NOTES:

(1) THE EXAM STRUCTURE IS GIVEN BELOW. THE GENERAL SCOPE OF THE QUESTION IS INDICATED FOR EACH QUESTION IN ITS TITLE. A LONG QUESTION HAS A VALUE OF 1. A SHORT QUESTION HAS A VALUE OF 0.5.

(2) PLEASE ATTEMPT QUESTIONS SO THAT YOU HAVE ANSWERED AN EQUIVALENT OF AT LEAST 6 TOTAL QUESTIONS. YOUR CHOICE MUST INCLUDE AT LEAST ONE SHORT QUESTION.

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

Structure of Exam:

Q1. Data Modeling (Long)
Q2. Database Design (Short, Short)
Q3. Physical Design/Query Processing (Long)
Q4. Concurrency Control and Recovery (Long)
Q5. Concurrency Control and Recovery (Long)
Q6. Data Warehousing (Short)
Q7. Data Mining (Short)
Q8. Web Data Management (Long)
Q9. Mobile Data Management (Long)
Q10. Applications and Future Research (Long)
Q1. DATA MODELING (Long)

“The traditional software configuration management (SCM) process is looked upon by practitioners as the best solution to handling changes in software projects. It identifies the functional and physical attributes of software at various points in time, and performs systematic control of changes to the identified attributes for the purpose of maintaining software integrity and traceability throughout the software development life cycle.” [Wikipedia]

Please do all of the following:

1) List 6-8 key functions supported by an SCM.

2) Design an ER/EER Diagram for a database supporting the management of data needed to support the above functions in an SCM.

3) Explain in details how each of the function would store/maintain/retrieve the data in the database.

4) Write a list advantages and disadvantages of using a database to support a SCM.
Q2. DATABASE DESIGN

THESE TWO QUESTIONS ARE INDEPENDENT. You may answer any one or both as independent questions.

A. (SHORT)

When normalizing a relational database,

- We know that we can always decompose to 3NF while preserving FDs and avoiding lossy joins.
- We also know that we can sometimes decompose to BCNF while preserving FDs and avoiding lossy joins.
- Now, there are situations where we cannot decompose to BCNF and still preserve all FDs.

In the latter case, there is a choice. We can either keep the 3NF relation and deal with redundancy. Or, we can choose two BCNFs if we are willing to join them together each time we insert or update the relations to check if the FDs are still preserved.

Please give examples of practical cases where we should choose one approach over the other. Give sufficient arguments to support your choice based on practical and semantic considerations.

B. (SHORT)

Consider the problem of designing a Data Warehouse. It typically has fact tables which may be pointed from multiple dimension tables.

- There has not been much popular theoretical work on normalization of data warehouses. Is it because it is equivalent to normalization of relations – so there is nothing special required for warehouses, or that the ideas of “normalization” and “decomposition” do not apply at all to data warehouses? Explain with proper examples.
- Can the ER model be of help in constructing a data warehouse? Take a sample DW star schema and show how it could have been arrived at starting with conceptualization in ER. Point out what may be some of the necessary extensions to the ER model to make it suitable to model data warehouses.
Q3. PHYSICAL DESIGN/QUERY PROCESSING  (LONG)

Consider a shared nothing parallel database system that has 10 nodes. Each of the processing nodes has its own local memory (1GB) and a disk (50GBs) and sends messages to communicate with the other nodes.

Suppose we want to store the following tables in our DBMS:

Customer (CID, CName, City, State, Phone) where the primary key is CID. The size of the Customer table is 25GBs. We will also support an index on the primary key.

Sales (CID, PID, SDate, Quantity) where the primary key is CID,PID,SDate. The size of the Sales table is 200GBs. We will also support an index on the primary key.

Suppose we have the following queries that need to be processed on the DBMS:

Select * From Customer Where CID = $CID  ($CID represents variable input)

Select City, Sum(Quantity) From Customer Natural Join Sales Group By City

(a) Describe two data placement strategies for these two tables, i.e., how are the rows in the tables assigned to the disks in the DBMS.

(b) Describe any advantages that one scheme may have over the other.

(c) Describe, in sufficient detail, how the two queries could be processed for your two data placement schemes.

If you need to make any assumptions, state them clearly.
Q.4 and 5 CONCURRENCY CONTROL AND RECOVERY

You may answer any one or both as independent questions.

Q.4. (LONG) Although concurrency control and recovery algorithms are described separately, they sometimes have interactions that complicate the design and implementation of database management systems. One example of such interactions is cascaded aborts, a situation where committed transactions may have to be aborted due to another transaction abort.

(1) [10%] Explain why the general 2PL concurrency control, where locks may be released prior to transaction end, may introduce cascaded aborts.

(2) [10%] Explain why the strict variant of Two-Phase Locking (2PL) concurrency control, where locks are released only when the transaction ends, prevents cascaded aborts.

(3) [30%] Outline the normal DO-REDO-UNDO recovery algorithm in sufficient detail to explain its capability to support normal transaction aborts.

(4) [30%] Explain the main reasons the normal DO-REDO-UNDO algorithm (outlined in your answer to Q.4.3) is not designed to support cascaded aborts.

(5) [20%] Outline a strategy to extend the normal DO-REDO-UNDO algorithm that makes the extended recovery algorithm support cascaded aborts, by addressing the issues raised in Q.4.4.

Q.5. (LONG) Consider an application built using service-oriented architecture (SOA, for concreteness you may consider web service APIs), in which databases are independent of each other, and the 2-phase commit protocol (2PC) is considered undesirable in SOA due to its tight coupling of components.

