Assignment 4
CIS 610 Big Data and Data Science, Fall 2016
due 11:59 pm, Friday, Nov 18th

1. See a relational DB on a company at `http://www.cs.uoregon.edu/classes/16F/cis610bdds/assignments/en_6e_fig_3-6.pdf` .

a) If you can choose to put all or some of tables into a Key-Value database, will you gain any benefit to do so? If yes, give one example key-value pair to show your idea of implementation.

b) If you can choose to put all or some of tables into a Document database, will you gain any benefit to do so? If yes, give one example Document (e.g., in JSON) to show your idea of implementation.

c) If you can choose to put all or some of tables into a Column-Family database, will you gain any benefit to do so? If yes, give one example Column Family to show your idea of implementation.

d) If you can choose to put all or some of tables into a Graph database, will you gain any benefit to do so? If yes, give a portion of the graph to show your idea of implementation.

2. Assume Mark created a Document database (e.g., CouchDB) for the company DB in 1). Then he wants to add more information in. Here is a possible employee document, which incorporates the works_on information. A project document would be separate.

```
{
   "fname": "Alicia",
   "minit": "J",
   "lname": "Zeleya",
   "ssn": "999887777",
   "bdate": {
    "year": 1961, "month": 1, "day": 19
   },
   "address": "3321 Castle, Spring, TX",
   "sex": "F",
   "salary": 25000,
   "works_on": {
       "10": 10,
       "30": 30
   },
   "superssn": "987654321",
   "dno": 4
}
```

Based on the above new document, you may figure out roughly what is Mark's design for this document database. Please describe example Map and Reduce steps to explain how to add a map-reduce function to determine the average number of hours that an employee works on each project. The idea is that the map part should emit a pair for each works_on entry (note: it must have one to be an interesting employee document). The key would be the project number and the value the number of hours worked. The reduce section will average the values for each key. You do not need to produce the project name.

3. Explain the Map step and Reduce step for the following relational DB operations.

a) Projection; b) Selection; c) Group by; d) Sort-merge join.

4. Suppose that the data mining task is to cluster the following nine points (with (x,y) representing location) into three clusters:

A_1(3,10), A_2(3,5), A_3(9,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(2,2), C_2(5,9), C_3(6,8)

Suppose initially we assign A_1, B_1 and C_1 as the center of each cluster, respectively. Please add a Map-reduce function for the K-means algorithm. Show the results for the first two iterations and explain how Map-reduce can help.

5. A database has six transactions. Let min_sup = 50%.

TID	items_sold
T001	A, B, C, D, E, F
T002	B, H, S, C, W, T
T003	A, U, O, F, W, D
T004	O, A, B, C, F, X
T005	O, A, C, D, F, Y
T006	B, C, X, E, W, Z

Please add a Map-reduce function for the Apriori algorithm to generate all frequent itemsets. Show the results for each step and explain how Map-reduce can help.

To turn in by emails: Email your answers to **dou@cs.uoregon.edu**. A pdf file is prefered. If you are using Word, you should be able to convert your word file to a pdf file.

To turn in by paper version: Ask Adriane or Cheri to put your answers to Prof. Dejing Dou's mailbox.