Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

Software Technologies
Assignment 1 (Hadoop)

Q1. Familiarize yourself with Hadoop installation and the 'word count' program available at http://wiki.apache.org/hadoop/WordCount. Modify it to find the top k frequent words for any given input parameter k.

Q2. Implement a distributed sort on Hadoop.

Hadoop Terasort approach: Your implementation could use the following quicksort approach as used by Hadoop Terasort. Instead of the standard single pivot for quicksort, randomly sample N pivots from the input keys. Now use these pivots to partition the input and send these partitions to reducers, which are sorted locally using inbuilt sort function. Report the sorting time for 10 Gig of records.
Ref:

1. http://grepcode.com/file/repository.cloudera.com/content/repositories/releases/com.cloudera.hadoop/hadoop-examples/0.20.2-737/org/apache/hadoop/examples/terasort/TeraSort.java?av=f
2. http://sortbenchmark.org/Yahoo2013Sort.pdf
3. http://www.slideshare.net/tungld/terasort

Input for sorting: Download 'gensort' program (http://www.ordinal.com/gensort.html) to generate sort input. The input is a collection of string records, each of 100 bytes, where the initial 10 bytes is the key. These strings should be sorted based on these 10 byte keys. Use 'valsort' available in the same gensort distribution to verify whether the output is a sorted collection. Check http://sortbenchmark.org/ for the current sort benchmarks.