CS3223: Database Management Systems
Tutorial 5
(Week 7, March 2022)

1. The GRACE hash join cannot produce answer tuples until the partitioning phase is completed. There is a variant called the Hybrid Hash Join that operates as follows. We pick an appropriate number of buckets, k, into which the two relations R and S are divided. Suppose S is the smaller relation. During the partitioning phase, one of the buckets from S, say bucket 0, is stored completely in main-memory buffers, while the other k-1 buckets of S are written out to disk as usual. When R is partitioned, those tuples that go into bucket 0 are not written out to disk, but are immediately used to probe for matches against the tuples of bucket 0 of S, which continue to reside in main memory. During the joining phase, only buckets 1-(k-1) of S need to be joined with the corresponding buckets of R (since the join between buckets 0 of R and S have already been performed). Assuming all records are uniformly distributed, what would the cost of this algorithm be? For simplicity, you can assume all partitions are of the same size.

2. Consider the join of two relations R and S whose keys are respectively rid and sid. A join index is a relation that contains the pairs (rid, sid) for the join results, i.e., the join index is a materialized join result that stores keys of matching tuples. In this way, if we want to perform the join on R and S, we only need to scan the join index, and for each pair retrieve the corresponding R and S tuples. Comment on the advantages and disadvantages of this method.

3. Consider the following relations:
   applicant(pid, cityid, income, gmat)
   location(cityid, country)

   In the relation applicant, we assume that an individual is uniquely identified by pid, resides in city, earns an annual salary given by income, and has a GMAT-score of gmat. The relation location identifies the country which a given city is in.

   a) Describe how the following query is evaluated, and its expected cost: "Find applicants that earn a salary greater than 60,000 and have GMAT scores higher than the average score of applicants from USA". You are not required to present the exact cost; rather, the components of the cost (e.g., scan the applicant relation once, etc.).

   b) Describe how the following query is evaluated, and its expected cost: "Find applicants that earn a salary greater than 60,000 and have GMAT scores higher than the average score of applicants from the same US city".