NOTES:

(1) THE EXAM IS DIVIDED INTO 6 LONG QUESTIONS (1-6) AND 4 SHORT QUESTIONS (7-10). THE GENERAL SCOPE OF THE QUESTION IS INDICATED FOR EACH QUESTION IN ITS TITLE.

(2) PLEASE ATTEMPT 4 OUT OF THE 6 LONG QUESTIONS AND 3 OUT OF 4 SHORT QUESTIONS.

(3) If you feel there is ambiguity in some question, please state a reasonable assumption and proceed.
Q1. DATA MODELING (Long)

A. We all know the mapping from an Extended-Entity Relationship Diagram to a Relational Model Schema. In this mapping, information is lost, and there is no corresponding inverse mapping of a Relational Model Schema back to an Extended Entity-Relationship Diagram.

Provide at least 5 examples illustrating how different EER constructs are mapped to the same Relational Model constructs, thus preventing us from later determining unambiguously which EER constructs the Relational Model constructs came from.

B. XML is gaining in popularity as a language which is mainly serving as a means of exchanging data in a format that can be defined in a hierarchical way. (1) Provide arguments both pro and con for the claim that “XML is another data model with its own schema facility.” (Hint: start by considering the essential ingredients for a data model).
Q2. QUERY PROCESSING  (Long)

Refer to the following relational schema and statistics:

STUDENT (SId, SName, MajorId, GradYear)
DEPT (DId, DName)
COURSE (CId, Title, DeptId)
ENROLL (EId, SectionId, StudentId, Grade)
SECTION (SectId, CourseId, Prof, YearOffered)

<table>
<thead>
<tr>
<th>TABLE</th>
<th>D(T)</th>
<th>R(T)</th>
<th>V(T, F)</th>
</tr>
</thead>
</table>
| STUDENT | 4,500 | 45,000 | 45,000 for P=SId  
44,960 for P=SName  
50 for P=GradYear  
40 for P=MajorId |
| DEPT | 2 | 40 | 40 for P=DId, DName |
| COURSE | 25 | 500 | 500 for P=CId, Title  
40 for P=ProfId |
| SECTION | 2,500 | 25,000 | 22,000 for P=SectId  
500 for P=CourseId  
250 for P=Prof  
50 for P=YearOffered |
| ENROLL | 50,000 | 1,560,000 | 1,560,000 for P=EId  
25,000 for P=SectionId  
45,000 for P=StudentId  
14 for P=Grade |

Figure 16-7
Example statistics about the university database
T is the table name,  \( B(T) \) is the number of blocks that store table \( T \),  \( R(T) \) is the number of rows in table \( T \) and  \( V(T, F) \) is the number of unique values for column \( F \) in table \( T \). You can assume that attribute values are uniformly distributed across the table. For example, the SECTION table contains 250 unique Prof values. Hence there are 25,000/250 or 100 rows in the SECTION table with any given value for Prof.

Use the following SQL query for this question:

```
Select DName, Title, Prof
From DEPT, COURSE, SECTION
Where DId = DeptId and CId = CoursId
```

(a) Draw one query tree for this SQL query.

(b) Given the query tree in (a), describe the options that exist for the optimization of the query?

(c) If we assume that a JOIN operation is implemented using a nested block join algorithm, then explain what effect the order of the operands has with regard to the join cost.

(d) What effect does the size of the available buffer pool have on the cost of processing this query? For example, consider the two extreme cases, a very small and a very large size buffer pool.

(e) Compute the total cost of executing the query as depicted in the query tree in (a), assuming that the query is executed in a pipeline fashion. You can assume that there is one block for each table (i.e., input) and one block for each operator (i.e., output).
Q3. CONCURRENCY CONTROL (Long)

3.A (5%) Most of database concurrency control algorithms are based on the serializability property. Define serializability and explain how it can prevent the lost update problem.

3.B (15%) Describe the two-phase locking (2PL) concurrency control algorithm and explain how it can preserve serializability.

3.C (15%) Describe an optimistic concurrency control (OCC) algorithm and explain how it can preserve serializability.

3.D (20%) Compare 2PL and OCC in terms of their implementation complexity, execution overhead, and the degree of concurrency allowed. Use quantitative arguments for your comparisons and cite relevant literature when appropriate.

3.E (5%) Serializability is a global property, in contrast to local properties such as linearizability. Define global and local properties. Give concrete examples to show that serializability is a global property.

3.F (15%) Give an example of concurrency control algorithms that preserve local serializability at each site, but not global serializability in a distributed database. Give a concrete example of execution logs at each location that show local serializability, but not globally serializable when considered as a whole distributed database.

3.G (25%) Give an example of distributed concurrency control (DCC) algorithms that preserve global serializability. Show that it disallows the scenario shown in 3.F. Explain the difference between this DCC algorithm and the one described in 3.F, and provide an argument why this DCC algorithm can preserve global serializability in general.

