Topper's Solutions

....In Search of Another Topper

Data Warehousing & Mining

Sem-8 (Computer)

(As Per Revised Syllabus w.e.f 2015 - 2016)

Data Warehousing & Mining

7.	Data Preprocessing	Why Preprocessing? Data Cleaning; Data Integration; Data Reduction: Attribute subset selection, Histograms, Clustering and Sampling; Data Transformation & Data Discretization: Normalization, Binning, Histogram Analysis and Concept hierarchy generation.	33
8.	Classification	8.1 Basic Concepts; **Classification methods:** 1. Decision Tree Induction: Attribute Selection Measures, Tree pruning. 2. Bayesian Classification: Naïve Bayes' Classifier. 8.2 **Prediction:** Structure of regression models; Simple linear regression, Multiple linear regression. 8.3 **Model Evaluation & Selection:** Accuracy and Error measures, Holdout, Random Sampling, Cross Validation, Bootstrap; Comparing Classifier performance using ROC Curves. 8.4 **Combining Classifiers:** Bagging, Boosting, Random Forests.	36
9.	Clustering	What is clustering? Types of data, Partitioning Methods (K-Means, KMedoids) Hierarchical Methods(Agglomerative , Divisive, BRICH), Density-Based Methods (DBSCAN, OPTICS)	44
10	Mining Frequent Pattern and Association Rule	Market Basket Analysis, Frequent Itemsets, Closed Itemsets, and Association Rules; Frequent Pattern Mining, Efficient and Scalable Frequent Itemset Mining Methods, The Apriori Algorithm for finding Frequent Itemsets Using Candidate Generation, Generating Association Rules from Frequent Itemsets, Improving the Efficiency of Apriori, A pattern growth approach for mining Frequent Itemsets; Mining Frequent Itemsets using vertical data formats; Mining closed and maximal patterns; Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules; From Association Mining to Correlation Analysis, Pattern Evaluation Measures; Introduction to Constraint-Based Association Mining.	52

Marks Distributions

#	Chapter Name	May 2016	Dec 2016
1.	Introduction to Data Warehousing.	15	15
2.	Dimensional Modeling.	15	15
3.	ETL Process.	10	15
4.	Online Analytical Processing (OLAP).	20	10
5.	Introduction to data mining.	10	15
6.	Data Exploration.	-	-
7.	Data Preprocessing.	10	-
8.	Classification.	15	25
9.	Clustering.	10	10
10.	Mining Frequent Pattern and Association Rule.	20	10
11.	Miscellaneous.	-	10
-	**Repeated Questions**	-	**20**

*** **Note:** If you need some additions questions which are not included then do send the questions on Support@ToppersSolutions.com or Whatsapp it on +917507531198 ***

If possible then we will provide softcopy for the same.

CHAPTER - 1: INTRODUCTION TO DATA WAREHOUSING

Q1] ILLUSTRATE THE ARCHITECTURE OF A TYPICAL DW SYSTEM. DIFFERENTIATE DW AND DATA MART.

Q2] DIFFERENTIATE DATA WAREHOUSE VS DATA MART.

ANS: **[Q1 | 10M – MAY16] & [Q2 | 5M – DEC16]**

DATA WAREHOUSE:

1. Data Warehouse is constructed by **integrating data** from multiple heterogeneous sources.
2. It is integrated, subject-oriented, time-variant and non-volatile collection of data.
3. It was defined by **Bill Inmon** in 1990.
4. Data Warehouse is a system used for **reporting** and **data analysis**.
5. It is considered as a core component of **business intelligence**.

ARCHITECTURE OF TYPICAL DATA WAREHOUSE:

Figure 1.1 shows Typical Data Warehouse Architecture.

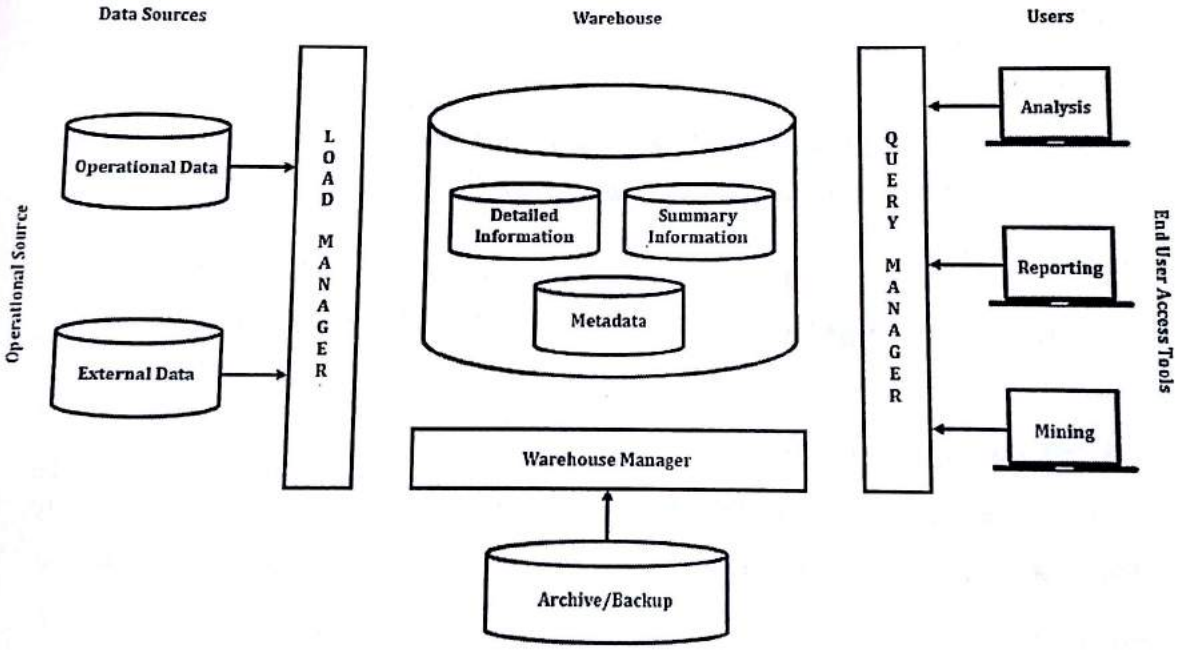


Figure 1.1: Typical Data Warehouse Architecture.

Data Warehouse Architecture consist of following components:

I) Operational Source:

- Operational Source is a data source consists of Operational Data and External Data.
- Data can come from Relational DBMS like Oracle, Informix.

II) **Load Manager:**

- Load Manager performs all the operations required to **Extract and Load Data.**
- The size and complexity of the load manager varies between specific solutions from one data warehouse to other.

III) **Warehouse Manager:**

- Warehouse Manager is responsible for the **warehouse management process.**
- Operations performed by warehouse manager includes:
  - Analysis of Data.
  - Transformation and Merging.
  - Generation of Aggregation.
  - Backing up and Archiving of Data.
  - De-normalization.

IV) **Query Manager:**

- Query Manager performs all the operations associated with management of user queries.
- The complexity of a query manager is determined by facilities provided by the end users access tools and database.

V) **Detailed Data:**

- It is used to store all the **detailed data** in the database schema.
- Detailed Data is loaded into the data warehouse to supplement the aggregated data.

VI) **Summarized Data:**

- Summarized Data is a part of data warehouse that **stores predefined aggregations.**
- These aggregations are generated by the warehouse manager.

VII) **Archive and Backup Data:**

- The Detailed and Summarized Data are stored for the purpose of archiving and backup.
- The data is transferred to storage archives such as magnetic tapes or optical disks.

VIII) **Metadata:**

- Metadata is basically **Data stored above Data.**
- It is used for extraction and loading process, warehouse management process and query management process.

IX) **End User Access Tools:**

- End User Access Tools consists of **Analysis, Reporting and Mining.**
- The users interacts with warehouse using end user access tools.

DIFFERENTIATE BETWEEN DATA WAREHOUSE AND DATA MART:

Table 1.1 shows the difference between Data Warehouse and Data Mart.

Table 1.1: Difference between Data Warehouse and Data Mart.

Parameters	Data Warehouse	Data Mart
Scope	Enterprise Level.	Department Level.
Approach	Top – Down Approach is used.	Bottom – Up Approach is used.
Centralized & Planned	Yes	No
Size	100 GB to 1 TB.	< 100 GB.
Initial effort, cost, Risk	Higher.	Lower.
Data Sources Used	Many Data Sources are required.	Few Data Sources are required.
Nature	Highly Flexible.	It is restrictive.
Implementation Time Required	Implementation takes Months to Year.	Implementation is done usually in months.
Subjects	Multiple Subjects.	Single Subject.
Data Available	Data is historical, detailed and summarized.	Data consists of some history, detailed and summarized.

Q3] WHAT IS MEANT BY METADATA IN THE CONTEXT OF A DATA WAREHOUSE? EXPLAIN THE DIFFERENT TYPES OF META DATA STORED IN A DATA WAREHOUSE. ILLUSTRATE WITH A SUITABLE EXAMPLE.

ANS: **[10M – DEC16]**

METADATA:

1. Metadata is simply defined as **Data about Data.**
2. The data that is used to represent other data is known as **metadata.**
3. For **example**, the index of a book serves as a metadata for the contents in the book.
4. In other words, we can say that metadata is the summarized data that leads us to detailed data.
5. In terms of data warehouse, we can define metadata as follows:
   a. Meta Data is the **road-map** to a data warehouse.
   b. Meta Data in a data warehouse defines the **warehouse objects.**
   c. Metadata acts as a **directory.** This directory helps the decision support system to locate the contents of a data warehouse.

ROLES OF METADATA:

1. The following figure 1.2 shows the roles of metadata.
2. Roles of metadata includes:
   a. It is used for query tools.
   b. It is used in extraction and cleansing tools.
   c. It is used in reporting tools.
   d. It is used in transformation tools.
   e. It plays an important role in loading functions.

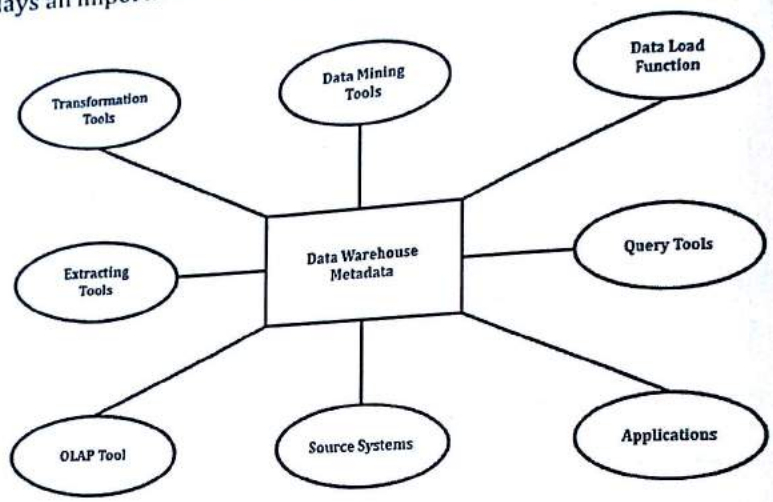


Figure 1.2: Roles of Metadata.

TYPES OF METADATA:

Metadata in a data warehouse fall into three major categories as shown in figure 1.3.

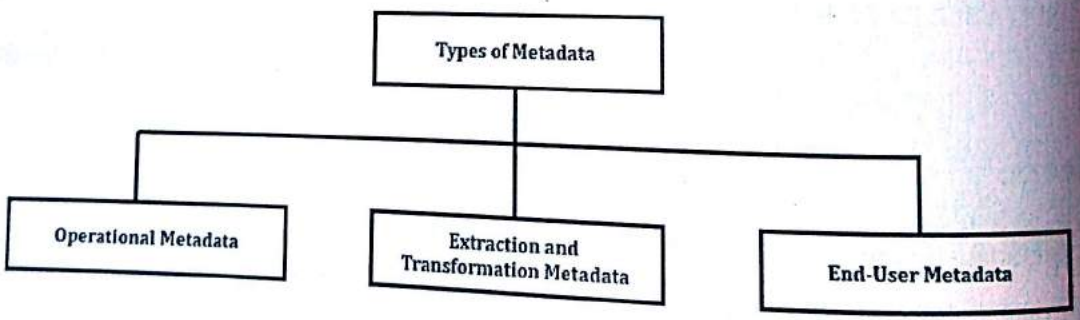


Figure 1.3: Types of Metadata.

