

1 Explain the similarities and differences between Database system, Data warehouse and Data Mining. List the difference between OLTP and OLAP **5 Marks**

2 Suppose that a data warehouse for *Big University* consists of the following four dimensions: *student*, *course*, *semester*, and *instructor*, and two measures *count* and *avg grade*. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg grade* stores the average grade for the given combination. **5 Marks**

(i) Draw a *snowflake schema* diagram for the data warehouse.

(ii) Starting with the base cuboid [*student*; *course*; *semester*; *instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of *CS* courses for each *Big University* student.

(iii) If each dimension has five levels (including all), such as "*student* < *major* < *status* < *university* < all", how many cuboids will this cube contain (including the base and apex cuboids)?

3 You are given the transaction data shown in the Table below from a fast food restaurant. There are 9 distinct transactions (order:1 – order:9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity we assign the meal items short names (M1 – M5) rather than the full descriptive names **5 Marks**

Meal Item	List of Item IDs	Meal Item	List of Item IDs
Order:1	M1, M2, M5	Order:6	M2, M3
Order:2	M2, M4	Order:7	M1, M3
Order:3	M2, M3	Order:8	M1, M2, M3, M5
Order:4	M1, M2, M4	Order:9	M1, M2, M3
Order:5	M1, M3		

Minimum support is 2 and the minimum confidence is 77%.

a. Apply the Apriori algorithm to the dataset of transactions and identify *all* frequent k-itemsets. Show all of your work.

b. Find all *strong* association rules of the form: $X \rightarrow Y \mid Z$ and note their confidence values

4 A database D has five transactions as shown below. Let min sup = 60% and min conf = 80%. **5 Marks**

TID	Items
T100	{B, O, N, E, C, O}
T200	{B, O, N, E, C, A}
T300	{C, A, N, E, C, A}
T400	{F, A, N, E, C, A}
T500	{F, A, C, A}

- (a) Using FP-growth algorithm, find all frequent 4- and 3-itemsets in the database D.
- (b) From the frequent itemsets you discovered, list all of the strong association rules matching the following metarule, where X is a variable representing customers, and $Item_i$ denotes variables representing items (e.g., “A”, “C”, etc.)

$$buys(X; item_1) \wedge buys(X; item_2) \rightarrow buys(X; item_3)$$