NOTES:

(1) THE EXAM IS DIVIDED INTO 5 LONG QUESTIONS AND 5 SHORT QUESTIONS. THE GENERAL SCOPE OF THE QUESTION IS INDICATED FOR EACH QUESTION IN ITS TITLE.

(2) PLEASE ATTEMPT 4 OUT OF THE 5 LONG QUESTIONS (QUESTIONS 1 THRU 5 ARE STATED AS LONG) AND 4 OUT OF 5 SHORT QUESTIONS (QUESTIONS 6 THRU 10 ARE STATED AS SHORT).

(3) In question no. 5, you have the alternative to answer one or two parts. It will still be treated as ONE question.

(4) Q6 is a short question – you are required to answer only one part.

(5) If you feel there is ambiguity in some question, please state a reasonable assumptions and proceed.
Q1. DATA MODELING (LONG)

The relational data model has only one type of relation and the only semantics other than that are the concept of a key and the primary-foreign key semantics.

Two other relationship types are important in other data models, the IS-A relationship type and the PART-OF relationship type.

To help enforce these two relationship types you are requested to do two things:

1) design a set of catalog relations using the relational model to store information about IS-A relationship types and PART-OF relationship types.
2) Express the semantic rules that govern the use of relations of the two new relationship types, IS-A and PART-OF, during insertions into, updates of and deletions from them.

Propose such a design and illustrate it with a short sample schema containing IS-A and PART-OF relationships.
Q2. DATA WAREHOUSING /DATABASE DESIGN (LONG)

A data warehouse is a DBMS with an added set of functionalities.

(i) Comment on what additional functions need to be added on top of an RDBMS to make it function as a warehouse.
(ii) What is a data cube? Can you propose at least 3 or 4 algebraic operators for data cubes with a precise definition of how they are specified and giving some idea about how they may be implemented. Which of these operators are commercially available in products?
(iii) Point out the significant differences in designing a database from a conceptual model and designing a data warehouse.

Optional:
(iv) How would a MOLAP engine differ from a ROLAP engine? Are there any MOLAP products around? What advantage do they provide?
Q3. QUERY PROCESSING (LONG)

Answer both parts:

(i) Consider the following relational schema:
    Employee (Ssn, Lname, Fname, Salary, Dnumber) where Ssn is the primary key.

    Describe two strategies in detail, one using an index on Dnumber and the other
    not using an index, for efficiently executing the following query:

    SELECT Dnumber, AVG (Salary)
    FROM Employee
    GROUP BY Dnumber
    HAVING COUNT (*) < 3

    Also explain under what conditions, each strategy would be the better than the other
    one for executing this query. Be as specific as possible.

(ii) Describe modifications to the Hash-join algorithm so that it can compute the full
     outer join operation.
Q4. DOCUMENT DATABASES AND INFORMATION RETRIEVAL (LONG)

Table 1 below represents a matrix of documents and terms. Each document is represented by a row. Each element in the row represents the frequencies of corresponding terms. In this table each element $$f_{ij}$$ represents the frequency of term $$t_j$$ in document $$d_i$$ for all $$i, 1 \leq i \leq 4$$ and $$j, 1 \leq j \leq 4$$.

<table>
<thead>
<tr>
<th></th>
<th>$$t_1$$</th>
<th>$$t_2$$</th>
<th>$$t_3$$</th>
<th>$$t_4$$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$$d_1$$</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$$d_2$$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>$$d_3$$</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$$d_4$$</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1

(a) Define Inverse Document Frequency (IDF) and Term Frequency (TF). Discuss why TF-IDF is an important metric in Information Retrieval Systems.

(b) Calculate IDF of terms $$t_1$$, $$t_2$$, $$t_3$$ and $$t_4$$.

*Note: the IDF should be calculated with 2-based logarithm, viz. $$\log_2(N/n_j)$$.*

(c) Let $$tf_{ij}$$ be weights of document $$d_i$$ for all $$i, 1 \leq i \leq 4$$ and $$j, 1 \leq j \leq 4$$, be defined as:

$$tf_{ij} = \alpha + \frac{(1-\alpha)f_{ij}}{\text{Max}(f_k | 1 \leq k \leq 4)}$$

when $$f_{ij} \neq 0$$ and $$tf_{ij} = 0$$ when $$f_{ij} = 0$$.

Given a fixed $$\alpha = 0.3$$, determine representation of each $$d_i$$, $$\forall i, 1 \leq i \leq 4$$ given by $$tf_{ij}$$-idf weights. The $$tf_{ij}$$-idf can be calculated as the product of $$tf_{ij}$$ * idf_j.