I) **Operational Metadata:**

- In Data Warehouse, Data comes from **several operational systems** of the enterprise.
- Different source systems contain **different data structures.**
- The data elements selected for the data warehouse have various field lengths and data types.
- So the information of operational data source is given by Operational Metadata.

II) **Extraction and Transformation Metadata:**

- Extraction and transformation metadata contains the information about **extraction of data** from heterogeneous source system.
- It also contains the information about **data transformation** in data staging area.

III) **End-User Metadata:**

- The end-user metadata is the **navigational map** of the data warehouse.
- It enables the end-users to find information from the data warehouse.
- The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

EXAMPLE OF METADATA:

Topper's Solutions Customer Sales Data Warehouse.

Entity Name: Customer.
Alias Name: Account, Client.
Definitions: A Person that purchases the solutions.
Source Systems: Online Sales.
Responsible User: Sagar Narkar.
Data Quality Reviewed: 07-Mar-2017.

Q4] OPERATIONAL VS. DECISIONAL SUPPORT SYSTEM.

ANS: [5M – MAY16]

COMPARISON BETWEEN OPERATIONAL SYSTEM AND DECISIONAL SUPPORT SYSTEM:

Table 1.2 shows the difference between Operational System and Decisional Support System.

Table 1.2: Comparison between Operational System and Decisional Support System.

Operational System	Decisional Support System
It is Application Oriented.	It is Subject Oriented.
It uses Detailed Data.	It uses Summarized Data.
It contains isolated data.	It contains integrated data.
It is used to run business.	It is used to analyze business.
It is performance sensitive.	It is not performance sensitive.

There is no data redundancy.	There is data redundancy.
It has repetitive access.	It has Adhoc access.
It has up to date data.	It has snapshot of data.
Database size is 100 MB – 100 GB.	Database size is 100 GB – few TB.
Only few records can be accessed at a time.	Large volume of data can be accessed at a time.

*** EXTRA QUESTIONS ***

Q1] DATA WAREHOUSE DESIGN STRATEGIES OR APPROACHES.

ANS:

TOP DOWN APPROACH:

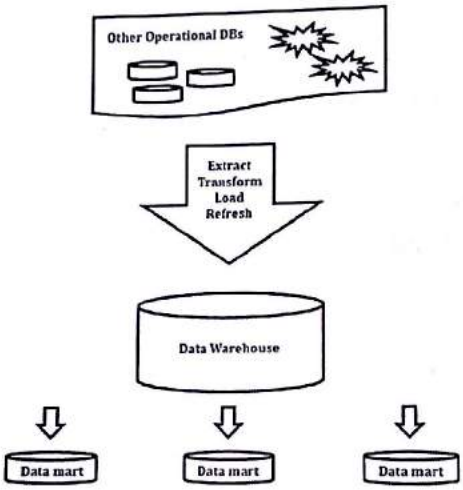


Figure 1.4: Top Down Approach.

1. Figure 1.4 shows the Top Down Approach for Data Warehouse.
2. In this approach, the data flow begins with **data extraction** from the operational data sources.
3. This data is then loaded into **staging area**.
4. It is then transferred to **Operational Data Store (ODS)**.
5. Sometimes the ODS step is skip, if it is replication of the operational databases.
6. Data is also loaded into data warehouse in a parallel process to avoid extracting it from the ODS.
7. Then the data mart is loaded with the data.
8. And finally OLAP environment is available to the users.

Advantages:

- The data is centralized.
- Results can be obtained quickly.

Disadvantages:

- Time consuming process.
- Failure risk is very high.

BOTTOM UP APPROACH:

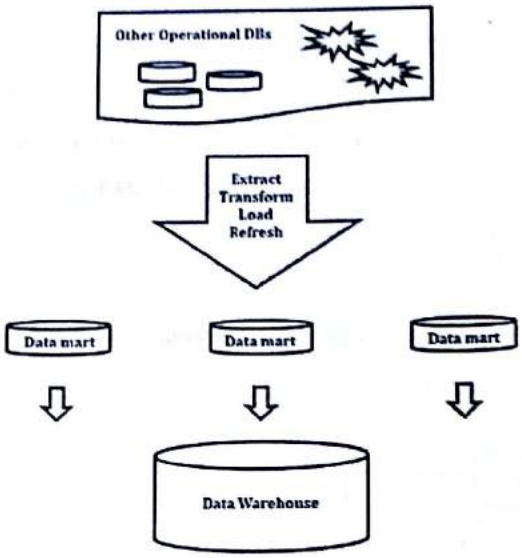


Figure 1.5: Bottom Up Approach.

1. Figure 1.5 shows the Bottom Up Approach for Data Warehouse.
2. The position of data warehouse and data mart are reversed in bottom up approach.
3. The data flow begins with **extraction of data** from operational databases into the staging area.
4. Data is then loaded into **Operational Data Store (ODS)**.
5. The data in ODS is appended to or replaced by the fresh data being loaded.
6. Once the ODS is refreshed the current data is once again extracted into the staging area.
7. It is then processed to fit into the data mart structure.
8. The data from the data mart is then extracted to the staging area aggregated, summarized and so on.
9. It is then loaded into the data warehouse.
10. Finally it is made available to the end user for analysis.

Advantages:

- Data Marts can be delivered more quickly.
- Risk of failure is low.

Disadvantages:

- Redundancy of data in data mart.
- It preserve inconsistent and incompatible data.

Q2] THREE TIER/ MULTI-TIER DATA WAREHOUSE ARCHITECTURE.

ANS:

DATA WAREHOUSE:

1. Data Warehouse is constructed by **integrating data** from multiple heterogeneous sources.
2. It is integrated, subject-oriented, time-variant and non-volatile collection of data.
3. It was defined by **Bill Inmon** in 1990.
4. Data Warehouse is a system used for **reporting** and **data analysis**.
5. It is considered a core component of **business intelligence**.

MULTI-TIER ARCHITECTURE OF DATA WAREHOUSE:

Figure 1.6 shows Multi-Tier Architecture of Data Warehouse.

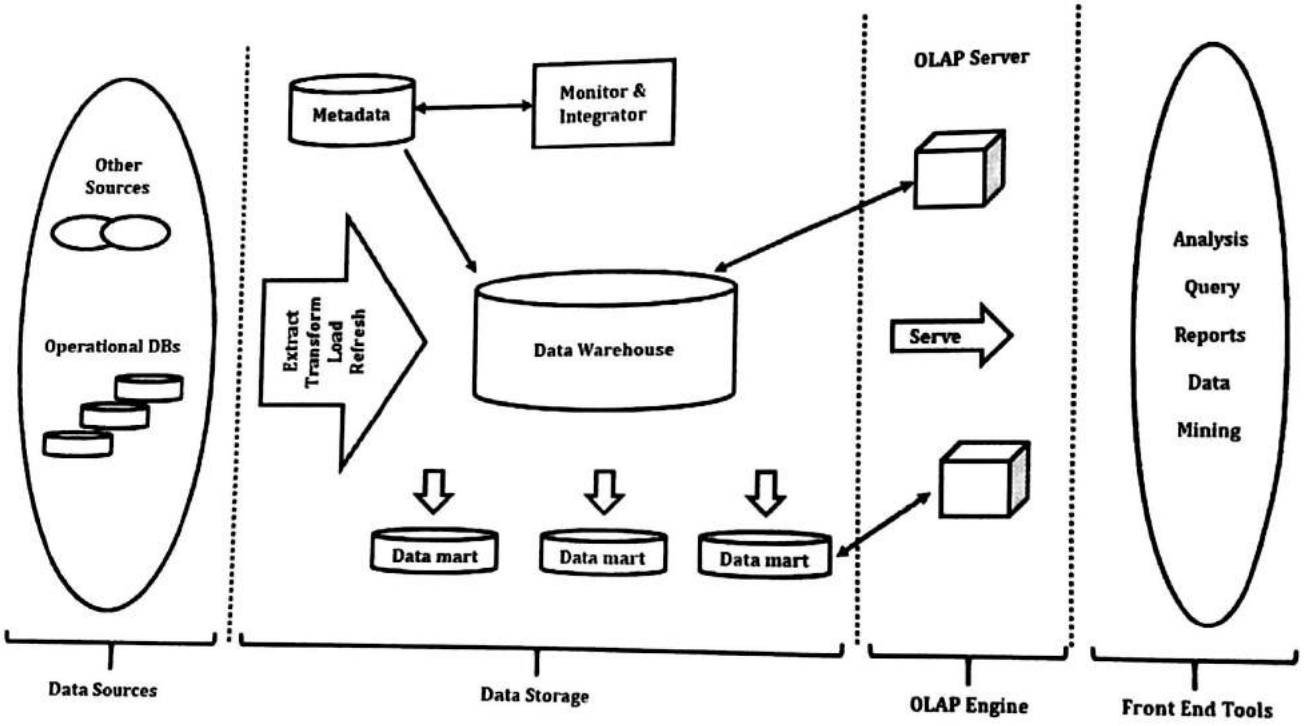


Figure 1.6: Multi-Tier Data Warehouse Architecture.

Multi-Tier Data Warehouse Architecture consists of following components:

I) **Bottom Tier:**

- Bottom Tier usually consists of **Data Sources** and **Data Storage**.
- It is warehouse database server. For Example: **RDBMS**.
- In Bottom Tier, using application program interface, data is extracted from operational and external sources.
- Application Program Interface likes ODBC, OLE-DB, JDBC is supported.

II) Middle Tier:

- Middle Tier usually consists of **OLAP Engine.**
- OLAP Engine is either implemented using **Relational OLAP (ROLAP) or Multidimensional OLAP (MOLAP).**

III) Top Tier:

- Top Tier includes **front end tools.**
- Front end tools includes query and reporting tools, analysis tools and data mining tools.
- There are three data warehouse models present.
- **Enterprise Warehouse:** The information of the entire organization is collected related to various subjects in enterprise warehouse.
- **Data Mart:** It is a subset of warehouse that is useful to a specific group of users.
- **Virtual Warehouse:** It is set of view over operational databases.

CHAPTER - 2: DIMENSIONAL MODELING

Q1] FACTLESS FACT TABLE **[5M – MAY16]**

ANS:

FACT TABLE:

1. Fact Table is a collection of **facts** and **measures.**
2. It is located at the center of a **star schema** or a **snowflake schema** surrounded by dimension tables.

FACTLESS FACT TABLE:

1. A Factless fact table is fact table that does not contain fact.
2. They contain only **dimensional keys.**
3. It captures events that happen only at information level.
4. A Factless fact table captures the **many-to-many relationships** between dimensions.
5. Factless fact tables are used for **tracking a process or collecting stats.**

TYPES OF FACTLESS FACT TABLE:

As shown in figure 2.1, there are two types of factless fact tables: those that describe events, and those that describe conditions.

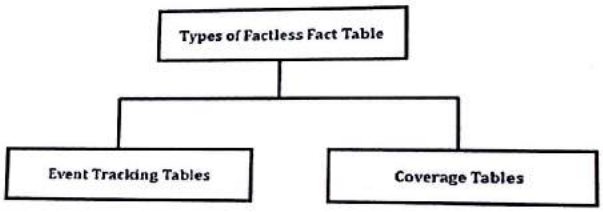


Figure 2.1: Types of Factless Fact Table.

I) **Event Tracking Tables:**

- Event Tracking Tables is used to track the event of interest.
- Many event-tracking tables in dimensional data warehouses turn out to be factless.

II) **Coverage Tables:**

- Coverage Tables was defined by Ralph.
- It is used to support negative analysis report.

EXAMPLE OF FACTLESS FACT TABLE:

- Tracking student attendance.
- List of people for the web click.

