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

P. T. O.

(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