1. **Programming Languages:**

   C/C++

   *(Note: This is a hard requirement and no other languages will be accepted.)*

2. **Datasets:**

   Use the DBGEN program of the TPC-H Benchmark to generate the LINEITEM table of the TPC-H schema, with the Scale Factor (SF) set to 1. The LINEITEM table in this configuration will have 6,001,215 tuples, where each tuple has 16 attributes.

   - The DBGEN program can be downloaded at the following URL:
     
     http://www.tpc.org/tpch/spec/tpch_2_12_0_b5.zip

   - The schema of LINEITEM can be found at Page 15 in the tpch2.12.0.doc, which can be downloaded at the following URL:
     
     http://www.tpc.org/tpch/spec/tpch2.12.0.doc

3. **Implementation Description:**

   **Task 1:** Use two alternatives of External Merge Sort (EMS), i.e., EMS without / with Double Buffering, to sort the LINEITEM table. Please
compare the number of passes and the running times of the two alternatives, and give your explanation for possible differences in the comparisons. (Tips: You should think about using multithreading program to implement Double Buffering.)

Consider two cases:

1) Using 12 buffer pages (That means there are only 12 buffer pages available in the whole memory, including both input and output buffers) in memory for the external merge sort;

2) Using 130 buffer pages (the same as above) in memory for the external merge sort.

Note: In EMS without Double Buffering, please do \((2K+1)\)-way merge-sort, where \(K\) is 5 or 64. In EMS with Double Buffering, please do \(K\)-way merge-sort, where \(K\) is 5 or 64. In the implementation, each buffer page occupies 8K bytes. The pair of the L_ORDERKEY and the L_PARTKEY (You can see more details in the document of TPC-H.) of the LINEITEM table is assumed to be the sort key in the external merge sort.

Task 2: Organizing the sorted LINEITEM table into disk pages

(This is an optional task. If you have enough time, you can do it and then you can earn a bonus point. Or please ignore it.)

Please use the page format for storing variable-length records to
organize the LINEITEM table sorted in Task 1. In the implementation, each disk page occupies 1K bytes. For each page we maintain a directory of slots, with a \(<\text{record offset}, \text{record length}\>\) pair per slot. Both “record offset” and “record length” are 4 bytes wide.

4. Technical Report:

Write the design of your project, including major data structures and major algorithms outlined in pseudo-code. You can use either English or Chinese to write the report.

5. Deadlines:

1) The Deadline is 12 pm, Nov 28, 2010.

Note: For each deadline, follow the requirements of Assignment Submission and submit your whole package (source code; compile scripts; technical report) to the chief of study affairs.