Q2] UPDATES TO DIMENSION TABLES.

ANS: **[5M – MAY16]**

DIMENSION TABLE:

1. A dimension table is a table in a **star schema** of a data warehouse.
2. A dimension table stores attributes, or dimensions, that describe the objects in a fact table.

UPDATES TO DIMENSIONS TABLES:

1. Over the time, every day as more and more sales take place, more and more rows get added to the fact table.
2. Updations due to change in fact table happens very rarely.
3. Now consider the dimension tables. Compared to the fact table, the dimension tables are more **stable** and **less volatile**.
4. Dimension table changes due to **change in attributes** themselves but not because of increase in number of rows.
5. Types of changes that affect dimension tables are as follows:

I) <u>Slowly Changing Dimensions:</u>

- Dimensions are generally constant over time, but if not constant then it may change slowly.
- **Example:** Customer ID of the record remain same but the **marital status or location** of customer may change over time.
- There are three different types:
  - **Type 1 Change:** It is related to **correction of errors** in source systems and changes are not preserved.
  - **Type 2 Change:** It is related to the **true changes** in source systems and changes are preserved.
  - **Type 3 Change:** It is related to **tentative changes** in the source systems and changes are preserved.

II) <u>Large Dimension Tables:</u>

- Large Dimensions tables are **very deep and wide.**
- Deep means it has large numbers of rows.
- Wide means it may have many attributes or columns.
- To handle large dimensions table, we can divide large dimension into some mini dimensions based on the interest.

II) <u>Rapidly Changing Dimensions:</u>

- If the dimension table changes rapidly then break the dimension table into one or more smaller dimension tables.

- Move the rapidly changing attributes in another dimension table and leave the original dimension table with slowly changing attributes.

IV) **Junk Dimensions:**

- Some textual data or flags cannot be the significant fields in major dimensions of source legacy systems.
- Although it cannot be discarded.
- So create a single **Junk dimension** and keep all meaningful text and flags into it.
- Such junk dimensions are useful to fire the quires based on flags or text values.

Q3] FOR A SUPER MARKET CHAIN, CONSIDER THE FOLLOWING DIMENSIONS NAMELY PRODUCT, STORE, TIME & PROMOTION. THE SCHEMA CONTAINS A CENTRAL FACT TABLE FOR SALES.

i. Design star schema for the above application.

ii. Calculate the maximum number of base fact tables records for warehouse with the following values given below:

- **Time period – 5 Years.**
- **Store – 300 stores reporting daily sales.**
- **Product – 40,000 products in each store (about 4000 sell in each store daily)**

[10M – MAY16]

ANS:

STAR SCHEMA:

1. Star Schema is the most popular schema design for a Data Warehouse.
2. It is called a star schema because the diagram resembles a star, with points radiating from a center.
3. The center of the star consists of **fact table** and the points of the star are the **dimension tables**.
4. Usually the fact tables in a star schema are in **third normal form (3NF)** whereas dimensional tables are de-normalized.

STAR SCHEMA FOR SUPER MARKET CHAIN:

Figure 2.2 shows the Star Schema for Super Market Chain.

Fact Table: Sales.

Dimension Table: Product, Store, Time and promotion.

There are 40,000 Products, 300 Stores and Time period of 5 Years.

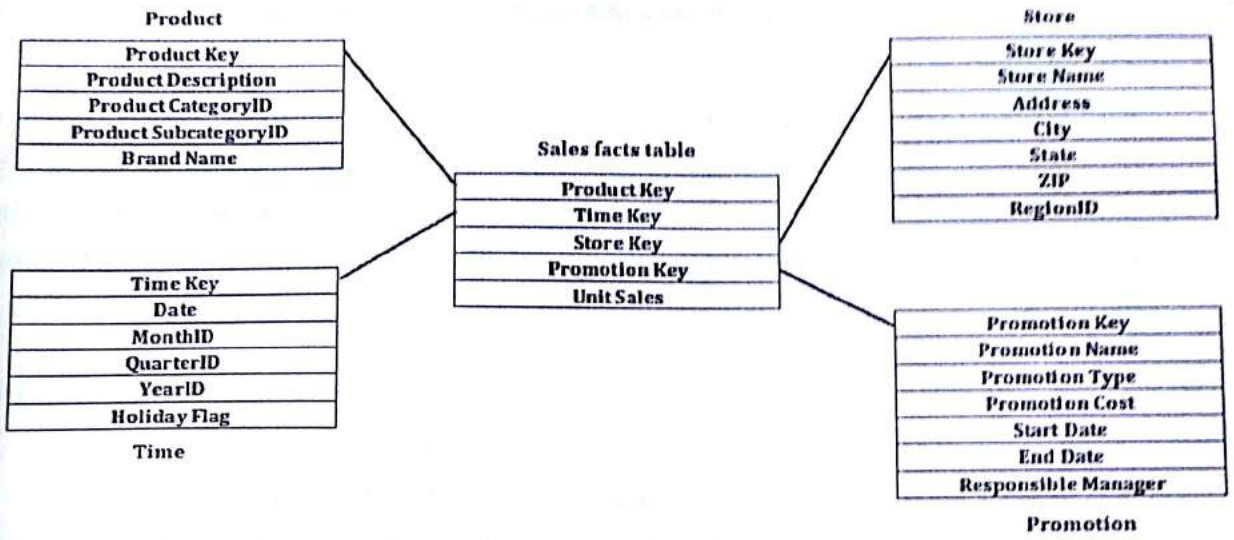


Figure 2.2: Sales Promotion Star Schema.

Maximum No. of fact table records:

Time Period = 5 Years x 365 Days

= 1825

No. of Stores = 300

Each Stores daily sale = 4000

Promotion = 1

Maximum No. of Fact Table Records = Time Period x No. of Stores x Daily Sale x Promotion.

= 1825 x 300 x 4000 x 1

= 2,190,000,000

Maximum No. of Fact Table Records = 2 Billion.

Q4] CONSIDER FOLLOWING DIMENSIONS FOR A HYPERMARKET CHAIN: PRODUCT, STORE, TIME AND PROMOTION.

With respect to this business scenario, answer the following questions. Clearly state any reasonable assumptions you make. Design a star schema. Whether the star schema can be converted to snowflake schema? Justify your answer and draw snowflake schema for the data warehouse (clearly mention the Fact table(s), Dimension table(s), their attributes and measures)

[10M – MAY16]

ANS:

STAR SCHEMA:

1. Star Schema is the most popular schema design for a Data Warehouse.
2. It is called a star schema because the diagram resembles a star, with points radiating from a center.
3. The center of the star consists of **fact table** and the points of the star are the **dimension tables.**
4. Usually the fact tables in a star schema are in **third normal form (3NF)** whereas dimensional tables are de-normalized.

STAR SCHEMA FOR HYPER MARKET CHAIN:

Figure 2.3 shows the Star Schema for Hyper Market Chain.

Fact Table: Sales.

Dimension Table: Product, Store, Time and promotion.

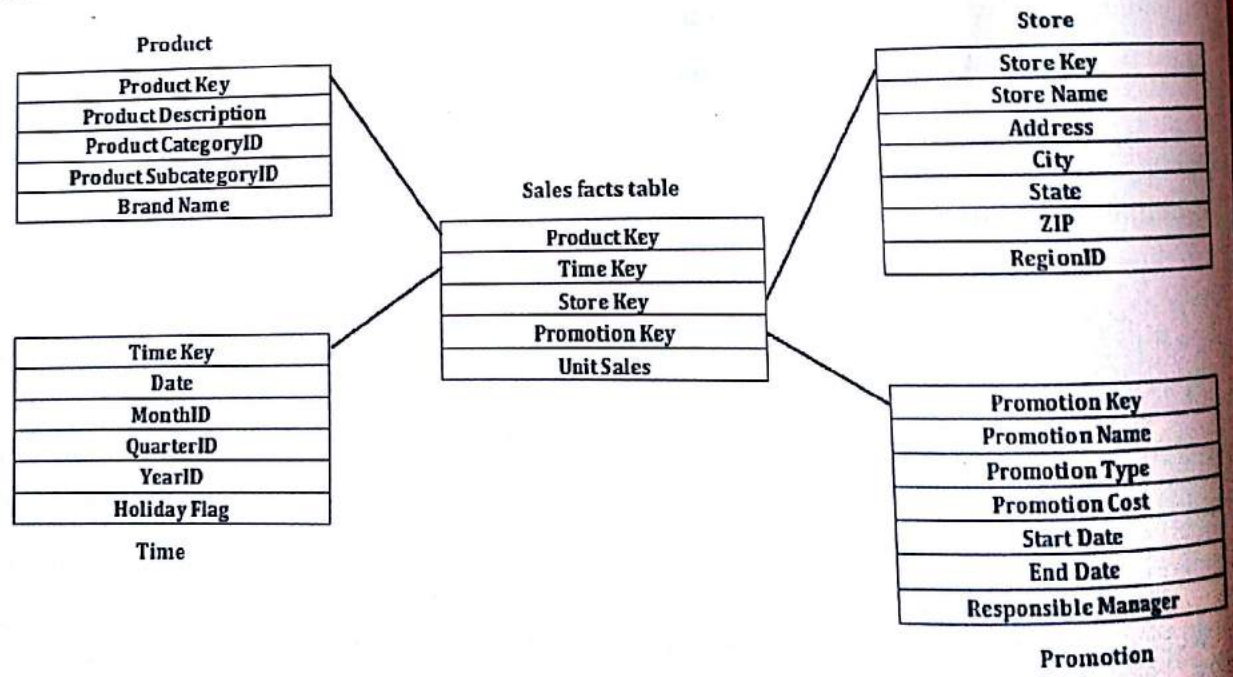


Figure 2.3: Sales Promotion Star Schema.

Whether the star schema can be converted to snowflake schema?

Yes the above star schema can be converted to snowflake schema by considering the following assumptions:

1. Product can be classified into category and subcategory.
2. Store belongs to a region, and a region dimension is not added in star schema.
3. Time Dimensions can be further divided into Month, Quarter and Year.
4. Promotion can be further classified into types.

SNOWFLAKE SCHEMA:

1. The snowflake schema is an **extension** of the star schema, where each point of the star explodes into more points.
2. In a star schema, each dimension is represented by a single dimensional table.
3. Whereas in a snowflake schema, that dimensional table is **normalized** into multiple lookup tables, each representing a level in the dimensional hierarchy.
4. Figure 2.4 shows the Snowflake Schema for Hyper Market Chain.

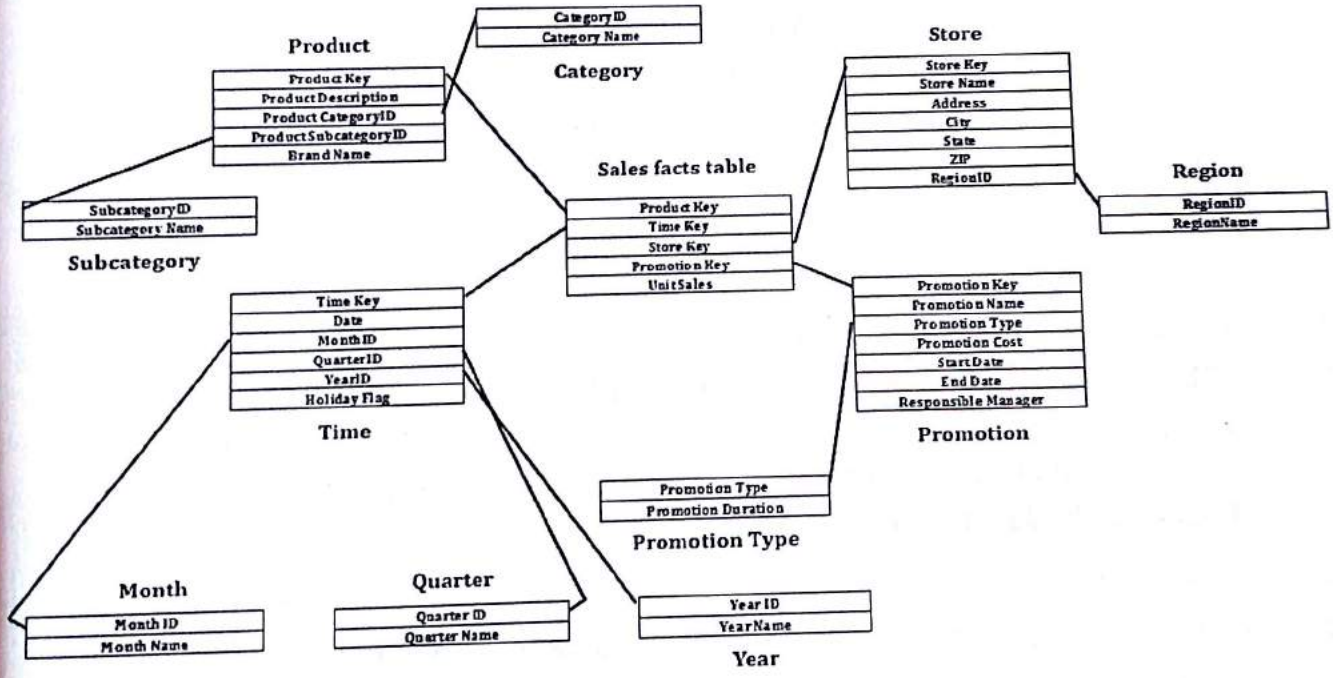


Figure 2.4: Sales Promotion Snowflake Schema.