(1) [20%] Assume that two services SA and SB both offer a compensation service (SA⁻¹ and SB⁻¹). You need to build an atomic composite service SF using both SA and SB. Explain how you can commit SA and SB separately, and then compensate for them using SA⁻¹ and SB⁻¹ if needed, to achieve an effect similar to an atomic commit involving SA and SB.

(2) [20%] Design a forward recovery algorithm that generalizes your answer to Q.5.1 for N participating services.

(3) [20%] Assume that two services SC and SD both offer an undo service Un-SC and Un-SD, which are able to undo the effects of SC and SD, even after they have committed. (There may be some practical limitations due to the coupling between SC and Un-SC, for example, but you may ignore them for this question.) Explain how you can build an atomic composite service SG using SC and SD, by using Un-SC and Un-SD in a way similar to your answer to Q.5.1.

(4) [20%] Design a backward recovery algorithm that generalizes your answer to Q.5.3 for N participating services.

(5) [20%] Compare the semantic and practical similarities and differences between backward recovery (Q.5.4) and forward recovery (Q.5.2). Specifically, explain how they are able to handle the service equivalent of cascaded aborts. Hint: Use concrete examples to explain the differences of these strategies on other concurrent and intervening services.
Q6. And Q7. DATA WAREHOUSING (SHORT) and DATA MINING (SHORT)

Q6. DATA WAREHOUSING (SHORT)

In Data warehousing, the common functionalities of drill-down, roll-up, pivoting etc. have been implemented. The OLAP operations further perform “data-cube” oriented operations. Consider a spatio-temporal data warehouse that keeps track of events in space and time. Propose an architecture for such a data warehouse focusing on what spatio-temporal functionality you would implement in such a warehouse and what the end-user will be able to do. Explain your choices clearly with proper justification.

Q7. DATA MINING (SHORT)

Consider the task of clustering a set of objects, based on their attribute values, into K disjoint sets of objects, i.e., clusters. We will consider the much used K-means algorithm where the grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

Suppose we have 4 objects and each object has 2 attributes as shown below.

<table>
<thead>
<tr>
<th>Object</th>
<th>Attribute 1</th>
<th>Attribute 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

(a) Illustrate the processing of the K-means algorithm for a value of K = 2 and initial centroid values of (1, 1) and (2, 1) which are randomly chosen from our set of objects.

(b) How does the choice of initial centroid values affect the algorithm? Are initial centroid values that are as far apart from each other as possible a better choice?
Q8. WEB DATA MANAGEMENT (LONG)

You are asked to design a Web service optimizer. Assume that the Web service providers communicate with relevant database providers and standard protocols and languages, such as Simple Object Access Protocol (SOAP), Web Services Description Language (WSDL), and the Universal Discovery Description & Integration (UDDI), will be used in the system. The set of Web services is provided by Web service providers. You are asked to use the following application to illustrate your design concept:

Suppose a restaurant chain plans to establish a new restaurant within a city and needs to determine the best location to generate the maximal revenue using information from several web service providers. Such information includes population (census), maps, and neighboring organizations/institutions/associations (e.g., residential, hotel, suppliers, partnerships, and entertainment). Based on such information, the new restaurant might be able to identify possibly partners for various ventures, or to locate nearby movie theatres or wholesale suppliers of beverages. (Note: You may disregard the access cost or assume a constant access cost to databases.)

You are asked to design a generic web service query optimizer that can collect the information useful for the best location decision above. Illustrate your approach using the above example.

(a) Propose a supporting architecture for the web service optimizer.
   • Analyze your proposed system architecture.
   • Present the considerations that went into designing your system.

(b) Design an algorithm for your query optimizer.
   • Use diagrams to illustrate the design of the algorithm.
   • Identify the strengths and shortcomings of your algorithm.

(c) Illustrate your approach using the above example. The constraint conditions are: type of restaurant, type of service food, and price ranges. (Thus, you might look for an optimal solution - best location to generate maximum revenue - under the constraint that all food is priced below the average for the area.)
Q9. MOBILE DATA MANAGEMENT (LONG)

Mobile data management has been around for more than 30 years. Although substantial efforts have been witnessed in mobile data management, most of existing research efforts in the first two decades had been dedicated to mobile location management techniques in centralized location monitoring systems for cellular networks, targeting at reducing the costs of location updates and location query processing, such as spatial indexing methods for efficient server side processing of location queries. The goal is to avoid constant re-evaluation of the pending location queries, while ensuring the correctness of their answers within certain tolerance bounds (delta). However, most of these indexing algorithms have to deal with the increased time complexity as the number of mobile objects becomes much larger than the number of queries in the system.

As mobility and Internet connectivity become pervasive, we see an increasing demand and interest in revisiting techniques and architectures for location based data management. You are asked to answer the following questions below with the best knowledge and research experiences you have had so far:

1. List the set of assumptions in the past location data management or mobile database research. [Hints: consider the mobile network, the location updates of moving objects, and the location query evaluation].

2. State which of the assumptions are no longer held today and why.

3. State the set of new assumptions and requirements in location based data management today and elaborate on your statements with some concrete examples.
Q 10. Applications AND Future Research (Long)

There are several recent developments that are not yet ready for including in standard DB textbooks yet. Consider the following popular terms:

- Key-value based data storage and retrieval (e.g., NoSQL databases)
- Map-reduce approach to efficient query processing (e.g., using key-value storage vs. columnar databases)
- Use of Cloud platforms for database applications, in both Infrastructure as a Service (Amazon EC2) and Platform as a Service (Microsoft Azure) styles

Discuss the essential ideas involved in each the above concepts; point out their current status, required further research, and possible future applicability. Give your assessment of what impact they will have on the future of data management.