Assignment 2
CIS 610 Big Data and Data Science, Fall 2016
due 11:59 pm, Friday October 21st

1. Explain the difference between data replication in a distributed system to the maintenance of a remote backup site? Compare the advantages and disadvantages of the data replication and data fragmentation approaches in a distributed system. When is it useful to have replication or fragmentation of data?

2. Consider the relations:

$employee(name, address, salary, plant\_number)$
$machine(machine\_number, type, plant\_number)$

Assume that the *employee* relation is fragmented horizontally by $plant\_number$, and that each fragment is stored locally at its corresponding plant site. Assume that the *machine* relation is stored in its entirely at the Armonk site. Describe a good strategy for processing each of the following queries.

a. Find all employees at the plant that contains machine number 1130.

b. Find all employees at plants that contain machines whose type is

"milling machine."

c. Find all machines at the Almaden plant,

d. Find employee $\bowtie$ machine.

3. For each of the strategies of problem 2, state how your choice of a strategy depends on:

a. The site at which the query was entered.

b. The site ate which the result is desired.

4. What is the difference and relationship between distributed database and data warehouse?

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.