CHAPTER - 3: ETL PROCESS

Q1] DESCRIBE THE STEPS OF ETL PROCESS. **[10M – MAY'16]**

ANS:

ETL:

1. ETL Stands for **Extract, Transform** and **Load.**
2. It is a process in data warehousing responsible for **pulling data out of the source systems and placing it** into a data warehouse.

ETL PROCESS:

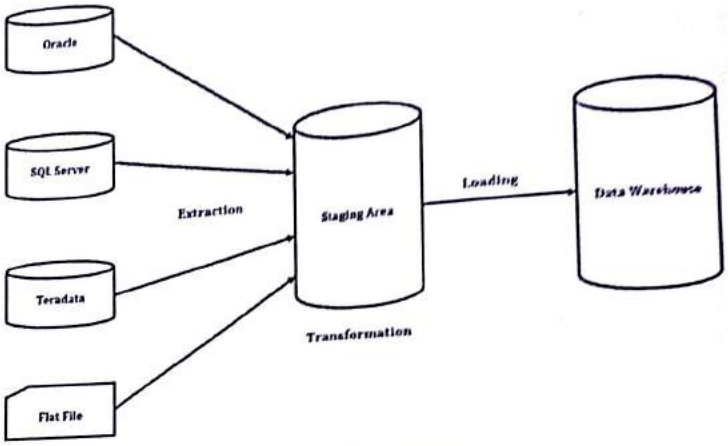


Figure 3.1: ETL Process.

1. Figure 3.1 represents the ETL Process.
2. ETL process is the way to move and prepare data for data analysis.
3. ETL process involves the following tasks:

I) <u>Extracting the data from different sources:</u>

- This is the first step in ETL process.
- Different data sources can be **RDBMS** or **files** like CSV, JSON, and XML etc.
- In this step, Data is **extracted** from source system.
- Data is also made accessible for further processing.
- The main objective of the extraction step is to **retrieve all required data** from source system.
- The extraction step should be designed in a way that it does not negatively affect the source system.
- Most data projects consolidate data from different source systems.
- Each separate source uses a different format.
- Common data-source formats include RDBMS, XML (like CSV, JSON).

- Thus the extraction process must convert the data into a format suitable for further transformation.

II) **Transforming the data:**

- This may involve **cleaning, filtering, validating and applying business rules.**
- In this step, certain rules are applied on the extracted data.
- The main aim of this step is to **load the data** to the target database in a cleaned and general format.
- This is because when the data is collected from different sources each source will have their own standards.
- For **example** if we have two different data sources A and B.
- In source A, date format is like dd/mm/yyyy, and in source B, it is yyyy-mm-dd.
- In the transforming step we convert these dates to a general format.
- The other things that are carried out in this step are:
  - **Cleaning** (e.g. "Male" to "M" and "Female" to "F" etc.).
  - **Filtering** (e.g. selecting only certain columns to load).
  - **Enriching** (e.g. Full name to First Name , Middle Name , Last Name).
  - **Splitting** a column into multiple columns and vice versa.
  - **Joining** together data from multiple sources.
- In some cases data does not need any transformations and here the data is said to be **"rich data"** or **"direct move"** or **"pass through"** data.

III) **Loading:**

- This is the final step in the ETL process.
- In this step, the extracted data and transformed data is **loaded** to the target database.
- In order to make data load efficient, it is necessary to **index the database** and disable constraints before loading the data.
- All the three steps in the ETL process can be run parallel.
- Data extraction takes time and so the second step of transformation process is executed simultaneously.
- This prepares data for the third step of loading.
- As soon as some data is ready, it is loaded without waiting for completion of the previous steps.

Q2] IN WHAT WAY ETL CYCLE CAN BE USED IN TYPICAL DATA WAREHOUSE, EXPLAIN WI SUITABLE INSTANCE. [10M – DEC

ANS:

ETL:

1. ETL Stands for **Extract, Transform** and **Load.**
2. It is a process in data warehousing responsible for **pulling data out** of the source systems a **placing it** into a data warehouse.

ETL CYCLE:

A typical ETL lifecycle consists of the following 10 steps of execution.

1. Initiation of cycle.
2. Building reference data.
3. Extracting data from different sources.
4. Validation of data.
5. Transforming data.
6. Staging of data.
7. Generation of audit reports.
8. Publishing data.
9. Archiving.
10. Cleanup.

ETL PROCESS:

Refer Q1.

USES OF ETL CYCLE IN TYPICAL DATA WAREHOUSE:

1. ETL is the most important step in data warehousing.
2. Data warehousing brings data from different sources onto a single platform and in a single for
3. So ETL makes analysis of the data easier and effective.
4. ETL is required in taking management decisions.
5. It is used in designing strategies and future plans.

Q3] DATA QUALITY **[5M – DEC16]**

ANS:

DATA QUALITY:

1. Data quality can simply be described as a **fitness for use of data.**
2. To be more specific every portion of data has to be accurate to clearly represent the value of itself.
3. **Data cleansing** may be required in order to ensure data quality.
4. Some of the reasons for Dirty Data are listed as follows:
   a. Dummy Values.
   b. Absence of Data.
   c. Non-Unique Identifiers.
   d. Cryptic Data.
   e. Multi-purpose Fields.

DATA QUALITY CYCLE:

Figure 3.2 shows the data quality cycle.

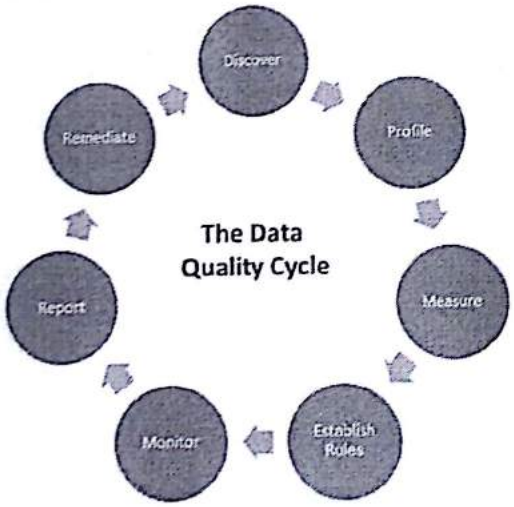


Figure 3.2: Data Quality Cycle.

Components of Data Quality Cycle includes:

I) **Data Discovery:** It is the process of finding, gathering, organizing and reporting metadata about data.

II) **Data Profiling:** It is the process of analyzing data in detail, comparing the data to its metadata, calculating data statistics and reporting the measures of quality for the data.

III) **Data Quality Rules:** Based on the business requirements for each Data Quality measure, the data quality rules are made.

IV) **Data Quality Monitoring:** It is the process of monitoring of Data Quality, based on the result executing the Data Quality rules.

V) **Data Quality Reporting:** Dashboards and scorecards are used to report Data Quality measur

VI) **Data Remediation:** It is the ongoing correction of Data Quality exceptions and issues as they reported.

CHAPTER - 4: ONLINE ANALYTICAL PROCESSING (OLAP)

Q1] DISCUSS VARIOUS OLAP MODELS.

ANS: **[10M – MAY16]**

OLAP:

1. OLAP Stands for Online Analytical Processing.
2. It is based on the multidimensional data model.
3. OLAP was defined by OLAP Council.
4. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

OLAP MODELS:

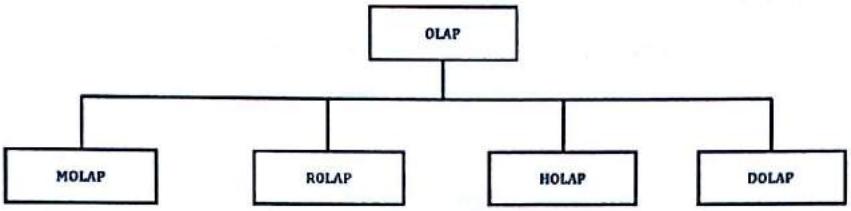


Figure 4.1: OLAP Models.

I) MOLAP:

- MOLAP Stands for **Multi-dimensional OLAP.**
- In MOLAP, data is stored in a **multidimensional cube.**
- It uses array-based multidimensional storage engines.
- The storage is not in the relational database, but in proprietary formats.
- Figure 4.2 shows MOLAP Process.

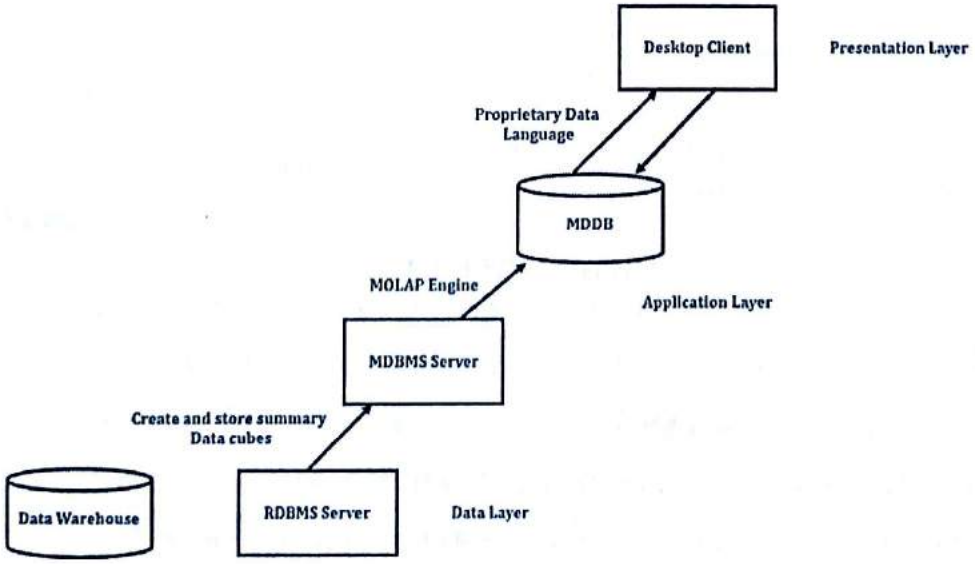


Figure 4.2: MOLAP Process.

Advantages:

- It can perform complex calculations.
- It has excellent performance.

Disadvantages:

- It can handle limited amount of data.
- It requires additional investment.

II) ROLAP:

- ROLAP Stands for **Relational OLAP**.
- ROLAP uses relational or extended relational DBMS.
- ROLAP servers are placed between **relational back-end server** and **client front-end** tools.
- Figure 4.3 shows ROLAP Process.

Advantages:

- It has higher scalability.
- It can handle large amount of data.

