`

Big Data
PES University
Bangalore

Extra Hands-On questions in MapReduce

PROBLEM STATEMENTS

1. Find the number of movie rating existence. I mean to say give the count of movies with rating 5 ,4 and so on.
   Dataset is available in the name of movielens.txt.

OUTPUT EXAMPLE

Movie	Number of Cars that use Gas
Rating 5	10
Rating 4	12

The table is solely for representational purposes. We expect the actual output to be in a text file with each line of the answer having the pair <Rating> <number>.

2. Find the Top 10 friend recommendations for each user.
   For a user X who is the best friend to recommend.
   Dataset is available in the name of friends.txt. where numbers represent the number of users and the comma separated values represent the list of friends.

OUTPUT EXAMPLE

```
9978    12361|12650|10275|14288|10498|11383|19469|22097|22946|23065
9979    21308|25866|8457|9970|10738|10824|10909|11299|11612|11635
998     1000|1001|1002|1003|1004|1005|1006|1007|1008|1009
9980    11612|14246|17435|351|4981|5069|52|7611|7651|8760
9981    30409|44358|127|13854|13913|14004|14105|14173|14182|14872
9982    31270|17999|30818|40610|40896|12060|12067|12179|12187|12195
9983    17199|4509|1973|17193|23993|37830|7327|4522|10622|18035
9984    4498|4509|11757|1421|10620|17198|17270|17272|1772|1382
9985    14411|16185|20050|22147|33461|40358|42471|11191|45825|522
9986    522|579|30860|4839|4841|14390|14411|15761|20500|21273
9987    34485|13134|37941|9992|34642|13478|13877|34299
9988    35667
9989    34299|34485|34642|37941|9992|13134|13478|13877
999     950|1000|1001|1002|1003|1004|1005|1006|1007|1008
9990    13134|13478|13877|34299|34485|34642|37941
9991    13478|13134|13877|34299|34485|34642|37941|9992|9993|9994
9992    9987|35667|9989|9991
9993    13134|13478|13877|34299|34485|34642|37941|9991
9994    13134|13478|13877|34299|34485|34642|37941|9991
9995    37356|37188|37135|10097|36679|40748|36905|37812|37597|10103
9996    36679|36854|10096|10000|44050|36669|37156|10008|36703|36765
```

3. Write a Map Reduce Program to count how many words belong to each of the following four length categories:
   tiny: 1 letter — small: 2–4 letters — medium: 5–9 letters — big: more than 10 letters
   Dataset is available in the name of word.txt.
   Output Example:

word	category
and	Medium
Reason	big

You are working for a large supermarket chain. They are interested in how much money their customers spend, on average, at different hours of the day. They give you three large tab-separated values (TSV) files containing millions of records as follows:

1: ReceiptItems.tsv

Receipt ID	Item ID
R1401	I306
R1401	I306
R1401	I504
R1402	I007
R1402	I306
R1403	I306
R1403	I504
...	...

2: ReceiptTimes.tsv

Receipt ID	Time
R1403	19:00
R1401	18:59
R1402	19:01
...	...

3: ItemDetails.tsv

Item ID	Name	Price ($)
I306	Zanahoria 500g	500
I504	CocaCola 3L	1400
I007	Comfort	1200
...	...	

In these tables, the Receipt ID column corresponds to an individual transaction, where a customer pays for their items. The Item ID corresponds to a unique identifier for each type of item. The same item may appear multiple times in a transaction. So in the table above, in transaction R1401, a customer buys 2 × Zanahoria 500g ($500) and 1 × CocaCola 3L ($1400), spending a total of $2400 at time 18:59. Likewise transaction R1402 spends $1700 at time 19:00 and transaction R1403 spends $1900 at time 19:01.

Given this input, your manager wants you to compute the total spent by customers of the supermarket chain each hour of the day. For example, just considering the three transactions above, the answer would be:

Output

Hour	Total
...	...
18:00–18:59	$2400
19:00–19:59	$3600
...	...

The output should then be sorted in descending order by total value.

4. Implement the same using the Map Reduce. I suggest you build a simple TSV for performing these operations.
5. You have just been hired by the NCDC2 to help with analyzing their large amounts of weather data (about 1GB per year). The NCDC produces CSV (Comma-Separated Values) files with worldwide weather data for each year.

   Each line of one of these files contains:

   • The weather station's code.

   • The date, in the ISO-8601 format.

   • The type of value stored in that line. All values are integers. TMIN (resp. TMAX) stands for minimum (resp. maximum) temperature.

   Temperatures are expressed in tenth of degrees Celsius. AWND stands for average wind speed, and PRCP stands for precipitation (rainfall), etc. Several other types of records are used (TOBS, SNOW, ...). • The next field contains the corresponding value (temperature, wind speed, rainfall, etc.) • All lines contain five more fields that we won't use in this exercise. We will work on the CSV file for 2013, which has been sorted by date first, station second, and value type third, in order to ease its parsing.

It can be found at the following location on the server: /cs/bigdata/datasets/ncdc-2013-sorted.csv
Here is a sample of that file: ... FS000061996,20130102,TMAX,206,,,S,
FR000007650,20130102,PRCP,5,,,S, FS000061996,20130102,TMIN,128,,,S,
FR000007650,20130102,TMAX,111,,,S, GG000037279,20130102,TMAX,121,,,S,
FR000007747,20130102,PRCP,3,,,S, GG000037308,20130102,TMAX,50,,,S,
FR000007747,20130102,TMAX,117,,,S, GG000037308,20130102,TMIN,-70,,,S,
FR000007747,20130102,TMIN,75,,,S, GG000037432,20130102,SNWD,180,,,S,
FR069029001,20130102,PRCP,84,,,S, GG000037432,20130102,TMAX,15,,,S,
FR069029001,20130102,TMAX,80,,,S, GG000037432,20130102,TMIN,-105,,,S,
FS000061996,20130102,PRCP,0,,,S, ...
As you can see, not all stations record all data. For instance, FR069029001 only recorded rainfall and maximum temperature on 01/02/2013. Not all stations provide data for every day of the year either.
The NCDC wants to plot the difference between the maximum and the minimum temperature in Central Park for each day in 2013. Write a The Map task(s) that send(s) (<tmin>, <tmax>) pairs to the Reducer. Temperatures are converted to degrees Celsius. The output will be:
(4.4, -3.3)
(6, -5.6)
(0, -4.4)
For each key/value pair, the Reduce task substracts the minimum temperature from the maximum temperature, converts it to degrees, and writes the result to a file.
Comment if the approach given is the right approach.