NOTES:

(1) THE EXAM CONSISTS OF 10 SEPARATE QUESTIONS. EACH IS MARKED AS LONG OR SHORT. IN SOME LONG QUESTIONS, THERE ARE 2 PARTS, EACH OF WHICH IS CALLED AS A SHORT QUESTION (YOU CAN ANSWER EITHER ONE – WHICH WILL COUNT AS HALF, OR BOTH WHICH WILL COUNT AS FULL).

(2) ANSWER ANY COMBINATION OF QUESTIONS SO THAT YOU HAVE ANSWERED AN EQUIVALENT OF 6 LONG QUESTIONS. (E.G., YOU MAY ANSWER 4 LONG ONES AND 4 SHORT ONES TO MAKE A TOTAL OF 6).

(3) If you feel there is ambiguity in some question, please state a reasonable assumption and proceed.

(THE TOTAL EXAM HAS AN EQUIVALENT OF 9 LONG QUESTIONS OF WHICH YOU ARE ANSWERING 6).
Q1. DATA MODELING (A. SHORT, B. SHORT – BOTH TOGETHER -LONG)

A. You are deeply familiar with the ER/EER Model and the mapping to the Relational Model. You also know that this mapping does not preserve all the information in the ER/EER Diagram in the resulting Relational Schema. Please generate a list of all the cases where information is lost in the mapping, i.e. cases where it is not clear from the resulting Relational Schema exactly which ER/EER Diagram it came from. Please illustrate each case by an example.

B. There is a very high degree of proliferation of what we can term as “unstructured data” in the form of text, XML, HTML, and multimedia data in the database applications today.

1) What are the difficulties in modeling this data in the context of the relational model – illustrate with at least 3 concrete examples.

2) Choose any one “data type” (e.g., text) to be added to a relational DBMS. Point out /suggest what additional functionality would be necessary to deal with that type in the relational DBMS as well as what extensions are needed at the SQL or Algebra level.
Q2. DATABASE DESIGN (A. SHORT, B. SHORT – BOTH TOGETHER -LONG)

Most database design tools in use today allow ER and EER based schemas or UML class diagrams etc. as input and they produce mappings into relational tables, SQL queries as well as constraints. Answer the following in this context:

A. Most tools try to achieve 3NF schemas. Point out in your own words the exact nature of inter-attribute problematic relationships that are removed by normalizing into 3NF. Why is BCNF considered a stronger form of 3NF? What type of redundancy is addressed by 4NF? Why are 4NF and 5NF not really applicable in most practical situations?

B. There is a need for doing a series of physical design decisions with respect to any database. Describe three important decisions. (e.g., choice of an index). Suppose we want to develop a tool that allows various design tradeoffs to be evaluated. Suggest how you will go about designing such a tool- mention how the tool would be designed to evaluate the tradeoffs for each of the physical design decision you mention. (what inputs are considered, what parameters are captured, what data characteristics are considered …..etc.)
Q3. METADATA MANAGEMENT (LONG)

Here is a quote from a recent NSF workshop:

“Cyberscience Needs and Drivers in Mathematical and Physical Sciences

As the technology available to researchers in the mathematical and physical sciences become more advanced, the opportunities for doing science also evolve, requiring development of the new computational tools that will enable the frontiers of science discoveries. An important role for the members of the scientific community is to identify needs for cyberscience, defined as the science that cannot be done without the advanced capability of cyberinfrastructure. Advances in information technology-enabled systems, tools and services are revolutionizing the practice of science and engineering research and education. Today, simulation and modeling can be as important to discovery and innovation as theory and experimentation. Advances in sensor technology and the availability of affordable mass data storage devices are making possible the collection, creation and federation of large, complex datasets, and pervasive networking technology has made possible broad access to them. Grids are providing scientists and engineers, both professional and amateur alike, with access to a wide range of valuable, heterogeneous computing and information resources as well as sophisticated research instrumentation. Information technology has also permitted the creation of collaboration-rich digital knowledge environments, where researchers and educators work together on projects of common interest.”

Quoting the same source:

“Data Management Recommendations:

1. Develop a general-purpose open-source distributed RAID file system with redundancy, put together by teams (application scientists and computer scientists).
2. Develop authentication systems to guarantee data quality and integrity.
3. Draw on the experience of the digital-library initiative to understand the issues of longterm integrity of data, data preservation, and data curation. The relevant domain science must be engaged in developing this understanding.
4. Establish a common framework for the development of metadata, with emphasis on automation. Such a framework will require standards and an enforcement mechanism.
5. Engage the computer science community in database research.”

QUESTION: Please write a 2-page essay about the challenges database researchers face regarding metadata management (Recommendation 4) in this new world of cyberscience.
Q4. QUERY PROCESSING (LONG)

(a) Consider a relation NUMS defined by the following SQL statement:
Create Table NUMS (ADAT Integer).

Using conventional SQL, write a query to compute the median value in NUMS. You may assume there are no duplicates in NUMS and that NUMS contains an odd number of tuples.

Note there is no “Median” operator in conventional SQL. Your answer will be graded on simplicity as well as correctness.

(b) Consider a conventional query optimizer that estimates cardinality and cost to decide among different access methods, join methods, and query plans. In this setting, nested-loops joins are considered more “dangerous” than simple hash joins in the presence of poor cardinality and cost estimates. Explain why.

(c) Consider executing the following query on the table Employee(SSN, DeptNum):
Select * From Employee Where DeptNum = 3

You are deciding whether to use an index scan on a 3-level B+ tree for the DeptNum = 3 condition or to use a full table scan on Employee. Assuming a typical I/O cost metric, should you use a table scan or an index scan? Justify your answer.