Disadvantages:

- Performance is slow.
- Limited SQL Functionality.

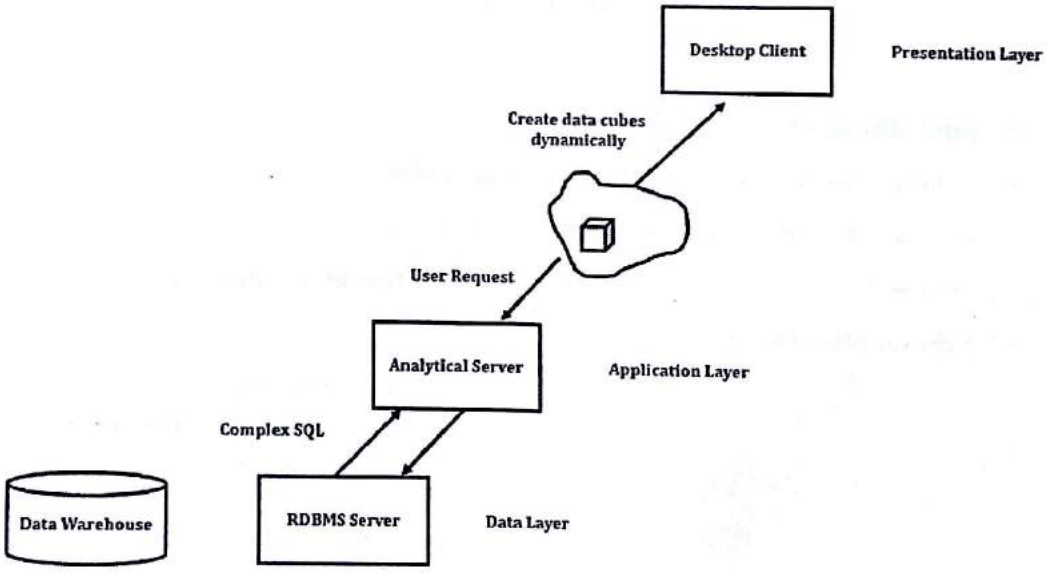


Figure 4.3: ROLAP Process.

III) HOLAP:

- HOLAP Stands for **Hybrid OLAP**.
- Hybrid OLAP is a combination of both ROLAP and MOLAP.
- It offers **higher scalability of ROLAP** and **faster computation of MOLAP**.
- HOLAP servers allows to store the large data volumes of detailed information.

V) **DOLAP:**

- DOLAP Stands for **Desktop OLAP.**
- It is variation of ROLAP.
- DOLAP requires only DOLAP software to be present on machine.
- It offers **portability to the users.**

Q2] **INDEXING OLAP DATA.**

ANS: **[5M – DEC16]**

INDEXING OLAP DATA:

1. Indexing is used to quickly locate data without having to search every row in a database.
2. Indexing provides the basis for both **rapid random lookups** and **efficient access of ordered records.**
3. Indexing OLAP Data includes **Bitmap Index and Join Indices.**

I) **Bitmap Index:**

- Bitmap Index is the **index on particular column.**
- Each value in the column has a bit vector.
- The length of the bit vector is number of records in the base table.
- The i^{th} bit is set if the i^{th} row of the base table has the value for the indexed column.
- It is not suitable for high cardinality domains.
- **Example** of Bitmap Index is shown in figure 4.4.

Base Table		
Customer	Region	Type
C1	Asia	Retail
C2	Europe	Dealer
C3	Asia	Dealer
C4	America	Retail
C5	Europe	Dealer

Index on Region			
RecID	Asia	Europe	America
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0

Index on Type		
RecID	Retail	Dealer
1	1	0
2	0	1
3	0	1
4	1	0
5	0	1

Figure 4.4: Example of Bitmap Index.

II) **Join Indices:**

- Traditional indices map the values to a **list of record ids.**
- But Join Indices map the values of the dimensions of star schema to rows in the fact table.
- Join Indices is: **JI (R-id, S-id) where R (R-id, ...) >< S (S-id, ...)**
- Join indices can span multiple dimensions.
- Figure 4.5 shows the Example of Join Indices.
- **Fact Table:** Sales and **Dimension Tables:** Location and Item.

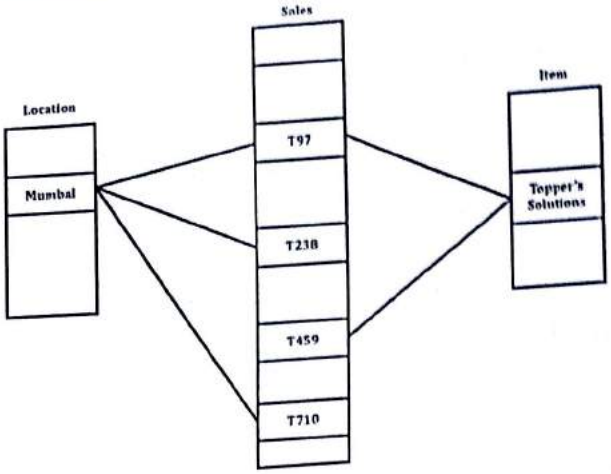


Figure 4.5: Example of Join Indices.

Q3] WE WOULD LIKE TO VIEW SALES DATA OF A COMPANY WITH RESPECT TO THREE DIMENSIONS NAMELY LOCATION, ITEM AND TIME. REPRESENT THE SALES DATA IN THE FORM OF A 3-D DATA CUBE FOR THE ABOVE AND PERFORM ROLL UP, DRILL DOWN, SLICE AND DICE OLAP OPERATIONS ON THE ABOVE DATA CUBE AND ILLUSTRATE.

[10M – MAY16]

ANS:

OLAP OPERATIONS:

1. OLAP Operations are implemented to retrieve the information from data warehouse into OLAP multi-dimensional databases.
2. Since OLAP servers are based on multidimensional view of data, so OLAP operations are performed in multidimensional data.
3. List of OLAP operations:

I) **Roll-up:**

- Roll-up performs aggregation on a data cube in any of the following ways:
  - By climbing up a concept hierarchy for a dimension.
  - By dimension reduction.
- The following figure 4.6 illustrates how roll-up works.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

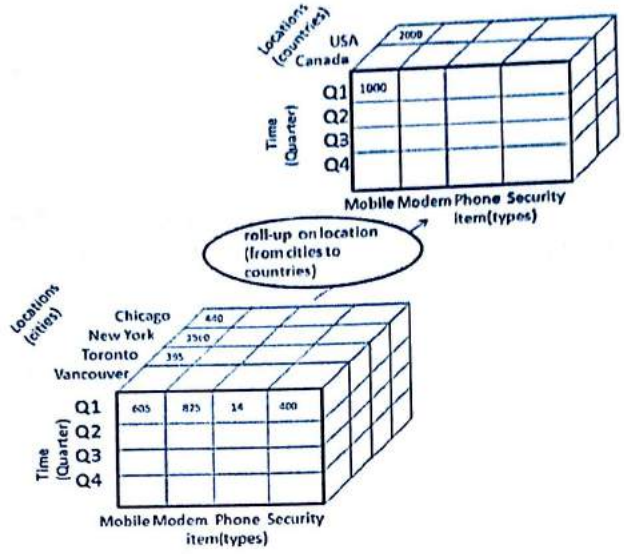


Figure 4.6: Roll-up Operation.

II) Drill-down:

- Drill-down is the reverse operation of roll-up.
- It is performed by either of the following ways:
  - By stepping down a concept hierarchy for a dimension.
  - By introducing a new dimension.
- The following figure 4.7 illustrates how drill-down works:
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

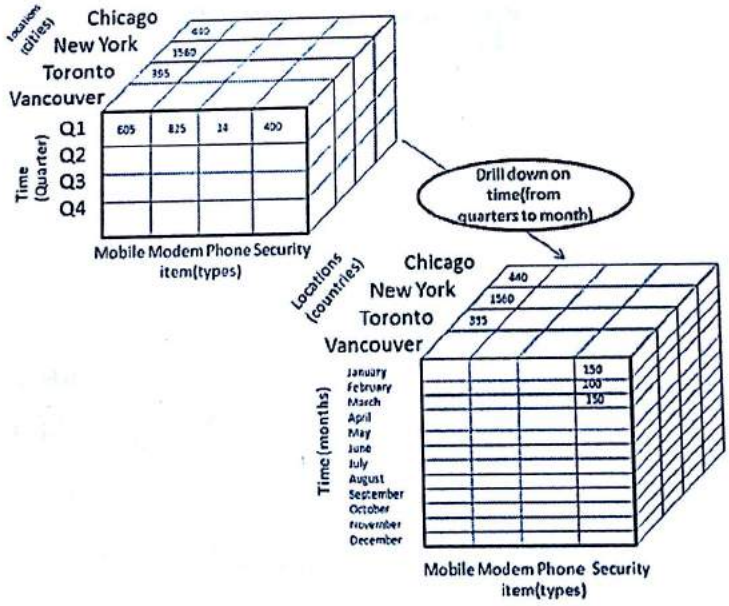


Figure 4.7: Roll-up Operation.

III) **Slice:**

- The slice operation selects one particular dimension from a given cube and provides a new sub-cube.
- Consider the following figure 4.8 that shows how slice works.
- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

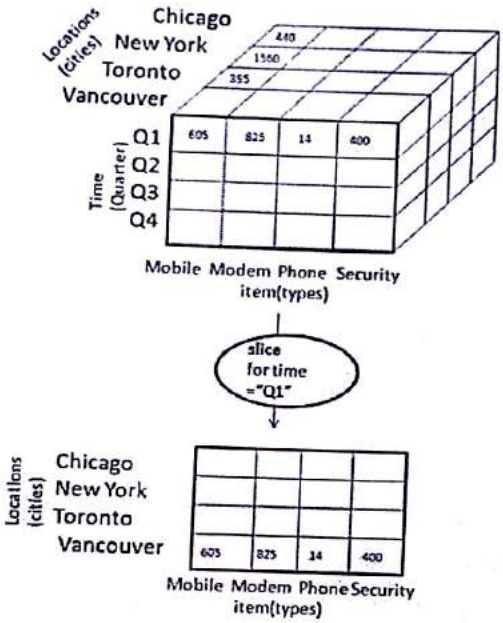


Figure 4.8: Slice Operation.

IV) **Dice:**

- Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- Consider the following figure 4.9 that shows the dice operation.

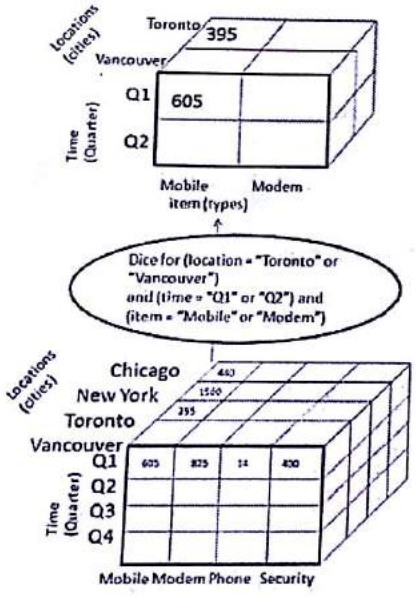


Figure 4.9: Dice Operation.

- The dice operation on the cube based on the following selection criteria involves three dimensions.
  - Location = "Toronto" or "Vancouver"
  - Time = "Q1" or "Q2"
  - Item =" Mobile" or "Modem"

V) Pivot:

- The pivot operation is also known as rotation.
- It rotates the data axes in view in order to provide an alternative presentation of data.
- Consider the following figure 4.10 that shows the pivot operation.
- In this the item and location axes in 2-D slice are rotated.

