## KUMARAGURU COLLEGE OF LIBERAL ARTS AND SCIENCE

**Affiliated to Bharathiar University**

# **Saravanampatti, Coimbatore – 641 049**

**Department of Data Science**

**MID Term: August 2025**

|  |  |  |
| --- | --- | --- |
| **Batch: 2024-2027** | **Class: II BSc Data Science** | **Subject Title: Data Engineering** |
| **Semester: III** | **Max. Marks: 50** | **Time Duration: 2 Hrs.** |
| **Name: Mugilan** |  |  |

**Section A**

**Answer all questions (10 x 1 = 10)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **What is the first step in data preprocessing?** | | | |
|  | **A** | **Data reduction** | **C** | **Data cleaning** |
|  | **B** | **Data transformation** | **D** | **Data integration** |
|  | | | | |
| **2** | **What are the two main types of data objects?** | | | |
|  | **A** | **Categorical and numerical** | **C** | **Scalar and vector** |
|  | **B** | **Linear and non-linear** | **D** | **Binary and multi-valued** |
|  | | | | |
| **3** | **What is the purpose of data discretization?** | | | |
|  | **A** | **To reduce noise in data** | **C** | **To transform data** |
|  | **B** | **To convert data into categorical form** | **D** | **To merge data from multiple sources** |
|  | | | | |
| **4** | **Which of the following is a common technique used in data cleaning?** | | | |
|  | **A** | **Discretization** | **C** | **Normalization** |
|  | **B** | **Integration** | **D** | **Removing duplicates** |
|  | | | | |
| **5** | **In a data warehouse, what is the purpose of a data cube?** | | | |
|  | **A** | **To store transactional data** | **C** | **To organize data in a multidimensional manner** |
|  | **B** | **To store relational data** | **D** | **To process data in real-time** |
|  | | | | |
| **6** | **What is a fact table in the context of data warehouse modeling?** | | | |
|  | **A** | **A table that stores metadata** | **C** | **A table that defines relationships between other tables** |
|  | **B** | **A table that contains summary data** | **D** | **A table that contains detailed transactional data** |
|  | | | | |
| **7** | **What is the main goal of data warehouse implementation?** | | | |
|  | **A** | **To enable efficient data analysis and reporting** | **C** | **To replace operational databases** |
|  | **B** | **To optimize transaction speed** | **D** | **To store unstructured data** |
|  | | | | |
| **8** | **What is the role of a “star schema” in data warehouse design?** | | | |
|  | **A** | **To support normalized data storage** | **C** | **To enforce strict data integrity rules** |
|  | **B** | **To simplify queries by structuring data in a central fact table surrounded by dimension tables** | **D** | **To replicate data across multiple locations** |
|  |  |  |  |  |
| **9** | **Which of the following is a characteristic of a relational data model?** | | | |
|  | **A** | **Document-based** | **C** | **Hierarchical data storage** |
|  | **B** | **Schema-less** | **D** | **Fixed schema with tables, rows, and columns** |
|  |  |  |  |  |
| **10** | **Which NoSQL data model is best suited for hierarchical data storage?** | | | |
|  | **A** | **Document store** | **C** | **Column store** |
|  | **B** | **Graph store** | **D** | **Key-value store** |

