University of Wisconsin-Madison
Computer Sciences Department

Database Qualifying Exam
Fall 05

GENERAL INSTRUCTIONS

Answer each question in a separate book.

Indicate on the cover of *each* book the area of the exam, your code number, and the question answered in that book. On *one* of your books list the numbers of *all* the questions answered. Return all answer books in the folder provided. Additional answer books are available if needed.

Do not write your name on any answer book.

SPECIFIC INSTRUCTIONS

Answer all five (5) questions (NOTE: this is different from some previous years, when you were only asked to answer 4 of 5.) Before beginning to answer a question make sure that you read it carefully. If you are confused about what the question means, state any assumptions that you have made in formulating your answer. Good luck!

The grade you will receive for each question will depend on both the correctness of your answer and the quality of the writing of your answer.

Policy on misprints and ambiguities:

The Exam Committee tries to proofread the exam as carefully as possible. Nevertheless, the exam sometimes contains misprints and ambiguities. If you are convinced a problem has been stated incorrectly, mention this to the proctor. If necessary, the proctor can contact a representative of the area to resolve problems during the *first hour* of the exam. In any case, you should indicate your interpretation of the problem in your written answer. Your interpretation should be such that the problem is nontrivial.

1. SQL

Consider the following schema, which describes sales of parts to customers. In the Sales table, all sales of a given part to a given customer on a given day are already aggregated into a single row (leading to the composite primary key), and *total* is the sum of these sales. It might differ from the *price* for the part multiplied by the *quantity* because of promotional discounts, which might vary from day to day and across products and customers.

Customers(cid, cname, caddress); key: cid
Parts(pid, pname, color, price); key: pid
Sales(cid, pid, date, quantity, total); key: <cid, pid, date>

1) How much did each customer save on doughnuts in 2004 because of discounts?

2) Print the name(s) of the customer(s) who bought doughnuts on the most number of days in 2004.

3) For each customer, print the monthly moving average of the amount spent on doughnuts in 2004. (That is, for each customer, for each day in 2004, print the total amount spend on doughnuts by that customer on the—at most—30 preceding days in 2004. Hint: This query requires you to use SQL:1999's extensions to query sequence data. Exact syntax is not important, but the correct use of appropriate constructs/concepts is.)

2. B-Tree Locking

This question deals with the B-link protocol proposed by Lehman and Yao.

1) In the B-link protocol, writers lock but readers get no locks. In view of this, how does the protocol insure that readers find what they are looking for in the presence of concurrent updates from writers?

2) Joe Qualtaker thinks he has discovered a bug in the protocol. Consider the following timeline:

   i) A reader reads a leaf page, finds a record for the search key it is looking for, of the form (searchKey, RID). The RID refers to a record on a data page that is distinct from the leaf page.

   ii) The reader is preempted, and some writer deletes the record designated by the RID from the data page.

   iii) The reader wakes up, tries to "dereference" the RID, and gets an error.

   2a) Is Joe Qualtaker correct that this is a bug?
   2b) If Joe is correct, propose a fix to the bug. If he is wrong, explain what he is missing.

3. Join Algorithms

Consider a self-join of the form R Join_(R.a = R.b) R. Compare the performance of the following two algorithms to evaluate the join:

1) A symmetric hash join, with R as the input for both the right and left inputs of the symmetric hash join.

2) A hybrid hash join, again with R as the input for both the right and left inputs of the hybrid hash join.

Your answer should cover when (if ever) you expect each to dominate the other.

4. Data Mining

Clustering, Frequent Itemsets, and Decision Trees are three widely used techniques in data mining. Answer the following questions about them briefly.

1) Each of these three techniques is applicable in certain situations, but not others. For each technique, describe a problem scenario (i.e., nature of data, desired goal) when it is the best technique, and another scenario in which it is not effective.

2) Frequent Itemsets originated in the database community, but Clustering and Decision Tree construction have been widely studied in Statistics and Machine Learning, and are well understood. What issues relating to Clustering and Decision Trees might be candidates to address using ideas from database research? More generally, what are your thoughts on what database researchers bring to the table in "Data Mining" research?

3) Consider the problem of finding itemsets that are frequent in *both* the Walmart sales database and the Target sales database. Discuss ways in which you might adapt and/or apply the a priori algorithm, which finds frequent itemsets in a single database, to this problem.

5. Database System Architectures

In recent years, disks have been doubling in capacity approximately every 12 months while continuing to drop in price. If this trend continues, a commodity disk drive will soon hold 640 GB.

1) How does this trend impact the use of mirroring versus RAID-5 for achieving robustness with respect to failures?

2) What are the implications of this trend on database system architectures and obtaining high levels of performance?

3) What solutions do you envision to the problems that 1/2 TB disk drives will impose? What opportunities will such drives makes feasible?