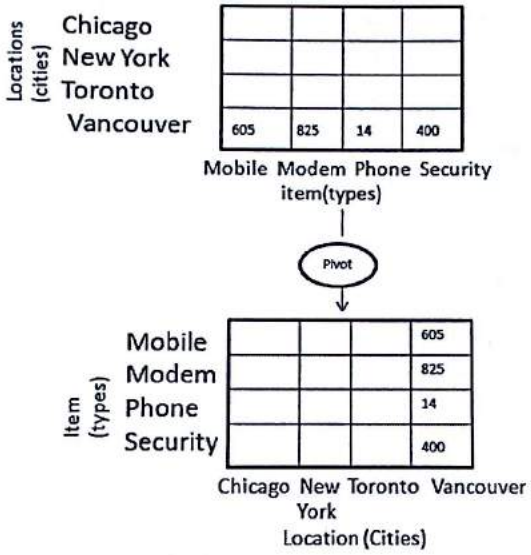


Figure 4.10: Pivot Operation.

Q4] DIFFERENTIATE OLTP VS OLAP

ANS: **[5M – DEC16]**

Table 4.1 shows the comparison between OLTP and OLAP.

Table 4.1: Comparison between OLTP and OLAP.

Parameters	OLTP	OLAP
Full Form	Online Transaction Processing.	Online Analytical Processing.
Oriented	Transaction Oriented.	Subject Oriented.

Characteristics	Operational Processing.	Informational Processing.
Data Redundancy	Data Redundancy is bad.	Data Redundancy is good.
Granularity	Few Levels of Granularity.	Multiple Level of Granularity.
Users	Many Users	Few Users.
Size	10 MB to GB	100 GB to TB.
Priority	High Performance and Availability.	High Flexibility.
Access	Read and write.	Mostly Read.
Function	It is used for Day to Day Operations.	It is used for long term informational requirements.

CHAPTER - 5: INTRODUCTION TO DATA MINING

Q1] DESCRIBE THE VARIOUS FUNCTIONALITIES OF DATA MINING AS A STEP IN THE PROCESS OF KNOWLEDGE DISCOVERY. **[10M – DEC16]**

ANS:

DATA MINING:

1. Data Mining is defined as the procedure of extracting information from huge sets of data.
2. It is a **non-trivial process**.

KDD:

1. KDD Stands for **Knowledge Discovery in Database.**
2. KDD is the process of discovering knowledge in data.
3. The main goal is to extract knowledge from large database.
4. KDD includes wide variety of application domains which includes Artificial Intelligence, Pattern Recognition, Machine Learning Statistics and Data Visualization.
5. Figure 5.1 shows the KDD Process.

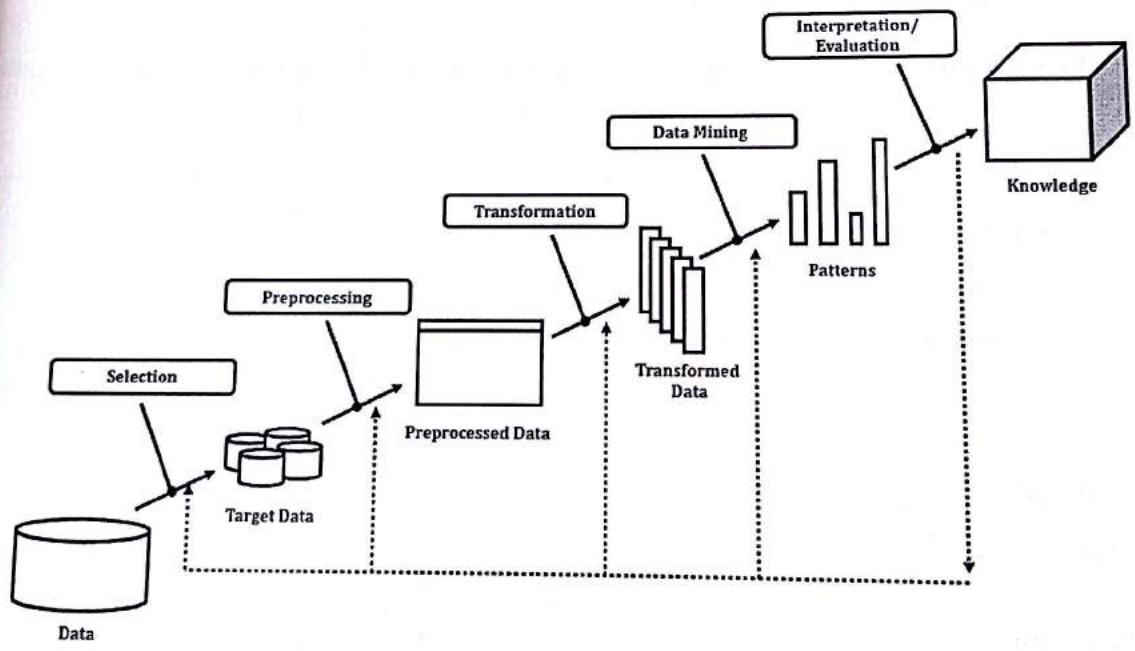


Figure 5.1: KDD Process.

List of steps involved in the knowledge discovery process:

I) Data Cleaning:

- In this step, the noise and inconsistent data is removed.

II) Data Integration:

- In this step, multiple data sources are combined.

III) **Data Selection:**

- In this step, data relevant to the analysis task are retrieved from the database.

IV) **Data Transformation:**

- In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

V) **Data Mining:**

- In this step, intelligent methods are applied in order to extract data patterns.

VI) **Pattern Evaluation:**

- In this step, data patterns are evaluated.
- It is used to identify the **truly interesting patterns** representing knowledge based on interesting measures.

VII) **Knowledge Presentation:**

- In this step, knowledge is represented.
- **Visualization** and **knowledge representation techniques** are used to present mined knowledge to users.

Q2] DISCUSS:

1. **THE STEPS IN KDD PROCESS.**
2. **THE ARCHITECTURE OF A TYPICAL DM SYSTEM.**

ANS:

[10M – MAY16]

DATA MINING:

Refer Q1.

KDD PROCESS:

Refer Q1.

ARCHITECTURE OF A TYPICAL DM SYSTEM:

Figure 5.2 shows the Architecture of a typical data mining system.

I) **Database, data warehouse, or other information repository:**

- This is **information repository.**
- Data cleaning and data integration techniques are performed on the data.

II) **Databases or data warehouse server:**

- It fetches the data as per the users' requirement which is need for data mining task.

III) **Knowledge base:**

- This is used to guide the search, and gives the interesting and hidden patterns from data.

IV) **Data mining engine:**

- It performs the data mining task such as **characterization, association, classification, cluster analysis** etc.

V) **Pattern evaluation module:**

- It is integrated with the mining module and it give the search of only the interesting patterns.

VI) **Graphical user interface:**

- This module is used to communicate between user and the data mining system.
- It allow users to browse databases or data warehouse schemas.

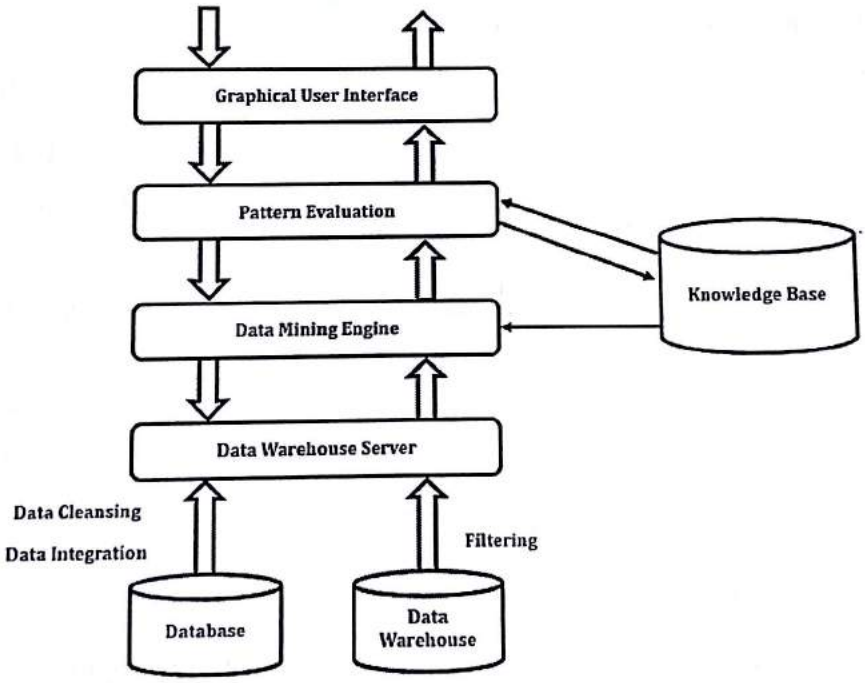


Figure 5.2: Architecture of Typical Data Mining System.

Q3] APPLICATION OF DATA MINING TO FINANCIAL ANALYSIS.

ANS:

[5M – DEC16

DATA MINING:

Refer Q1.

APPLICATION OF DATA MINING TO FINANCIAL ANALYSIS:

1. The financial data in banking and financial industry is generally **reliable and of high quality.**
2. So it facilitates the systematic data analysis and data mining.
3. Some of the typical cases are as follows:
   a. Design and construction of data warehouses for multidimensional data analysis and da mining.
   b. Loan payment prediction.
   c. Customer credit policy analysis.
   d. Classification and clustering of customers for targeted marketing.
   e. Detection of money laundering and other financial crimes.

Other Applications of Data Mining includes:

- Retail Industry.
- Telecommunication Industry.
- Biological Data Analysis.
- Other Scientific Applications.
- Intrusion Detection.

CHAPTER - 6: DATA EXPLORATION

No Questions were asked from this Chapter in May 16 and Dec 16 Paper of Mumbai University.

CHAPTER - 7: DATA PREPROCESSING

Q1] DISCUSS DIFFERENT STEPS INVOLVED IN DATA PREPROCESSING. **[10M – MAY16]**

ANS:

DATA PREPROCESSING:

1. Data pre-processing is an important step in the data mining process.
2. Data preprocessing involves **transforming** raw data into an understandable format.
3. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.
4. Data preprocessing is a proven method of resolving such issues.
5. Data preprocessing prepares raw data for further processing.

STEPS INVOLVED IN DATA PREPROCESSING:

I) Data Cleaning:

- Data Cleaning is also known as **Scrubbing**.
- It is a technique that is applied to **remove the noisy data** and **correct the inconsistencies** in data.
- It involves filling missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Steps in data cleansing:
  - **Parsing:** Parsing is the process in which individual data elements are located and identified in the source systems and then these elements are isolated in the target files.
  - **Correcting:** In this step, using data algorithm the individual data elements are corrected.
  - **Standardizing:** In standardizing process, conversion routines are used to transform data into a consistent format using both standard and custom business rules.
  - **Matching:** Matching process involves eliminating duplications by searching and matching records.
  - **Consolidating:** Consolidating process involves merging the records into one representation by analyzing and identifying relationship between matched records.

- Figure 7.1 shows example of data cleaning process.

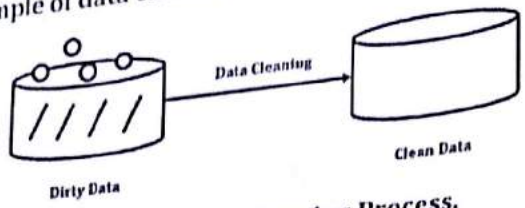


Figure 7.1: Data Cleaning Process.

II) Data Integration:

- Data integration involves combining data residing in different sources and providing users with a unified view of these data.
- Sources may include multiple databases, data cubes or data files.
- Data Integration removes the duplicate and redundant data.
- Figure 7.2 shows example of data integration process.

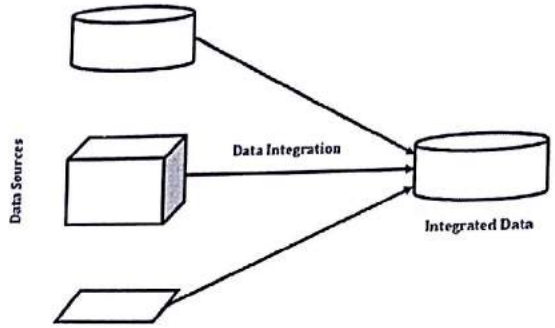


Figure 7.2: Data Integration Process.

