T.E. Com / Sem VI / Rev.
Data warehousing & Mining

Con. 3881-10. **(REVISED COURSE)** **AN-4472**

(3 Hours) **[Total Marks : 100**

N.B. (1) Question No. **1** is **compulsory**.
(2) Attempt any **four** questions out of remaining **six** questions.

1. (a) Define Data Warehouse. Explain the architecture of data warehouse with suitable block diagram. **10**
   (b) Explain data mining as a step in KDD. Give the architecture of typical DM system. **10**

2. (a) How are top-down and bottom-up approaches for building data warehouse differ ? Discuss the merits and limitation of each approach. **10**
   (b) What is K-means clustering ? Confer the K-means algorithm with the following data for two clusters. Data set { 10, 4, 2, 12, 3, 20, 30, 11, 25, 31 } **10**

3. (a) Give information package for recording information requirement for "Hotel Occupancy" considering dimensions like time, Hotel etc. Design star schema from the information package. **10**
   (b) Explain HITS algorithm. **10**

4. (a) What is Classification ? What are the issues in classification ? Apply statistical based algorithm to obtain the actual probabilities of each event to classify the new tuple as tall. Use the following data – **10**

Person ID	Name	Gender	Height	Class
1	Kristina	Female	1·6 m	Short
2	Jim	Male	2 m	Tall
3	Maggi	Female	1·9 m	Medium
4	Marya	Female	2·1 m	Tall
5	Stephanie	Female	1·7 m	Short
6	Bob	Male	1·85 m	Medium
7	Catherine	Female	1·6 m	Short
8	Dave	Male	1·7 m	Short
9	Wilson	Male	2·2 m	Tall

   (b) Define Metadata. What are the different types of metadata stored in a data warehouse ? Illustrate with a simple customer sales data warehouse. **10**

[TURN OVER

5. (a) What is Clustering Techniques ? Discuss the Agglomerative algorithm using following data and plot a Dendrogram using single link approach. The following figure contains sample data items indicting the distance between the elements :– **10**

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

(b) All electronics company have sales deptartment Sales consider three dimensions namely **10**

(i) Time (ii) Product (iii) Store.

The schema contain a central fact table sales with two measures.

(i) dollars–cost and (ii) units-sold

Using the above example describe the following OLAP operations :–

(i) Dice (ii) Slice (iii) Roll-up (iv) Drill-down

6. (a) Explain ETL of data warehousing in detail. **10**

(b) Consider the following transactions :– **10**

TID	Items
01	1, 3, 4, 6
02	2, 3, 5, 7
03	1, 2, 3, 5, 8
04	2, 5, 9, 10
05	1, 4

Apply the Apriori Algorithm with minimum support of 30% and minimum confidence of 75% and find the large item set L.

7. Write short notes on any **four** :– **20**

(a) Trends in data warehousing

(b) Decision tree based classification approach

(c) Key restructuring

(d) Crawlers

(e) Web personalization.