No. of Printed Pages : 3

MCS-221

MASTER OF COMPUTER APPLICATIONS (MCA-NEW)

Term-End Examination

June, 2023

MCS-221 : DATA WAREHOUSING AND DATA MINING

Time : 3 Hours *Maximum Marks : 100*

Weightage : 70%

Note : *(i) Question No. 1 is compulsory.*

(ii) Answer any ***three*** *questions from the rest.*

---

1. (a) Discuss the following data preprocessing stages briefly : 10

   (i) Data Cleaning

   (ii) Data Integration

   (iii) Data Reduction

   (iv) Data Transformation

(b) Define Aggregate fact tables and derived dimension tables. What are their significance ? Give an example. Also mention their advantages and disadvantages. 10

(c) Enumerate the key challenges in data warehouse design. 5

(d) Differentiate between a data lake and a data warehouse. 5

(e) Write and explain the Apriori algorithm to identify the most frequently occurring elements and meaningful associations in a dataset. 10

2. (a) What is Cluster Analysis ? How is this used in Data Mining ? Give an example. Also mention few applications of cluster analysis in data mining. 10

(b) List and discuss various types of Web-mining. 10

3. (a) Explain the following techniques for Dimensionality Reduction : 10

(i) Feature Selection

(ii) Feature Extraction

(b) Discuss the layered implementation of ETL in a Data Warehouse. 10

4. (a) Define OLAP. Differentiate between Multicube and Hypercube. Mention the applications of OLAP reporting system. 10

   (b) List and explain the following types of Data Warehouses : 10

      (i) Enterprise Data Warehouse (EDW)

      (ii) Operational Data Store

5. Write short notes on the following : 4×5=20

   (a) ELT *vs.* ETL

   (b) Data Marts

   (c) Applictions of Data Mining

   (d) OLAP data cube operations

No. of Printed Pages : 4

MCS–221

MASTER OF COMPUTER APPLICATIONS (MCA-NEW)

Term-End Examination

December, 2023

MCS-221 : DATA WAREHOUSING AND DATA MINING

Time : 3 Hours *Maximum Marks : 100*

(Weightage : 70%)

Note : *(i) Question No. 1 is compulsory.*

(ii) Answer any ***three*** *questions from the rest.*

---

1. (a) With the help of a diagram, describe the Conceptual Architecture of Hadoop Data Warehouse. 10

   (b) Draw and explain star schema diagram and snow-flake schema diagram for the dimensions (Products, Customers, Time, Locations) and fact (Sales-Items) for the measures namely Quantity-sold and Amount-sold for a manufacturing company data warehouse dimensional modeling. 10

(c) Define Noisy data while doing data pre-processing. Delete the noise with Binning smoothing techniques for the following details using partition in Bins (Equal-frequency) :

4, 2, 6, 10, 8, 16, 12, 24, 22, 14, 26

stored price details (in dollars). 10

(d) Define Clustering in Data Mining. Write and explain k-means clustering algorithm. List its advantages and disadvantages. 10

2. (a) What is Web-Mining ? List various web-mining tasks. Also, discuss the following types of web-mining : 10

(i) Web content mining

(ii) Web usage mining

(b) With the help of an example, explain rule-based classification. 5

(c) What are the various steps involved in building a classification model ? Explain with the help of an example. 5

3. (a) With the help of an example, explain Market Basket Analysis. 5

(b) Write and explain Apriori algorithm used to identify the most frequently occurring elements and meaningful associations in any dataset. 10

(c) List and discuss any *two* popular data mining tools. 5

4. (a) Discuss ETL and its need. Explain in detail, all the steps involved in ETL with the help of a suitable diagram. 10

(b) List and explain any *three* key challenges of Data Warehouse. 3

(c) With reference to Alex Gorelik, explain the following additional data lake stages : 7

(i) Data Puddle

(ii) Data Pond

(iii) Data Lake

(iv) Data Ocean

5. Write short notes on the following : 4×5=20

   (a) Aggregate fact table and derived dimensional tables

   (b) Data swamp

   (c) Data Preprocessing stages

   (d) Agglomerative approach of Hierarchical method

No. of Printed Pages : 3

MCS-221

MASTER OF COMPUTER APPLICATION (MCA-NEW)

Term-End Examination

June, 2024

MCS–221 : DATA WAREHOUSING AND DATA MINING

Time : 3 Hours *Maximum Marks : 100*

Weightage : 70%

Note : *Question number-1 is compulsory. Attempt any three questions from the rest.*

---

1. (a) With the help of a diagram, explain the real-time data warehouse architecture. Also, discuss its trade-off. 10

   (b) Describe fact constallation scheme. How is it different from snow-flake scheme ? With the help of an example explain fact constallation schema. Mention its advantages and limitations. 10

(c) Write and explain K-NN classification algorithm in Data-Mining. Discuss its advantages and disadvantages. 10

(d) Discuss vector space modeling for representing text documents. With reference to this modeling, explain TF-IDF and Inverse Document Frequency (IDF). 10

2. (a) Discuss various factors to be considered to improve the performance of ETL. 7

(b) Define a Data Mart. How is it different from a centralized data warehouse ? Discuss the structure of Data Mart with the help of an example. 8

(c) Discuss the following OLAP data cube operations : 5

(i) Roll-up (ii) Drill-down.

