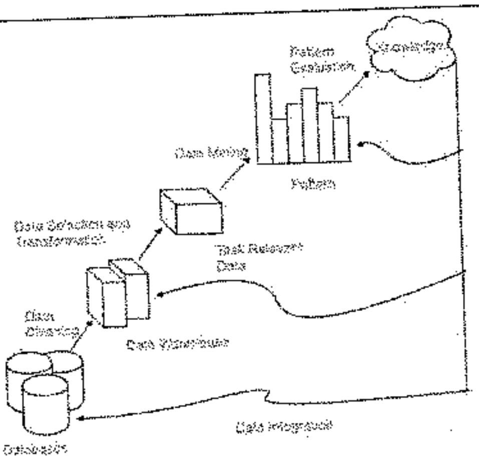
Shri Vile Parle Kelavani Mandal's
**DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

**Academic Year (2022-23)**
**Year: 3 Semester: V**

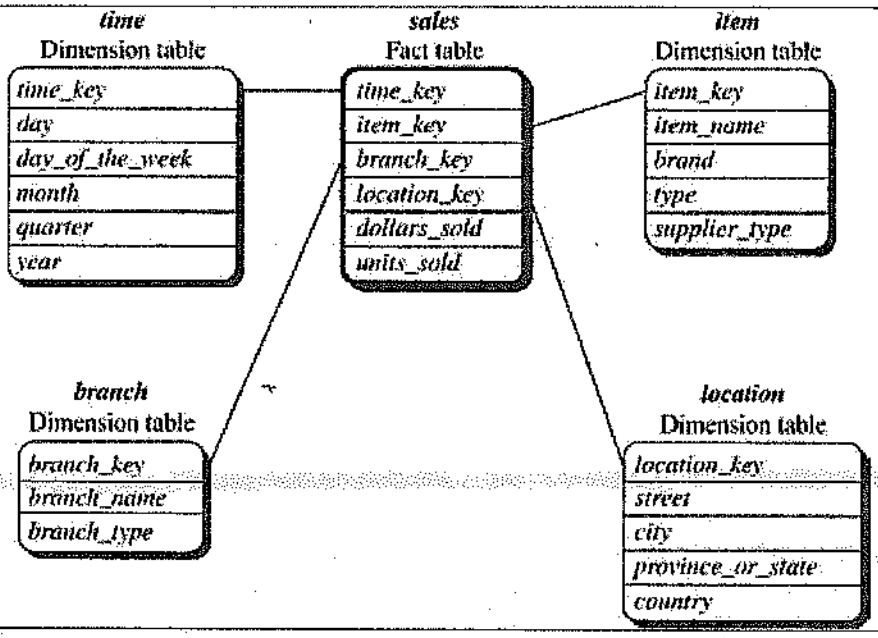
**Program: B. Tech. (Computer Engineering)**
**Subject: Data Mining and Warehouse**
**Date:**

**Max. Marks: 75**
**Time: 10: 30 am to 1:30 pm**
**Duration: 3 Hours**

**ANSWER KEY**
**Set 2**

| **Question No.** | | **Max. Marks** |
|---|---|---|
| Q1 (a) | A Data Warehouse is a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management's decisions.<br>he key features of a data warehouse are:<br>• **Subject Oriented** – A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.<br>• **Integrated** – A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.<br>• **Time Variant** – The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.<br>• **Non-volatile** – Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in operational database is not reflected in the data warehouse. | [10 ] |
| | **OR** | |
| Q1 (a) |  Knowledge<br>Pattern Evaluation<br>Data Mining<br>Patterns<br>Data Selection and Transformation<br>Task Relevant Data<br>Data Cleaning<br>Data Warehouse<br>Data Integration<br>Databases<br>**Data cleaning** (to remove noise and inconsistent data)<br>**Data integration** (where multiple data sources may be combined)<br>**Data selection** (where data relevant to the analysis task are retrieved from the database)<br>**Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)<br>**Data mining** (an essential process where intelligent methods are applied to extract data patterns)<br>**Pattern evaluation** (to identify the truly interesting patterns representing knowledge-based on interestingness measures.<br>Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users) Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledgebase. The preceding view shows data mining as one step in the knowledge | [10] |

Q2 (b) [08]



Q3 (a)

i. Boxplot displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. [02]

ii. LR 99, LDA 96.8, KNN 96.8, CART 98.5, SVM 96.8, RF 98.5 and ADA 99.6 [04]

iii. ADA boxplot has low IQR and low Range. Also median is more than or slighthly less than max accuracy of all other models. Hence, ADA. [04]

**OR**

Q3 (a)

Binning is a top-down splitting technique based on a specified number of bins. Data smoothing methods are also used as discretization methods for data reduction and concept hierarchy generation. For example, attribute values can be discretized by applying equal-width or equal-frequency binning, and then replacing each bin value by the bin mean or median, as in smoothing by bin means or smoothing by bin medians, respectively. These techniques can be applied recursively to the resulting partitions to generate concept hierarchies. Binning does not use class information and is therefore an unsupervised discretization technique. It is sensitive to the user-specified number of bins, as well as the presence of outliers. [04]

❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins: [02]

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

* Smoothing by **bin means**: [02]

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* Smoothing by **bin boundaries**: [02]

to the supermarket? This information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

Let's look at an example of how market basket analysis can be useful.