We have the following assumptions:
• Employee has 100 pages.
• On average there are 10 records on a page.
• Record values are distributed randomly.
• You expect approximately 1% of the records to satisfy DeptNum = 3.
• A linear table scan is 5 times as fast per page as random page accesses.
Q4. WEB DATA MANAGEMENT (LONG)

You are asked to design a modern Web data management system for a large multi-volume encyclopedia like the Encyclopaedia Britannica. Describe in detail the ways in which you would organize and index the encyclopedia text, and the facilities you would offer the system user. Justify your choices. Comment on the issues that would arise if you wanted to evaluate system performance.

Q5. DATA MINING & ASSOCIATION RULES (SHORT)

Consider the task of finding association rules in market-basket data. One of the basic algorithms for doing this is the Apriori algorithm. Using the Apriori algorithm and the market basket transaction data shown in the table below, answer the following questions.

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{Milk, Diaper}</td>
</tr>
<tr>
<td>2</td>
<td>{Bread, Butter, Milk}</td>
</tr>
<tr>
<td>3</td>
<td>{Milk, Diaper}</td>
</tr>
<tr>
<td>4</td>
<td>{Bread, Butter, Cookies}</td>
</tr>
<tr>
<td>5</td>
<td>{Beer, Cookies, Diaper}</td>
</tr>
<tr>
<td>6</td>
<td>{Milk, Diaper, Bread, Butter}</td>
</tr>
<tr>
<td>7</td>
<td>{Bread, Butter, Diaper}</td>
</tr>
<tr>
<td>8</td>
<td>{Beer, Diaper}</td>
</tr>
<tr>
<td>9</td>
<td>{Milk, Diaper, Bread, Butter}</td>
</tr>
<tr>
<td>10</td>
<td>{Beer, Cookies}</td>
</tr>
</tbody>
</table>

(a) List the candidate frequent item sets of size 2 that would be examined by the Apriori algorithm for a minimum support > 30%. Which of those item sets of size 2 satisfy the minimum support, i.e., become frequent item sets. Show all work.

(b) List the association rules using your frequent item sets of size 2 from part (a) that have a confidence > 60%. Show all work.
Q6. INFORMATION RETRIEVAL (LONG)

You work as an IT support team member for a large company. You are told that the company has many meetings, both of internal staff and between staff and external clients, and meetings are recorded in formal minutes.

The company keeps an archival of minutes in files such that the executives can check back on decisions taken early in large projects.

You are asked to design a retrieval system so that company staff can locate minutes on a particular topic. Because of the legal implications that past discussions and decisions may have, the company is particularly concerned that the new retrieval system will be reliable and effective.

(1) Outline the design of your system, indicating the particular features and functionality it will have that are intended to meet the company’s requirements.

(Hint: you can assume that minutes are clearly dated and have explicit lists of participants).

(2) The company is willing to allow the installation of a pilot system so your approach can be evaluated under realistic conditions. Describe, in detail, your design for the evaluation in terms of what data, operational conditions and aspects of your system would you consider, and why? What performance measures would you apply, and why?
Q7. DISTRIBUTED DATABASES (SHORT)

How does the distribution of data across multiple databases affect the relational database normalization process?

Q8. APPLICATIONS OF DATABASES AND TECHNOLOGIES (LONG)

It is probably fair to say that the recent database technology has been impacted by developments such as social networking, mobile and location based applications, sensor based data and cloud computing. Point out at least 4 distinct problem areas related to these developments that the database technology must address for future products and give your viewpoints and suggestions regarding the important issues related to them. Be sufficiently clear in identifying the crucial new research needed.
Q9. MULTI-VERSION CONCURRENCY CONTROL (LONG)

A. (20%) Many database concurrency control algorithms (e.g., two-phase locking) are described using the in-place update assumption, where only one version is maintained. In classic database literature, one exception to this rule is multi-version timestamp-based concurrency control. Outline the multi-version timestamp-based concurrency control algorithm and explain intuitively why it maintains the serializability property.

B. (10%) Another exception to the single-version assumption is snapshot isolation (SI), in which two versions are maintained for two-phase locking (e.g., as in Oracle). Outline the SI property and a locking algorithm that preserves SI.

C. (20%) In theoretical and simulation studies, multi-version concurrency control has been shown to allow higher degree of concurrency than typical one-version concurrency control such as two-phase locking. Summarize the research findings and practical systems (e.g., Immortal DB by Microsoft Research) in an intuitive explanation for this higher degree of concurrency achieved by multi-version concurrency control.

D. (30%) There are similarities and differences between a system with infinite number of versions (e.g., Immortal DB) and a system with a small number of versions (e.g., SI with two versions). Refine your explanation of item (1.C) to describe the similarities and differences between a small number of versions (e.g., 2) and a large number of versions.
Q10. DISTRIBUTED DB AND DATABASE RECOVERY (LONG)

A. (20%) Explain succinctly (no more than 2 pages) the recovery algorithm (DO-
REDO-UNDO) of typical single-version database management systems for in-
place updates. You only need to consider transaction aborts and system crashes
(no media failure).

B. (40%) Explain how such recovery algorithms can be simplified in a multi-version
database such as the Immortal DB. [Hint: Rewrite the recovery algorithm of item
(2.A) using the information provided by the versions.]

C. (40%) Explain how you may achieve distributed recovery (of distributed
transactions in a distributed database) by adding a synchronization mechanism to
the recovery algorithms in (2.A) and (2.B).