M.TECH. COMPUTER TECHNOLOGY FIRST YEAR SECOND SEMESTER – 2024

DATABASE TECHNOLOGY AND DATA MINING

Time: Three Hours   Full Marks: 100

Instructions
- The examination is divided into two groups A and B, each carrying 50 marks.
- Use separateAnswerscripts for Group A and Group B.
- Please specify Group A or Group B clearly on the top sheet of each Answer script.
- Submit both Answer scripts together at the end of the examination.
- Each group has specific instructions regarding number of questions to be answered therein.

Group A: Database Technology

Group A instruction:Answer Question No. 1 and any three from Questions 2-6.

1. Indicate whether the statement is "True" or "False". For each false statement, rewrite the false statement in its nearest correct form.   [14]
   (a) Normalization always leads to a higher degree of redundancy in a database.
   (b) Indexes in a database can only be created on primary keys.
   (c) In a database, the term 'NULL' is equivalent to 'zero' or an empty string (").
   (d) In SQL, the ORDER BY clause is used to specify the columns for which data should be retrieved from a table.
   (e) Triggers in a database can only be fired before an operation (e.g., BEFORE INSERT, BEFORE UPDATE).
   (f) In EER modeling, specialization always results in disjoint subclasses.
   (g) In a relational database, a transaction is a sequence of operations that must be executed as a single unit.

2. (a) Discuss how the Three Schema Architecture enables data independence in DBMS when adding a new attribute to an existing entity. Explain the procedural steps at each level (logical, physical, and view) to accommodate this modification.   [6]
   (b) A financial institution wants to implement strict data security policies to comply with regulatory requirements. Explain how views at the external schema level can be leveraged to enforce access controls and confidentiality measures on sensitive financial data. Provide examples to illustrate your answer.   [6]

3. (a) Discuss the concept of the client-server architecture in DBMS. Compare and contrast the two-tier and three-tier client-server architectures, providing advantages and disadvantages of each.   [6]
   (b) Explain the various types of joins in relational databases, including natural join, inner join, outer join, and self join. Provide example(s) and use case(s) for each type of join.   [6]

4. You are tasked with managing a hotel's booking system. The system keeps track of various aspects such as customers, rooms, bookings, and payments. The database consists of several tables with defined relationships. The following tables are used in the hotel booking system.

[Turn over

Customer(CustomerID, FirstName, LastName, Email, Phone)
Room(RoomID, RoomType, PricePerNight, Capacity)
Booking(BookingID, CustomerID, RoomID, CheckInDate, CheckOutDate, TotalAmount)
Payment(PaymentID, BookingID, PaymentDate, Amount, PaymentMethod)
where the primary keys are underlined, and Booking.CustomerID, Booking.RoomID and Payment.BookingID are foreign key to Customer.CustomerID, Room.RoomID and Bookig.BookingID respectively.

For each part of the following questions (considered independently of the other parts), write a single SQL statement that accomplishes the given requirements.
(a)   List the total amount spent by each customer along with their names.   [4]
(b)   List the payment methods and the total amount paid through each method.   [4]
(c)   Find the room number and type of the cheapest room available in the hotel.   [4]

5.   (a)   Given the following unnormalized table for a bookstore:

Order ID	Customer Name	CustomerAddress	Book ID	BookTitle	Author	Price	Qty
1	Sujoy Das	23A Central Rd	B1	DBMS	Desai	450	2
1	Sujoy Das	23A Central Rd	B2	Data Mining	Basu	525	1
2	Greg Stupka	45 Golf Club Rd	B1	DBMS	Desai	450	1

Normalize this table to 3NF.   [6]

(b)   In a company's database, an employee can supervise other employees. Design a recursive ER diagram to represent this supervisor-subordinate relationship and explain how you would implement it in a relational database.   [6]

6   (a)   Consider the following relational algebra expression and assume a relational database with the schemas: R(A, B, C) and S(C, D)

The expression is: $\pi_{A,B}\left(\sigma_{R.A=S.C \land R.C>10}(R \times S)\right)$

Construct a query tree for the given expression by clearly showing the relational algebra operations and their sequence   [6]

(b)   Given the following database schema and SQL query, draw the initial (canonical) query tree for the SQL query:
Database Schema
Employee(EmpID, EmpName, DeptID, Salary)
Department(DeptID, DeptName, Location)
Project(ProjID, ProjName, DeptID)
SQL Query:
SELECT E.EmpName, D.DeptName
FROM Employee E, Department D, Project P
WHERE E.DeptID = D.DeptIDAND D.DeptID = P.DeptID
AND P.ProjName = 'ProjectX';   [6]

Group B: Data Mining

Group B instruction: Answer Question No. 1 and any two from Questions 2-5.

1. Either support or refute **any six** of the following statements with **very brief** reasonings:- [6x3]
    (a) Clustering is used to clean data at the pre-processing stage.
    (b) Binary data is a special type of categorical data.
    (c) Data cubes can only be used to store data having three dimensional aggregate values.
    (d) A galaxy schema contains a single fact table shared between multiple dimension tables.
    (e) Data integration is a data pre-processing step achieved through correlation analysis.
    (f) The Apriori property follows the antimonotonicity rule.
    (g) The ID3 algorithm is a recursive one.
    (h) Dendograms are associated with the Density-based unsupervised learning technique.

2. (a) Explain one Lazy Supervised Learning Technique with example. [8]
    (b) Relate the working principle of one classification scheme with Baye's theorem. [8]

3. (a) Describe one Frequent Pattern Mining algorithm without candidate generation. [10]
    (b) Discuss what is a null-invariant measure in Correlation Analysis. [6]

4. (a) Outline a simple partitioning method in unsupervised learning. [8]
    (b) Elaborate on how the above technique can be adapted for fuzzy clustering. [8]

5.   Write short notes on **any four** of the following topics :- [4x4]
    (a) Data Matrix versus Dissimilarity Matrix
    (b) Overview of Data Reduction Strategies
    (c) Typical OLAP operations
    (d) Measures of association rule interestingness
    (e) Some Performance Evaluation Metrics for Classifiers
    (f) Outlier Detection – one statistical technique