**Example 6.1** **Market basket analysis.** Suppose, as manager of an *AllElectronics* branch, you would like to learn more about the buying habits of your customers. Specifically, you wonder, *"Which groups or sets of items are customers likely to purchase on a given trip to the store?"* To answer your question, market basket analysis may be performed on the retail data of customer transactions at your store. You can then use the results to plan marketing or advertising strategies, or in the design of a new catalog. For instance, market basket analysis may help you design different store layouts. In one strategy, items that are frequently purchased together can be placed in proximity to further encourage the combined sale of such items. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items.

In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading toward the software display to purchase antivirus software, and may decide to purchase a home security system as well. Market basket analysis can also help retailers plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers *as well as* computers. ■

If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently *associated* or purchased together. These patterns can be represented in the form of **association rules.** For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in the following association rule:

$$computer \Rightarrow antivirus\_software \ [support = 2\%, confidence = 60\%]. \qquad (6.1)$$

Rule **support** and **confidence** are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Rule (6.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a **minimum support threshold** and a **minimum confidence threshold.** These thresholds can be a set by users or domain experts. Additional analysis can be performed to discover interesting statistical correlations between associated items.

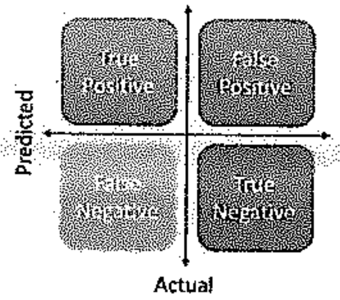
| itemset | sup_coun | itemset | sup_count |
|---|---|---|---|
| I1,I2,I3 | 2 | I1,I2,I3 | 2 |
| I1,I2,I5 | 2 | I1,I2,I5 | 2 |

Q4 (b)

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



Precision = 90/230 = 39.13% [02]

Recall = 90/300 = 30.00% [02]

Accuracy = 96.40 % [02]

□ The old value in the source system needs to be discarded
□ The change in the source system need not be preserved in the data warehouse

***Applying Type 1 Changes to the Data Warehouse.*** Please look at Figure 11-2 showing the application of Type 1 changes to the customer dimension table. The method for applying Type 1 changes is:

□ Overwrite the attribute value in the dimension table row with the new value □ The old value of the attribute is not preserved
□ No other changes are made in the dimension table row
□ The key of this dimension table or any other key values are not affected
□ This type is easiest to implement

**Type 2 Changes: Preservation of History**

***Nature of Type 2 Changes.*** Go back to the change in the marital status for Kristin Samuelson. Assume that in your data warehouse one of the essential requirements is to track orders by marital status in addition to tracking by other attributes. If the change to marital status happened on October 1, 2000, all orders from Kristin Samuelson before that

he types of changes we have discussed for marital status and customer address are Type 2 changes. Here are the general principles for this type of change:

□ They usually relate to true changes in source systems
□ There is a need to preserve history in the data warehouse
□ This type of change partitions the history in the data warehouse □ Every change for the same attribute must be preserved

***Applying Type 2 Changes to the Data Warehouse.*** Please look at Figure 11-3 showing the application of Type 2 changes to the customer dimension table. The method for applying Type 2 changes is:

□ Add a new dimension table row with the new value of the changed attribute □ An effective date field may be included in the dimension table
□ There are no changes to the original row in the dimension table
□ The key of the original row is not affected
□ The new row is inserted with a new surrogate key **Type 3 Changes: Tentative Soft Revisions**

***Nature of Type 3 Changes.*** Almost all the usual changes to dimension values are either Type 1 or Type 2 changes. Of these two, Type 1 changes are more common. Type 2 changes preserve the history. When you apply a Type 2 change on a certain date, that date is a cut-off point. In the above case of change to marital status on October 1, 2000, that date is the cut-off date. Any orders from the customer prior to that date fall into the older orders group; orders on or after that date fall into the newer orders group. An order for this customer has to fall in one or the other group; it cannot be counted in both groups for any period of time.

ere are the general principles for Type 3 changes:

- □ They usually relate to "soft" or tentative changes in the source systems
- □ There is a need to keep track of history with old and new values of the changed attribute
- □ They are used to compare performances across the transition
- □ They provide the ability to track forward and backward

  ***Applying Type 3 Changes to the Data Warehouse.*** Please look at Figure 11-4 showing the application of Type 3 changes to the customer dimension table. The methods for applying Type 3 changes are:

- □ Add an "old" field in the dimension table for the affected attribute
- □ Push down the existing value of the attribute from the "current" field to the "old" q4field
- □ Keep the new value of the attribute in the "current" field
- □ Also, you may add a "current" effective date field for the attribute
- □ The key of the row is not affected

⍰ No new dimension row is needed
⍰ The existing queries will seamlessly switch to the "current" value
⍰ Any queries that need to use the "old" value must be revised accordingly
⍰ The technique works best for one "soft" change at a time
⍰ If there is a succession of changes, more sophisticated techniques must be devised

| | | |
|---|---|---|
| Q5 (b) | | [07] |