INF 551 – Fall 2016 (Afternoon)

Quiz 11: Query execution (10 points)

10 minutes

1.

Step 1: Use 1 block in memory as input buffer and rest 100 as output buffers. Choose a hash function h which can hash the values of join attribute(B) to 100 buckets. Read from R block by block into the input buffer and apply h to each record, the hash results are then placed into the corresponding output buffers and written to disk when buffers are full. Apply the same hash function h to records in blocks of S. The output in this step are 100 buckets containing records from R (R_1 – R_{100}, each roughly size of 50 blocks) and 100 buckets containing records from S (S_1 – S_{100}, each roughly size of 200 blocks). In this step, 1 pass through R and S, and 2B(R) + 2B(S) = 50000 block I/O's.

Step 2: Now, the records to be joined are hashed into buckets (partitions) with same subscripts (R_j and S_j). For each j from 1 – 100, if neither R_j nor S_j is empty, read the smaller one of R_j and S_j into memory (can occupy as many as 99 blocks, assume no smaller partition exceeds this capacity) and hash the records with a different function h' (to make the join more efficient, you may not mention hash again in this quiz), and then scan the other one block by block to check and join, use 1 block as output buffer. The output is the join result. In this step, roughly 1 full pass through R and S, and at most B(R) + B(S) = 25000 block I/O's (if some partitions of one data sets are empty, the matching partitions of the other need not be read, since no join could happen here).

To sum up, the algorithm makes **2 passes** through R and S, and requires **75000 block I/O's**.

Rubrics: need to mention for step 1 M-1 = 100 blocks are used for hashing and R and S are read into memory 1 block per time; for step 2 join within corresponding buckets, and read the smaller one (can assume R here) fully and the larger one 1 block per time.
-0.5 for each of the above not mentioned or wrong.
-2 if not mention hashing at all

2.

3B(R) + 3B(S) = 75000

3.
No.
Because B(R) + B(S) > M^2

Rubrics: -2 if say yes. -1 or -0.5 if reason not correct or partly correct.