Assignment 1
CIS 670 Data Science, Winter 2019
due 11:59 pm, Monday, January 28th

1. Give a real life example of Big data application that has not been mentioned in the slides. (The slides mentioned CERNs Large Hydron Collider (LHC), EarthScope, LIGO etc). Explain why your example have three V's as challenges? If your example does not have three V's, you need to give more than one example to cover them (e.g., one example has the Volume and Velocity challenges, and another example has the Volume and Variety challenges).

2. What are the relationships between parallel databases and Big data with respect to three V's?

3. Consider join processing using symmetric fragment and replicate with range partitioning. How can you optimize the evaluation if the join condition is of the form $|r.A - s.B| \leq k$, where k is a small constant? Here, $|x|$ denotes the absolute value of x. A join with such a join condition is called a band join.

4. Describe a good way to parallelize each of following:

   a. The difference operation

   b. Aggregation by the **count** operation

   c. Aggregation by the **avg** operation

   d. Left outer join, if the join condition involves only equality

   e. Full outer join, if the join condition involves comparison other than equality

5. The attribute on which a relation is partitioned can have a significant impact on the cost of a query.

   a. Given a workload of SQL queries on a single relation, what attributes would be candidates for partitioning?

   b. How would you choose between the alternative partitioning techniques, based on the workload?

   c. Is it possible to partition a relation on more than one attribute? Explain your answer?

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