III) Data Transformation:

- In Data Transformation, data are transformed or consolidated into forms appropriate for mining.
- Data transformation involves:
  - Smoothing.
  - Aggregation.
  - Generalization.
  - Normalization.
- Figure 7.3 shows the example of data transformation process.

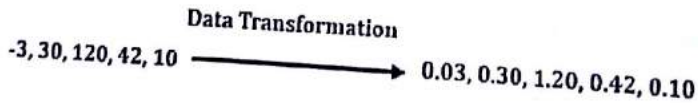


Figure 7.3: Data Transformation Process.

IV) Data Reduction:

- Data Reduction is used to obtain a reduced representation of the data set that is much smaller in volume.

- Strategies for data reduction includes:
  - **Data Cube Aggregation:** In Data Cube Aggregation, aggregation operations are applied to the data in the construction of a data cube.
  - **Attribute subset selection:** This process is used to detect and remove irrelevant, weakly relevant, or redundant attributes or dimensions.
  - **Dimensionality Reduction:** In this process encoding mechanisms are used to reduce the data set size.
  - **Numerosity Reduction:** In this process, the data are replaced by alternative, smaller data representations such as parametric models and non-parametric models like clustering.
- Figure 7.4 shows data reduction process example.

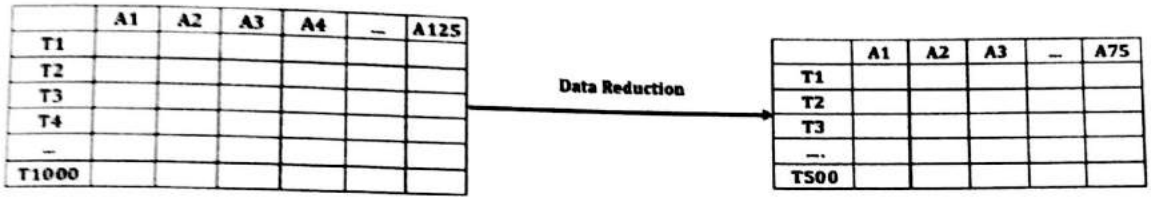


Figure 7.4: Data Reduction Process.

) **Data Discretization:**

In Data Discretization, the range of a continuous attribute is divided into intervals.

By discretization the size of the data is reduced.

In this process, the data is prepared for further analysis.

Discretization process is applied recursively on an attribute.

Three types of attributes:

- **Nominal:** Values from an unordered set.
- **Ordinal:** Values from an ordered set.
- **Continuous:** Real numbers.

CHAPTER - 8: CLASSIFICATION

Q1] DECISION TREE BASED CLASSIFICATION APPROACH.

ANS: **[5M – DEC16**

DECISION TREE BASED CLASSIFICATION:

1. It is one of the most important classification and prediction method in data mining.
2. A decision tree represents **rules.**
3. Rules are easy to understand and can be directly used in SQL to retrieve the records fro database.
4. A decision tree classifier has **tree type structure.**
5. It has leaf nodes and decision nodes.
6. A leaf node is the last node of each branch and indicates value of target attribute.
7. A decision node is the node of tree which has leaf node or sub-tree.
8. Figure 8.1 shows the representation of decision tree for tennis play.
9. As shown in figure 8.1, Humidity, Outlook and Wind is Attribute.
10. High, Normal, Strong, Weak, Sunny, Rain and Overcast is Value.
11. Yes and No is classification.

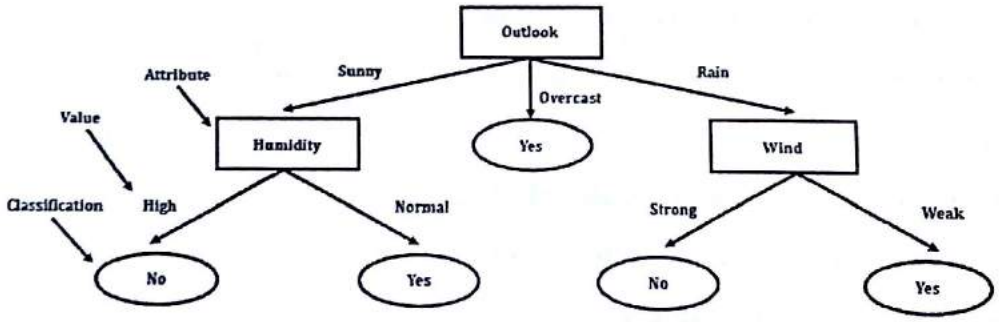


Figure 8.1: Decision tree for tennis play.

Q2] METRICS FOR EVALUATING CLASSIFIER PERFORMANCE.

ANS: **[5M – MAY1**

METRICS FOR EVALUATING CLASSIFIER PERFORMANCE:

1. **Sensitivity:** Sensitivity is defined as True Positive recognition rate which is the proportion positive tuples that are correctly identified.

$$\text{Sensitivity} = TP/P$$

2. **Specificity:** Specificity is defined as True Negative recognition rate which is the proportion negative tuples that are correctly identified.

$$\text{Specificity} = TN/N$$

3. **Classifier Accuracy:** It is percentage of test set tuples that are correctly classified.

$$\text{Accuracy} = (TP + TN) / (P + N)$$

4. **Error Rate:** It is percentage of error made over the whole set of instances used.

$$\text{Error Rate} = 1 - \text{Accuracy}$$

5. **Precision:** It is percentage of tuples which are correctly classified as positive are actual positive. It is the measure of exactness.

$$\text{Precision} = \frac{|TP|}{|TP|+|FP|}$$

6. **Recall:** It is percentage of positive tuples which the classifier labelled as positive. It is a measure of completeness.

$$\text{Recall} = \frac{|TP|}{|TP|+|FN|}$$

7. **F Measures:** It is Harmonic mean of precision and recall

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$$

Note:

TP: Class Members which are classified as class members.

TN: Class Non-Members which are classified as class non-members.

FP: Class Non-Members which are classified as class members.

FN: Class Members which are classified as class non-members.

P: Number of positive tuples.

N: Number of negative tuples.

Q3] WHY NAÏVE BAYESIAN CLASSIFICATION IS CALLED "NAIVE"? BRIEFLY OUTLINE THE MAJOR IDEAS OF NAIVE BAYESIAN CLASSIFICATION.

ANS: **[10M – DEC16]**

NAÏVE BAYESIAN CLASSIFICATION:

1. Naïve Bayesian Classification is based on **Bayes Theorem.**
2. Bayesian classifiers are the **statistical classifiers.**
3. Naïve Bayesian Classification is referred as **Naïve** because it makes the assumption that each of its inputs are independent of each other, an assumption which rarely holds true.
4. For **example,** a fruit may be considered to be an apple if it is red, round, and about 4" in diameter.
5. A Naive Bayes Classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

6. This assumption is made to reduce computational costs, and hence is considered naïve.

MAJOR IDEAS OF NAIVE BAYESIAN CLASSIFICATION:

Bayes Theorem:

1. It is also known as Bayes Rule.
2. It is used to find **conditional probabilities**.
3. **Bayes Theorem:** $P(H|X) = P(X|H)\ P(H) / P(X)$
4. An initial probability is called as **Apriori Probability** which we get before any additional information is obtained.
5. A probability is called as **Posterior Probability** which we get after any additional information is obtained.
6. P (H|X) is Posterior Probability of H and P (X|H) is Posterior Probability of X.
7. P (H) is Apriori Probability of H and P (X) is Apriori Probability of X.

Q4] DEFINE LINEAR, NON-LINEAR AND MULTIPLE REGRESSIONS. PLAN A REGRESSION MODEL FOR DISEASE DEVELOPMENT WITH RESPECT TO CHANGE IN WEATHER PARAMETERS.

[10M – DEC16]

ANS:

REGRESSION:

1. Regression is the method for **prediction** in data mining.
2. Regression shows a relationship between the average values of two variables.
3. Thus regression is very useful in estimating and predicting the average value of one variable for a given value of other variable.
4. The estimate or prediction may be made with the help of a regression line.
5. Regression may be used to determine for e.g. price of commodity, interest rates etc.

TYPES OF REGRESSION:

I) Linear Regression:

- If the regression curve is a **straight line** then there is a linear regression between two variables.
- The relationship between dependent and independent variable is described by straight line and it has only one independent variable.

$$Y = \alpha + \beta X$$

Where Y is dependent variable and X is independent variable and α, β are parameters.

II) Non-Linear Regression:

- If the curve of regression is not a straight line then it is called as non-linear regression.

Regressions tries to find the mathematical relationship between variables, if it gives a curved line then it is a non-linear regression.

It is also known as **Curvilinear Regression.**

iii) **Multiple Regression:**

Multiple Regression is given by following formula

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Multiple regression includes more than one predictor variable.

Q5] **A SIMPLE EXAMPLE FROM THE STOCK MARKET INVOLVING ONLY DISCRETE RANGES HAS PROFIT AS CATEGORICAL ATTRIBUTE, WITH VALUES {UP, DOWN} AND THE TRAINING DATA SET IS GIVEN BELOW.**

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Apply decision tree algorithm and show the generated rules.

ANS: **[10M – MAY16]**

DECISION TREE BASED CLASSIFICATION:

Refer Q1.

***** Note: Even if the sum is asked, write a short theory explaining about the content. *****

DECISION TREE FOR ABOVE EXAMPLE:

Figure 8.2 shows the decision tree for stock market case.

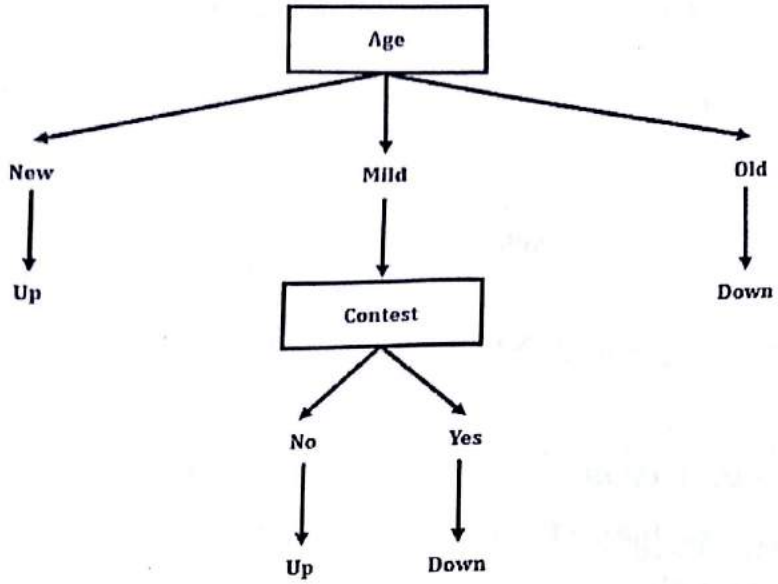


Figure 8.2: Decision tree for stock market case.

RULES:

1. IF Age = New THEN Profit = Up.
2. IF Age = Mild and Contest = No THEN Profit = Up.
3. IF Age = Mild and Contest = Yes THEN Profit = Down.
4. IF Age = Down THEN Profit = Down.

*** EXTRA QUESTIONS ***

Q1] WHAT IS CLASSIFICATION? WHAT ARE THE ISSUES IN CLASSIFICATION?

ANS:

CLASSIFICATION:

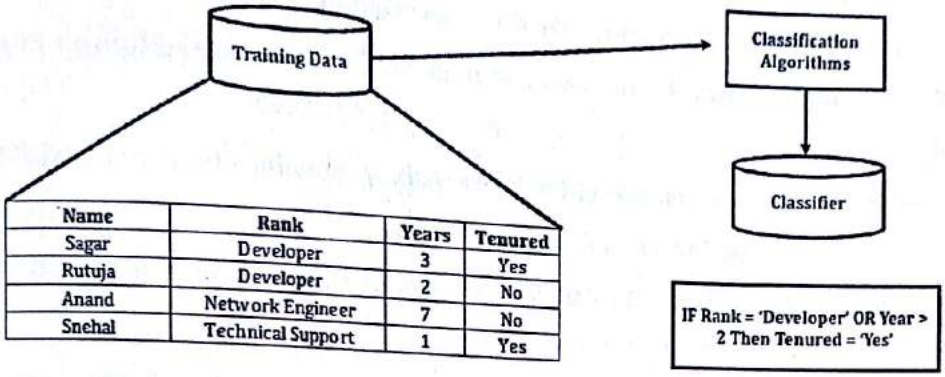
1. Classification is the form of **data analysis.**
2. Classification constructs classification model based on **training data set.**
3. Using this model it classifies the new data.
4. Classification models **predict categorical class labels.**
5. For example, we can build a classification model to categorize bank loan applications as eith safe or risky.

