Spring 2013
W03 (26 points)
Due: May 15 @ noon

Database Writing Activity

Learning objectives

- To become familiar with the MapReduce computational model

Reading

- Please read the following paper, which can be found on the course website:

  Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 1 (January 2008), 107-113. DOI=10.1145/1327452.1327492 http://doi.acm.org/10.1145/1327452.1327492

Reading questions

1. What kind of data processing problems led to the invention of MapReduce? (2 points)

2. What does the term "fault tolerant" mean? (2 points)

3. In your own words describe as precisely as you can the programming model for MapReduce. (2 points)

4. Describe the *Map* function, the *Reduce* function, and how the *Map* and *Reduce* functions work together in order to complete a programming task. (6 points)

5. Suppose you have a large number of text documents and you want to compute a frequency distribution of the length of the words in each document. For example, the number of words of length 1, 2, and 23 might be 100,001, 200,000, and 4 respectively, as shown in this table:

Length of Word	**Number of Occurrences**
1	100,001
2	200,000
...	
23	4

   Write the pseudo code for the *map* function and for the *reduce* function in order to compute this result. Please (a) show the arguments and the types for each of these functions; and (b) show what each function produces. (6 points)

6. The *MapReduce* programming model depends on partitioning the data inputs. Please explain. (2 points)

7. Why is data locality important in MapReduce implementations? (2 points)

8. What is the name of the open source system that provides an implementation of MapReduce? (1 point)

9. What is Amazon Web Services? Does AWS provide an implementation of MapReduce? (3 points)