**Section B - Answer All questions (5 x 4 = 20)**

|  |  |
| --- | --- |
| **11 A** | **A marketing analytics team is trying to extract valuable insights from a large customer transaction database to improve campaign targeting. In this context, what are the main stages involved in the Knowledge Discovery in Databases (KDD) process, and can you briefly illustrate each stage with a relevant example? (OR)** |
| **B** | **A retail analytics company is working with a chain of supermarkets to improve customer satisfaction, optimize inventory, and personalize marketing strategies. The team wants to explore various sources of data available across the business. In this context, what are the different types of data that can be mined, and how can each type be described in terms of its structure and potential use?** |
|  |  |
| **12 A** | **How does data integration support the data mining process, and why is it crucial for generating accurate and comprehensive insights? (OR)** |
| **B** | **A healthcare analytics team is preparing patient records for a machine learning model to predict disease risk. The data includes numerical values like blood pressure and cholesterol levels, as well as categorical information like age groups and diagnosis codes. To ensure the model performs accurately, the team needs to preprocess the data appropriately. What roles do data transformation and discretization play in the data preprocessing phase, and why are they important?** |
|  |  |
| **13 A** | **A logistics company is planning to implement a data warehouse to centralize and analyze its historical shipping and inventory data. What are the core concepts that define a data warehouse, and how do these concepts support effective data storage and decision-making?**  **(OR)** |
| **B** | **ABC company is analyzing its sales data across multiple regions, time periods, and product categories to make strategic decisions. To support this analysis, what types of OLAP operations can be used, and how would each operation apply in this business context?** |
|  |  |
| **14 A** | **A financial institution is deciding between using OLAP or OLTP systems to manage its data for daily transactions and long-term trend analysis. In this context, how do OLAP and OLTP systems differ, and what are the specific use cases each is best suited for?**  **(OR)** |
| **B** | **A company’s data warehouse stores detailed customer transaction records. To generate high-level summaries for decision-making, the team wants to simplify the data without losing essential patterns. What is the significance of using data generalization through attribute-oriented induction, and how does it support strategic analysis?** |
|  |  |
| **15 A** | **A software development team is designing a new application that will handle large volumes of user data. To ensure efficient data organization and retrieval, they need to plan the structure of their database carefully. What is the importance of data modeling, and what role does it play in effective data management? (OR)** |
| **B** | **A startup is designing a database to manage customer orders, products, and inventory efficiently. To choose the right data model, the team is considering the relational approach. What are the fundamental characteristics of a relational data model, and how do they support effective data organization and retrieval?** |

**Section C - Answer All questions (2 x 10= 20)**

|  |  |
| --- | --- |
| **16 A** | **Consider the scenario that a “Mom and Me” store must improve their sales in the next quarter year of 2021. The technical team has decided to apply data mining techniques to attract new customers to their shop. So, the technical team started collecting the data from various corporate hospitals and by extracting and mining the data of pregnant women, they can find the customers who may buy the products from their shop. The data collected from different hospitals in different formats.**   1. **Summarize the seven steps of knowledge discovery process on the above data to mine interesting patterns.** 2. **Suppose there are any missing value in any of the data like no data filled for delivery and month of delivery. Identify the ways to handle those data before applying mining techniques.**   **(OR)** |
| **B** | **Analyze the architecture of a data warehouse system within a business organization. Your case study should include a detailed explanation of the components, such as data sources, ETL processes, staging areas, the central data warehouse, metadata repository, OLAP engine, and end-user access. Additionally, provide a diagram to illustrate the architecture. Discuss how this architecture supports the organization's data consolidation, quality management, advanced analytics, and reporting needs.** |
|  |  |
| **17 A** | **Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.**   1. **Draw a star schema diagram for the above data warehouse.** 2. **Reconstruct the above star schema by adding a new dimension table medicine with the features medicine\_id, medicine\_name, supplier\_id, supplier\_name, supplier\_address and supplier\_phone. (OR)** |
| **B** | **Perform a comprehensive case study on data modeling for a specified business scenario. Your study should cover the entire process, from gathering requirements to creating conceptual, logical, and physical data models. Discuss the techniques used for normalization, entity-relationship (ER) diagram creation, and the mapping of business rules into data structures. Provide diagrams to illustrate your models and explain how they support the business’s data management, integrity, and retrieval needs. Additionally, evaluate the impact of your data model on system performance and scalability.** |

**Name & Signature of QP Setter Name & Signature of HOD / Reviewer**

**Dr.M.C.S.Geetha Dr.K.S.Narayanan**