**CSC 555: Mining Big Data**

Project phase 1

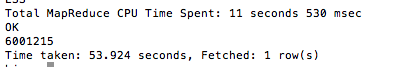
Name: LI HUANG

Student ID: 1506873

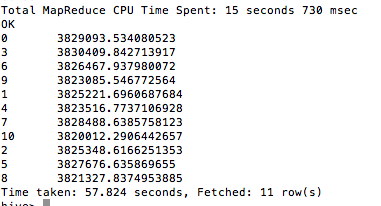
**1) hive**

Run the following seven (0.1, 0.2, 0.3, 1.2, 1.3 and 2.1, 2.2) queries in Hive and record the time they take to execute:

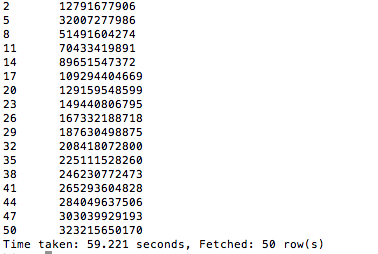
**0.1**

****

**0.2**

****

**0.3**

****

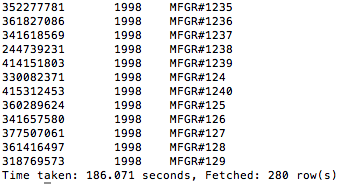
**1.2**

**Macintosh HD:Users:cathy:Desktop:Screen Shot 2017-04-24 at 10.22.00 AM.png**

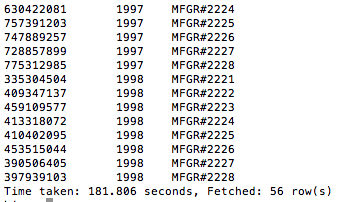
**1.3**

**Macintosh HD:Users:cathy:Desktop:Screen Shot 2017-04-24 at 10.23.42 AM.png**

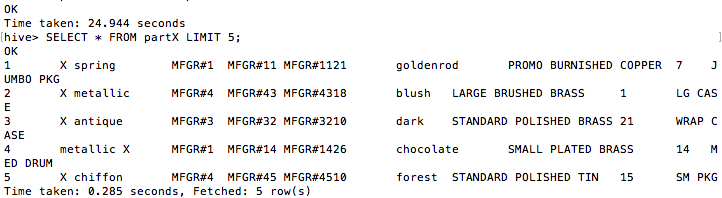
**2.1**

****

**2.2**

****

Use Hive user defined function (i.e., SELECT TRANSFORM, Slide #17 from Lecture4, weekday mapper is available here: <http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/weekday_mapper.py>) to perform the following transformation on Part table (creating a PartX table): in the 2nd column/p\_name replace the shorter word by X in the new table. For example, rose moccasin would become X moccasin or honeydew dim would be honeydew X. Please remember that split/print is always using ‘\t’ even if the source data is |-separated.

****

python code:

#!/usr/bin/python

import sys, datetime

for line in sys.stdin:

line = line.strip()

vals = line.split('\t')

ts = vals[1]

tswords=ts.split()

sht=min(tswords, key=len)

tswords =[word.replace(sht,'X') for word in tswords]

vals[1] =' '.join(tswords)

print '\t'.join(vals)

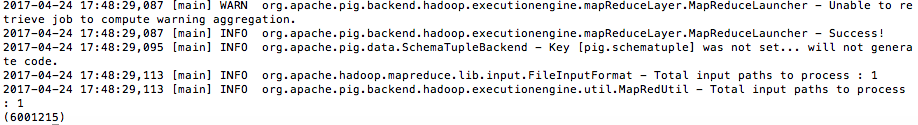
**2) pig**

**0.1**

lineorderG = GROUP lineorder ALL;

count = FOREACH lineorderG COUNT(lineorder.lo\_revenue);

DUMP count;

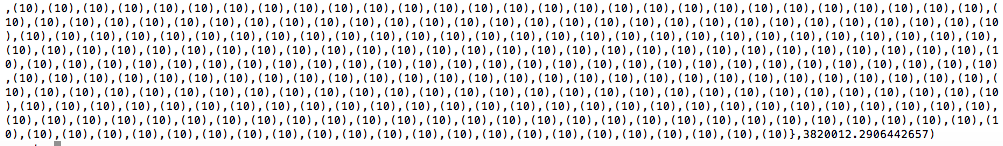
****

**0.2**

LG = GROUP lineorder BY lo\_discount;

avgle = FOREACH LG GENERATE lineorder.lo\_discount, AVG(lineorder.lo\_extendedprice);

DUMP avgle;

****

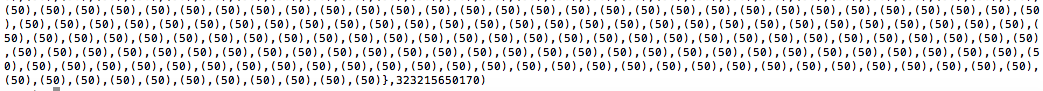
**0.3**

lineorder\_F = FILTER lineorder BY lo\_discount < 4;

GL = GROUP lineorder\_F BY lo\_quantity;

lineorder\_S = FOREACH GL GENERATE lineorder\_F.lo\_quantity, SUM(lineorder\_F.lo\_revenue);

DUMP lineorder\_S;



**3) streaming**

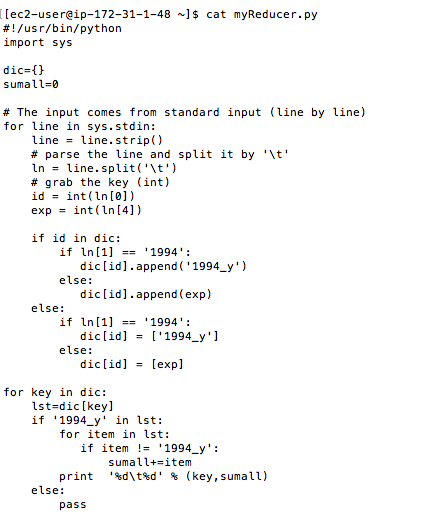
Implement query **1.1** using Hadoop streaming with python.

I am using one mapper and one reducer to finish this steaming as below:

myMapper.py code:



myReducer.py code:



Apply Mapper and Reducer on the two datasets dwdate.tbl & lineorder.tbl, ouput the desired result: group by lo\_orderdate, and lo\_orderdate is also the key, sum(lo\_extendendedprice) in the group as value.

The result is as following:

