`University of Wisconsin-Madison
Computer Sciences Department

Database Qualifying Exam
Fall 2007

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. Data Models

a) Compare and contrast the data model and query language extensions first proposed in the GEM paper with those later proposed by advocates of object-relational database systems such as POSTGRES.

b) What was a key technical limitation of the relational data model that lead to the design and development of database systems based on an object-oriented data model such as O2 and Object Store in the late 1980s and early 1990s?

c) Why do you believe that these systems failed to gain any significant market traction?

d) Based on what happened to such systems do you believe that XML is likely to replace the relational data model as the dominant model? Why or why not?

2. Bitmap Indexes

Consider the following radical idea. To store a table R(A1, A2, ..., Ak), we build bitmap indexes on each of A1, A2, ..., then "throw away" the original table, so we are left with only the bitmap indexes. Call this the "bitmap representation of R".

a) There is enough information in this bitmap representation to reconstruct the original table R (that is, no information has been lost.) Explain why this is the case.

b) Give a sketch of an algorithm to reconstruct the traditional (file of tuples) representation of R from the bitmap representation of R.

c) Which representation (bitmap or traditional file of tuples) would you expect to take more space on disk? If your answer is "it depends", explain on which factors this depends.

d) Compare the cost of answering a "select A_i from R" query (where A_i is some attribute of R) from the bitmap representation to the cost of answering the same query from the traditional file of tuples representation of R. You don't need to give precise formulas, but rather you should explain what properties of R and A_i impact the tradeoffs between the two approaches.

e) Compare the cost of answering a "select A_i ... A_k from R" query (where A_i ... A_k are some subset of attributes of R) from the bitmap representation to the cost of answering the same query from the traditional file of tuples representation of R. Again, you don't need to give precise formulas, but rather you should explain what properties of R and A_i ... A_k impact the tradeoffs between the two approaches.

f) Compare the cost of answering a "select A_i, ..., A_k from R where R.A_i = c" query (where A_i ... A_k are some subset of attributes of R, and c is some constant appearing in column A_i of R) from the bitmap representation to the cost of answering the same query from the traditional file of tuples representation of R. Once again, you don't need to give

precise formulas, but rather you should explain what properties of R, the A_i ... A_k, and c impact the tradeoffs between the two approaches.

3. Recovery

Briefly answer the following questions

a) Explain what happens when a checkpointing is taken in a database system that implements the ARIES recovery protocol?

b) Checkpointing can also be done as follows: Quiesce the system so that only checkpointing activity can be in progress. Then write out copies of all dirty pages and include the dirty page table and the transaction table in the checkpoint record.

What are the pros and cons of this approach versus the checkpointing approach used by ARIES?

c) What happens if a second begin-checkpoint record is encountered during the Analysis phase of the ARIES algorithm after a crash?

d) Can a second end-checkpoint be encountered during the Analysis phase?

e) Why is the use of CLRs important for the use of undo actions that are not physical in nature?

f) Give an example that illustrates how the paradigm of repeating history and the use of CLRs allow ARIES to support locks of finer granularity than a page.

4. Schema Integration

Let S be a relational database with two tables. The first table is HOUSES, with attributes **location**, **price**, and **agent-id**. This table has two tuples:

("Atlanta, GA"; 360,000; 32)
("Raleigh, NC"; 430,000; 15).

The second table is AGENTS, with attributes **id**, **name**, **city**, **state**, and **fee-rate**. This table has two tuples:

(32; "Mike Brown"; Athens; GA; 0.03)
(15; "Jean Laup"; Raleigh; NC; 0.04)

Let T be another relational database with a single table LISTINGS. This table has attributes **area**, **list-price**, **agent-address**, and **agent-name**. It has two tuples:

("Denver, CO"; 550,000; "Boulder, CO"; "Laura Smith")
("Atlanta, GA"; 370,800; "Athens, GA"; "Mike Brown")

a) Suppose we want to copy all data from database S to database T. Write a single SQL query that when executed over database S would transform all data of S into the format of T. That is, the query would create tuples for table LISTINGS of T from the data in S.

b) In practice, writing such SQL queries to copy data from one database to another is very time consuming. To save time, a user can employ a schema matching tool (such as those described in the Rahm/Bernstein paper) to find semantic matches between S and T. Examples of such matches are: **location** = **area** and **name** = **agent-name**. The user then employs a tool such as Clio (described in the Rahm/Bernstein paper) to elaborate these matches into SQL queries (that can then be executed to copy data).

Using these tools, can the above process be completely automated? If yes, why? If not, why not? In that latter case, discuss at which points in the process the user must be involved, what the user must do, and why.

c) Suppose the user has copied data from database S to database T, and has also copied data from database T to another database U. While doing this, the user has established that attribute x of S matches y of T, and attribute y of T matches z of U. Can the user conclude that attribute x of S matches z of U? If yes, why? If not, why not?

5. Information Retrieval and Web Search

a) Given a document collection, suppose you want to allow keyword searches such as "find all documents that contain the phrase 'data integration'" and "find all documents where the words "uncertainty" and "lineage" occur within a three-word distance". Design an appropriate inverted index, and discuss how you use that index to solve this problem. Outline the main data structures that you would use, and discuss any potential limitations of your solution.

b) Why is it that traditional keyword search techniques that work well in a controlled document collection (as discussed in the Singhal paper) often do not work well on the Web? Briefly describe the PageRank idea and explain how keyword search based on PageRank addresses the problems of traditional keyword search techniques.

c) Do you think keyword search based on PageRank would work well in organizational intranets? If yes, why? If not, why not?