,3}. Reducer won't need to perform actions on same key again and again

f) The higher the replication factors, the more files a node can have. Hence, mappers will have access to more files and don't need to transform files from other disks. This way will save some time.

3)

a)

i) HDFS will look for the locations of other copies of the file which node fails to process

ii) It will look for another available node which has a copy of the file to perform MapReduce.

b) Memory

c) Reducers won't have input to perform sort and reduce function unless have all the inputs.

4)

```
[ec2-user@ip-172-31-5-249 ~]$ cat assignment3.py
#!usr/bin/python
import sys

for line in sys.stdin:
        line = line.strip()
        column = line.split('\t')

        col = column[0:1] + column[1].split(' ') + column[2:6] + column[6].split
(' ')+ column[7:8]
        print('\t'.join(col))
Time taken: 0.158 seconds
hive> CREATE TABLE Part(partkey int, name varchar(30), mfgr varchar(10),category
 varchar(10), brand varchar(10), color varchar(15),type varchar(25),size int,con
tainer varchar(10)) ROW FORMAT  DELIMITED FIELDS TERMINATED BY '|' STORED AS TEX
TFILE;
OK
Time taken: 0.23 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/part.tbl' OVERWRITE INTO TABLE Part
;
Loading data to table default.part
OK
Time taken: 1.064 seconds
ED   BY  in serde properties specification
hive> CREATE TABLE TPart(partkey int, name1 varchar(20), name2 varchar(20), mfgr
 varchar(10), category varchar(10), brand varchar(10), color varchar(15), type1
varchar(15), type2 varchar(15), type3 varchar(15), size int, container varchar(1
0)) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
OK
Time taken: 0.071 seconds
hive> add file /home/ec2-user/assignment3.py;
Added resources: [/home/ec2-user/assignment3.py]
```

```
Time taken: 0.188 seconds
hive> INSERT OVERWRITE TABLE TPart SELECT TRANSFORM (partkey, name, mfgr, catego
ry, brand, color, type, size, container) USING 'python assignment3.py' AS (partk
ey, name1, name2, mfgr, category, brand, color, type1, type2, type3, size, conta
iner) FROM Part;
```

5)

34174 rows

```
Desktop — ec2-user@ip-172-31-5-249:~/pig-0.15.0 — ssh -i CSC555v2.pem...
necting to ResourceManager at localhost/127.0.0.1:8032
2019-02-11 14:54:34,617 [main] INFO  org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2019-02-11 14:54:34,693 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at localhost/127.0.0.1:8032
2019-02-11 14:54:34,705 [main] INFO  org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2019-02-11 14:54:34,790 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 5 time(s).
2019-02-11 14:54:34,793 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2019-02-11 14:54:34,795 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2019-02-11 14:54:34,796 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2019-02-11 14:54:34,820 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2019-02-11 14:54:34,821 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(34174)
grunt>
```

File exits; Size is 0 bite.

```
[ec2-user@ip-172-31-5-249 ~]$ hadoop fs -ls /user/ec2-user/
Found 3 items
drwxr-xr-x   - ec2-user supergroup          0 2019-02-12 10:20 /user/ec2-user/Th
reeCol
drwxr-xr-x   - ec2-user supergroup          0 2019-02-12 10:18 /user/ec2-user/ou
t
-rw-r--r--   1 ec2-user supergroup   11766581 2019-02-11 14:52 /user/ec2-user/ve
hicles.csv
grunt> VehicleData = LOAD '/user/ec2-user/vehicles.csv' USING PigStorage(',')
>> AS (barrels08:FLOAT, charge120:FLOAT, city08:FLOAT);
2019-02-12 09:52:04,893 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> DESCRIBE VehicleData;
VehicleData: {barrels08: float,charge120: float,city08: float}
```

```
grunt> cat ThreeCol;
,,
15.689436,0.0,0.0
29.950562,0.0,0.0
12.19557,0.0,0.0
29.950562,0.0,0.0
17.337486,0.0,0.0
14.964294,0.0,0.0
13.1844,0.0,0.0
13.73375,0.0,0.0
12.657024,0.0,0.0
13.1844,0.0,0.0
12.657024,0.0,0.0
15.689436,0.0,0.0
13.73375,0.0,0.0
```