Q4. DATABASE RECOVERY (Long)

- 4.A (10%) Define the cascaded aborts problem.
- 4.B (20%) Give an example of concurrency control algorithms that preserve serializability and will allow cascaded aborts. Provide a concrete scenario in which an execution is serializable, but cascaded aborts happen. Describe the scenario at the granularity of read and write operations in each transaction.
- 4.C (20%) Characterize the property that will disallow cascaded aborts. Explain the reasons why the scenario in 4.B did not satisfy your property and therefore allows cascaded aborts.
- 4.D (30%) Describe a recovery algorithm that can restore database consistency after the cascaded aborts scenario you gave in 4.B. Provide an argument of why your algorithm can claim the restoration of database consistency in general, not just the scenario in 4.B.
- 4.E (20%) Explain the conceptual difficulties in reconciling cascaded aborts with the ACID properties. Use concrete scenarios in applications to illustrate these difficulties.
Q5. DATA WAREHOUSING AND MINING  (Long)

A. Data warehousing and OLAP
(1) Data Warehouses need to go through an important step called ETL (Extract, Transform, and Load). The ETL problem is continuously getting more complex due to the variety of formats /standards and systems used to store the data. Point out the problems/issues involved in this process in a clear way and the approaches taken to solve them.
(2) Most data warehousing accompanied by OLAP is using the “cube” approach of summarizing data. Point out the issues/difficulties and limitations of this approach. In particular, what approaches exist for materializing these cubes and what are their limitations?

B. Data Mining: Association Rules
The Apriori algorithm is a simple algorithm for computing frequent item sets. We can treat the market-basket data as a relation such as Baskets (basket, item). On the other hand, we can store the data as a file consisting of records, one for each basket, where a record is a bit string of N bits (N being the total number of items). A ‘1’ in position \( m \) indicates the presence of item \( m \) in the basket and a ‘0’ indicates its absence. Discuss how these two representations affect the cost of the Apriori algorithm. In particular, you need only consider the cost of computing pairs of frequent items.
Q6. INFORMATION MANAGEMENT, STREAMING DATA, ACTIVE DATABASES (Long)

An emergency event, man-made or natural, can seriously affect or damage a nation’s critical infrastructure and cause tremendous loss of human lives and huge material damage. Any damage to public infrastructure - e.g., power grid, transportation, telecommunication - may have profound and wide-spread effects on the population because the public infrastructure represents one of the vital components of everyday life. Fast growth of urban areas places tremendous pressure on the public infrastructure. Emergency events and emergency response activities bring a new level of complexity and often overwhelm public infrastructure decision-making systems. A huge amount of public infrastructure stream data - e.g., video, audio, and sensor measurements – is generated and collected, but access to that data is often limited, in part because of the inability to process and present large volumes of stream data in an efficient and effective way, particularly for non-technical users. Presenting only raw stream data is often perceived to be inadequate. Such data and information often needs to be processed fast and be provided in a timely manner, so that analyses and predictions can be carried out. The provision of useful information, based on the processed stream data, and the conduct of such analyses, may only be possible through significant advancements in stream data collection, data integration, spatial-temporal data management, stream data mining, web services, and other types of information technology.

Considering the above, you will be focusing on managing traffic light control systems across various jurisdictions. You are asked to design a data processing system to manage the flow of traffic during an emergency event.

(a) Propose a model of the data management system
- Identify the potential users of this system, and explain their corresponding roles.
- Provide a high level diagram illustrating the overall system architecture.
- Describe each of the main components and processing flow of your solution.
- Describe by way of an exhaustive scenario how your system works.

(b) Design a centralized implementation of the proposed emergency response system.
Your solution should address the following aspects:
- Motivation for the proposed design.
- Identifying the strengths and shortcomings of the proposed system.

(c) Develop a distributed implementation of the proposed emergency response system.
Your solution should address the following issues:
- Motivation for the proposed design.
- Identifying the strengths and shortcomings of the proposed system.
Q7. DATABASE DESIGN (Short)

A. Prove that the proper mapping from an Extended Entity-Relationship Diagram to a Relational Model Schema always results in a set of relations in Boyce-Codd Normal Form. You may divide the answer into two parts: consider the ER model and its mapping; then bring in the additional features of Enhanced ER and its mapping.

Q8. DISTRIBUTED DATABASES (Short)

Describe the effect of data replication on SELECT and UPDATE processing in a distributed database system. We will assume that the granularity for replication is an entire table. If we have a given workload of SELECT and UPDATE transactions, describe how we could determine the amount of data replication we should have. Consider both I/O and transmission costs in your answer.

Q9. WEB DATA MANAGEMENT: (Short)

The so-called six degrees of separation, viz., everyone on Earth is separated from anyone else by no more than six degrees of separation, is just another manifestation of network combinatorics: if there are at least n links coming out of each node in a network, then the number of nodes that can be reached by traversing just m links in any directions is n**m; so if n is 100 and m is 6, n**m is 100**6, which is more than an order of magnitude higher than the human population.

The above combinatorial explosion makes finding a link between two given persons on the internet a very challenging problem. What method can be used to control this combinatorial explosion and attack the problem productively? Give the advantages and disadvantages of each method your propose.

Q10. APPLICATIONS/TECHNOLOGIES (Short)

Do you feel that more research is needed to deal with meta-data management in general to make applications provide a richer experience to the user? Argue at length by choosing any application area where you feel you have a pretty good understanding of the domain (e.g., GIS, Mobile Apps, Biological Apps, ERP,…..). First identify the issues, then point out some sample metadata of interest that would be applicable to the chosen area of application. Give examples of scenarios/use cases where current systems/applications are falling short in terms of lack of metadata and point out what specific problems exist related to metadata. Propose some approaches to deal with these problems. If you are aware of any commercial systems/solutions, mention them.