CS3223: Database Management Systems
Tutorial 6
(Week 8, March 2022)

1. Consider the relations R1(A,B,C), R2(C,D,E) and R3(E,F), with primary keys A, C, and E respectively. Assume that R1 has 10000 tuples, R2 has 15000 tuples and R3 has 7500 tuples. For simplicity, assume that all attributes have the same size, and that each page can contain 10 tuples of R1. Consider the query: R1 ⋈ R2 ⋈ R3.
   a. Assume that the data are uniformly distributed, estimate the result size of the query.
   b. List all possible plans assuming only left-deep search space is considered. You may assume that cross products are to be avoided.
   c. Compute the cost for each of the above plans you listed to determine the optimal plan. For simplicity, you may assume that only the nested-block join is supported, the buffer size is 100 pages, and all intermediate results are to be stored in secondary storage.

2. Discuss the motivation for the "avoid cross products" heuristics in multi-join query optimization. Does it always lead to better plans? Justify your answer.

3. Starting with an expression $\pi_L(R(a, b, c) \bowtie S(b, c, d, e))$, push the projection down as far as it can go if L is:
   - $b + c \rightarrow x, c + d \rightarrow y$
   - $a, b, a + d \rightarrow z$

4. We wish to join the relations R(a,b), S(b,c) and T(c,d) under the following assumptions:
   a. There are B buffers available to hold pages of data from these three relations.
   b. The relations occupy R, S, and T pages, respectively.
   c. R is stored sorted by b; the other relations are unsorted.
   d. The following strategy is used to perform the join:
      i. Perform the first phase of two-phase multi-way merge sort on S. That is, as many times as necessary, load all buffers from S, sort the tuples on b, and write out the sorted sublist.
      ii. Load T entirely into main memory, using as many buffers as needed.
      iii. Merge (and join when appropriate) R and the sorted sublists of S, and compare each of the resulting tuples with the tuples of T. Any join tuple in the result is stored in an output buffer, and not counted among the B buffers available for this process.

   Provide an expression (inequality) in terms of B, S, T and R such that this sequence of steps can be carried out as described.