(d) Assuming that the vector representing $$d_1$$ with $$tf_{ij}$$ weights stand for components of $$d$$ along term vectors, and that a query is given by $$q=(1, 0, 0, 1)$$, what is the RSV of $$d_1$$ with respect to $$q$$? [Hint: RSV stands for Retrieval Status Value and Computing RSV in this case is to compute the vector distance between document $$d_1$$ and q. Use simple retrieval function (i.e. the inner product)]

(e) Discuss how the decision on setting $$\alpha$$ in (c) would impact the TF and RSV of $$d$$ with respect to $$q$$. 

PhD-Qualifier-April4'08 Page 5 of 11 Database Area
Q5. DISTRIBUTED DB / CONCURRENCY CONTROL (LONG)

Note: You may attempt one or both questions.

Most of the centralized concurrency control algorithms preserve serializability, a well-defined correctness criterion. One reason many distributed concurrency control algorithms have been proposed to preserve serializability in distributed databases is that serializability requires coordination of all participants. For a distributed transaction with sub-transactions running on different nodes, simply maintaining local serializability in every database site does not always produce globally serializable transaction execution histories.

Question A. Distributed concurrency control.

(1) [20%] Give an example (as concrete as you can) of locally serializable transaction execution histories that are not globally serializable when viewed together as a distributed database. Hint: it is easier to use an example with local histories produced by 2PL (see the next item).

(2) [30%] Explain why a centralized concurrency control algorithm such as Two-Phase Locking (2PL) may produce local histories such as those in the above item (serializable in each node, but not globally). You will need histories produced by 2PL to answer this item.

(3) [50%] The strict 2PL concurrency control, where locks are released only when the transaction ends, achieves global serializability. Explain the reasons strict 2PL can guarantee global serializability by sketching a proof.

Question B. Comparison of distributed concurrency control algorithms.

(1) [40%] Design an additional layer of distributed concurrency control and coordination that will “fix” this problem for general 2PL, i.e., with this additional layer, the distributed database can achieve global serializability with each node running general 2PL.

(2) [30%] Compare the concurrency of transaction execution allowed by your “fixed” version with strict 2PL. The comparison can be qualitative (e.g., using a Venn diagram to describe the sets of histories allowed).

(3) [30%] Compare the performance costs of your “fixed” version with strict 2PL, in terms of number of messages exchanged that were introduced by your fix.
Q6. SPATIAL INDEXING AND SPATIAL QUERY PROCESSING (SHORT)

ANSWER ONLY ONE PART:

A. You have a spatial data set that contains WiFi Access Points (APs) with their 2D GPS coordinates and other AP information such as signal strength. Spatial queries of the form “Find all moving objects or all gas stations within in 5 miles of a given location (x1, y1)”. You are considering the use of an R-Tree index.

   i) What would the entries in the leaves contain?
   ii) What would the entries in the parents of the leaves contain?
   iii) What can you say about performance of range queries in comparison with a Grid File?

OR

B. Your DBMS only supports Grid File indexes. You want to store a collection of 2-D spatial objects. It is suggested that you do this by entering each object in every grid block that it overlaps with.

   i) What are the potential problems with this approach?
   ii) Can you see a situation (data/queries) in which this could be superior to R-trees?
Q7. CONCURREN CY CONTROL (SHORT)

Assume that the majority of transactions processed by our database system only read values from the database. What type of concurrency control mechanism would be best and why? When you answer this, you should explain why other competing approaches would not be as good. Be as specific as possible.
Q8. DATABASE RECOVERY (SHORT)

Discuss the interaction of the buffer manager in a DBMS and the recovery scheme in a DBMS. In particular, how is the output of log records to disk influenced by the buffer size? Also, discuss the role of latches in the recovery scheme.
Q9. DATA MINING (SHORT)

Consider the task of association rule mining. What makes the discovery of association rules from a large database a difficult task? (be precise and point out considerations regarding the parameters that govern the mining and the algorithmic consideration of reducing the complexity of this task.) What proposals have been made, if any – either as enhancement of SQL or as stand-alone languages to incorporate data mining into database management systems? Briefly comment on such language enhancements.
Q10. INFORMATION ACCESSIBILITY AND ARCHIVING -GENERAL QUESTION (SHORT)

(1) With the growing information accessibility over the past decade, there has been a fundamental shift away from ownership to access. Discuss how this fundamental change would impact the data and information management in the future.

(2) Google has recently announced the collaboration with a number of major academic and public libraries to digitize, and provide access to, large portions of their print material. This effort has been widely heralded as a boon to enhancing access to these library materials, but some information professionals, as well as publishers, have expressed concerns about this effort. Identify and discuss the major issues this initiative raises as well as the impact you believe this project may have.