3. (a) Discuss the following Data Integration issues : 10

(i) Schema Integration and object modeling

(ii) Redundancy

(iii) Detection and Resolution of Data Value Conflicts.

(b) Write and explain Decision Tree Classifer algorithm with the help of an example. 10

4. (a) Discuss Linear Discriminant Analysis (LDA) and Principal Component Analysis (PDA) feature extraction techniques. 10

(b) Explain Association Rule Mining (ARM). Also, explain the two phases involved in ARM process with the help of an example. 10

5. Write short notes on the following : $4 \times 5 = 20$

(a) ELT and its need

(b) Metadata and Data Warehousing

(c) Rule based classification

(d) Naive-payes Classifier

No. of Printed Pages : 5

MCS–221

MASTER OF COMPUTER APPLICATIONS

(MCA–NEW)

Term-End Examination

December, 2024

MCS-221 : DATA WAREHOUSING AND DATA MINING

Time : 3 Hours *Maximum Marks : 100*

Weightage : 70%

Note : *Question No. 1 is compulsory. Answer any* ***three*** *questions from the rest.*

---

1. (a) Draw snowflake schema diagram for hotel occupancy, where there are four dimensions namely hotel details, room

details, time and customer with two measures namely occupied rooms and vacant rooms. Also, explain the snowflake schema diagram.

Assumptions can be made wherever necessary. 10

(b) What is ETL in Data Warehousing ? Why do you need ETL ? Describe the layered implementation of ETL in a data warehouse. 10

(c) What is Noisy Data ? How binning method solves the problem of noisy data ? Illustrate with an example. 10

(d) Write and explain K-nearest neighbour's algorithm. Write its advantages and disadvantages. 10

2. (a) What is text mining ? With reference to text mining, explain the following techniques :

   (i) Information Extraction

   (ii) Text Summarization

   (iii) Text Categorization

   (iv) Text Clustering

   Also, mention any ***four*** applications of text mining. 10

   (b) Explain the following clustering methods briefly : 10

   (i) Partitioning method

   (ii) Density-based method

3. (a) With reference to data warehousing, explain the following terms : 10

   (i) Metadata and Data warehousing

(ii) Data Granularity

(iii) Operational data store

(iv) Data Mart

(b) Briefly explain the single-tier and three-tier data warehouse architectures with the help of a suitable diagram for each. 10

4. (a) What is Data Mining ? Briefly explain the following data mining techniques with the help of an example for each : 10

(i) Association Rule Mining

(ii) Outlier Detection

(iii) Regression Analysis

(b) Write and explain Apriori algorithm to identify the most frequently occurring elements and meaningful associations in a dataset. 10

5. Write short notes on any ***four*** of the following :

4×5=20

(a) Mining multilevel association rules

(b) Data reduction (with reference to data pre-processing)

(c) Mining multimedia data on the web

(d) K-Medoids algorithm of clustering

(e) Rule-based classification

× × × × × × ×

No. of Printed Pages : 6 **MCS–221**

MASTER OF COMPUTER APPLICATIONS

(MCA–NEW)

Term-End Examination

June, 2025

MCS-221 : DATA WAREHOUSING AND DATA MINING

Time : 3 Hours *Maximum Marks : 100*

Weightage : 70%

Note **:** *(i) Question No.* ***1*** *is compulsory.*

(ii) Answer any ***three*** *questions from the rest.*

1. (a) "Resampling method is a common approach that can be used for model selection in supervised learning." With

reference to resampling method, explain the following approaches : 10

(i) K-fold cross validation

(ii) Leave-one-out method

(iii) Random split

(iv) Time-based split

(b) What is frequent pattern mining ? Briefly explain the following classifications of frequent pattern mining along with an example for each : 10

(i) Based on the levels of abstraction involved in the rule-set.

(ii) Based on the types of values handled in the rule.

(iii) Based on the kinds of rules to be mined.

(c) What is Online Analytical Processing (OLAP) ? Explain briefly the following OLAP data cube operations with the help of an example for each : 10

(i) Roll-up

(ii) Drill-down

(iii) Slice and Dice

(d) What is a star-schema dimensional modeling ? What are its characteristics ? Draw a star schema for four dimensions–Time, Item, Branch, Location with 2 measures namely 'Units-sold' and 'Amount-sold' and 'sales' is the Fact Table. Also, explain the star-schema diagram. Assumptions can be made wherever necessary. 10

2. (a) With the help of a suitable diagram for each, explain Top-down Approach (proposed by Bill Inmon) and Bottom-up Approach (proposed by Kimball) of Data Warehouse design. 10

   (b) What are Datamarts ? How are they different from a Data Warehouse ? Write and explain all the steps involved in a datamart design. 10

3. (a) Write and explain DBSCAN algorithm a density based method in clustering. Also, mention any *two* advantages and disadvantages. 10

(b) What are outliers in data mining ? How are they handled ? Explain briefly any *two* outlier detection techniques. 10

4. (a) What is a data lake ? Draw a diagram and explain all the four layers in data lake architecture. 10

(b) With reference to HADOOP architecture, explain the following : 10

(i) Name Node

(ii) Data Nodes

(iii) File System Namespace

(iv) Data Replication

5. Write short notes on any ***four*** of the following : 4×5=20

(a) Data Warehouse key challenges

(b) Cloud Data Warehousing

(c) ELT and its benefits

(d) Data Mining Life Cycle

(e) Data Integration issues in Data Preprocessing

× × × × ×