CLASSIFICATION PROCESS:

Classification is a two-step process:

I) **Model Constructions:**

- This step is the learning step.
- In this step the classification algorithms build the **classifier.**
- The classifier is built from the **training set** made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class.
- Figure 8.3 shows example of model construction.

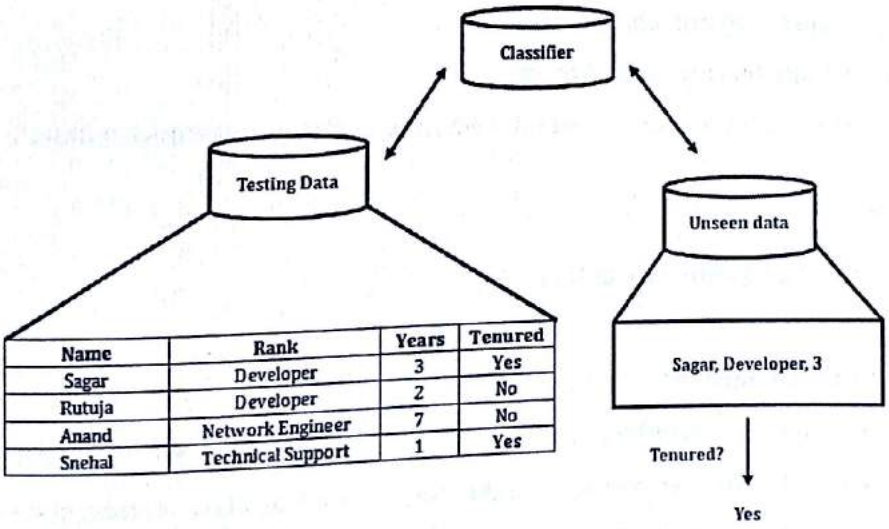


Name	Rank	Years	Tenured
Sagar	Developer	3	Yes
Rutuja	Developer	2	No
Anand	Network Engineer	7	No
Snehal	Technical Support	1	Yes

Figure 8.3: Example of model construction.

II) **Model Usage:**

- In this step, the classifier is used for classification.
- Here the test data is used to estimate the accuracy of classification rules.
- The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.
- Figure 8.4 shows example of model usage.



Name	Rank	Years	Tenured
Sagar	Developer	3	Yes
Rutuja	Developer	2	No
Anand	Network Engineer	7	No
Snehal	Technical Support	1	Yes

Figure 8.4: Example of model usage.

ISSUES IN CLASSIFICATION:

The major issue is preparing the data for classification. Preparing the data involves the following activities:

1. **Data Cleaning:** Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.
2. **Relevance Analysis:** Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
3. **Data Transformation and reduction:** The data can be transformed by any of the following methods.
    a. **Normalization:** Normalization involves scaling all values for given attribute in order to make them fall within a small specified range.
    b. **Generalization:** The data can also be transformed by generalizing it to the higher concept.

Q2] EXPLAIN ID3 ALGORITHM OR CLASSIFICATION ALGORITHM.

ANS:

ID3:

1. ID3 Stands for **Iterative Dichotomiser 3.**
2. It is an algorithm to build decision tree.
3. It was developed by J. Ross Quinlan in 1980.
4. ID3 adopt a greedy approach.
5. In this algorithm, there is no backtracking.
6. The trees are constructed in a top-down recursive divide-and-conquer manner.

ID3 ALGORITHM:

IDA (Examples, Target_attribute, Attributes)
Begin

- Create a Root node for the tree.
- If all Examples are positive, Return the single-node tree Root, with label = +
- If all Examples are negative, Return the single-node tree Root, with label = -
- If Attribute list is empty, Return the single-node tree Root, with label = most common value of Target_attribute in Examples.

- Otherwise Begin
    - A ← the Attribute that best classifies Examples.
    - The decision attribute for Root ← A
    - For each possible value V_i, of A.
    - Add a new tree branch below Root, corresponding to the test A = V_i
    - Let Example V_i be the subset of Examples that have value V_i for A.
    - If Example V_i is empty
        - Then below this new branch add a leaf node with label = most common value of Target attribute in Examples.
        - Else below this new branch add the sub-tree ID3 (Example V_i, Target_attributes, Attributes {A})
- End.
- Return Root.

Advantages:

- ID3 builds a short tree.
- ID3 builds a fastest tree.

Disadvantages:

- It perform poorly with many class and small data.
- Computationally expensive to train.

CHAPTER - 9: CLUSTERING

Q1] EXPLAIN K-MEANS CLUSTERING ALGORITHM? APPLY K-MEANS ALGORITHMS FOR THE FOLLOWING DATA SET WITH TWO CLUSTERS. DATA SET = {1, 2, 6, 7, 8, 10, 15, 17, 20} [10M – MAY16]

ANS:

CLUSTERING:

1. Clustering is **unsupervised learning problem.**
2. It is data mining technique used to place data elements into related groups without advanc knowledge of the group definitions.
3. It is a process of portioning data objects into sub classes which are called as clusters.
4. Clustering Algorithms are used in Marketing, Biology, and Insurance etc.

K-MEANS CLUSTERING ALGORITHM:

1. K-Means Clustering is one of the **partitioning method.**
2. It is simplest **unsupervised learning algorithm.**
3. K-Means Clustering aims to partition 'n' observations into 'k' clusters in which each observatio belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
4. This results in a partitioning of the data space.
5. K is positive integer number.

K-Means Clustering Process:

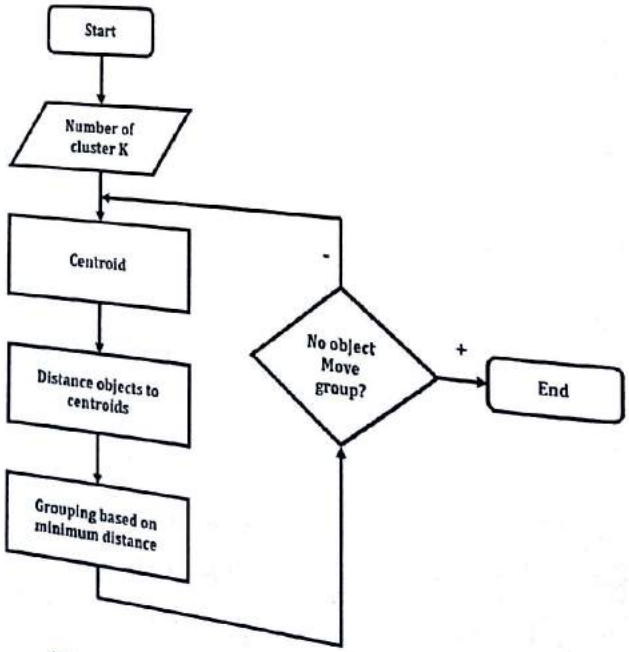


Figure 9.1: Flowchart for K-Means Clustering.

1. Figure 9.1 shows the flowchart for K-Means Clustering.
2. Define K centroids for K clusters which are generally far away from each other.
3. Then group the elements into clusters which are nearer to the centroid of that cluster.
4. After this first step, again calculate the new centroid for each cluster based on the elements of that cluster.
5. Follow the same method, and group the elements based on new centroid.
6. In every step, the centroid changes and elements move from one cluster to another
7. Do the same process till no element is moving from one cluster to another.

EXAMPLE:

Given:

Data Set = {1, 2, 6, 7, 8, 10, 15, 17, 20}

No. of clusters = 2

Solution:

Step-1: (Define K Centroid)

Consider initial two centroids for two clusters $C_1 = 6$ and $C_2 = 15$

Step-2: (Randomly assign data to two clusters)

$K_1 = \{1, 2, 6, 7, 8, 10\}$
$K_2 = \{15, 17, 20\}$

Step-3: (Calculate Mean)

No. of clusters = 2

Therefore $K_1 = \{1, 2, 6, 7, 8, 10\}$ $C_1 = \text{Mean} = 34/6 = 5.67$

$K_2 = \{15, 17, 20\}$ $C_2 = \text{Mean} = 52/3 = 17.33$

Step-4: (Reassign)

$K_1 = \{1, 2, 6, 7, 8, 10\}$
$K_2 = \{15, 17, 20\}$

As no elements is moving from cluster, so the final answer is **$K_1 = \{1, 2, 6, 7, 8, 10\}$ and $K_2 = \{15, 17, 20\}$**

Q2] WHAT IS CLUSTERING TECHNIQUES? DISCUSS THE AGGLOMERATIVE ALGORITHM WITH THE FOLLOWING DATA AND PLOT A DENDROGRAM USING SINGLE LINK APPROACH. THE TABLE BELOW COMPRISES SAMPLE DATA ITEMS INDICTING THE DISTANCE BETWEEN THE ELEMENTS.

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

[10M – DEC16]

ANS:

CLUSTERING:

1. Clustering is **unsupervised learning problem.**
2. It is data mining technique used to place data elements into related groups without advance knowledge of the group definitions.
3. It is a process of portioning data objects into sub classes which are called as clusters.
4. Clustering Algorithms are used in Marketing, Biology, and Insurance etc.

CLUSTERING TECHNIQUES:

Clustering Techniques can be classified into the following categories:

I) Partitioning Method:

- In Partitioning based approach, various partition is created.
- Each partition represents a **cluster.**

II) Hierarchical Method:

- This method creates a **hierarchical decomposition** of the given dataset of objects.
- There are two approaches – Agglomerative approach and Divisive approach.

III) Density - Based Method:

- This method is based on the **notion of density.**
- The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold.

IV) Grid - Based Method:

- In this, the various objects together **form a grid.**
- The object space is quantized into finite number of cells that form a grid structure.

VI) **Model-Based Method:**

- In this method, a model is hypothesized for each cluster to find the best fit of data for a given model.
- This method uses **density function** to locate clusters.

VII) **Constraint-based Method:**

- In this method, the clustering is performed by the incorporation of **constraints**.
- Constraints can be user-oriented or application-oriented.

AGGLOMERATIVE ALGORITHM:

1. Agglomerative Algorithm is used in **Hierarchical based clustering**.
2. It is also known as **AGNES (agglomerative nesting)**.
3. This approach is also known as the **bottom-up approach**.
4. In this, we start with each object forming a separate group.
5. It keeps on merging the objects or groups that are close to one another.
6. It keep on doing so until all of the groups are merged into one or until the termination condition holds.
7. A Hierarchical Agglomerative Clustering is typically visualized as a Dendrogram as shown in figure 9.2.
8. Dendrogram is tree like structure used to illustrate hierarchical clustering technique.

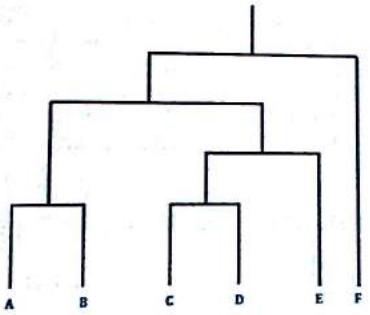


Figure 9.2: Dendrogram.

EXAMPLE:

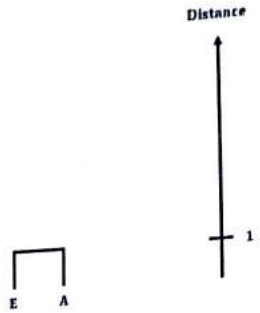
Given:

Distance Matrix:

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

Step - 1:

From above given distance matrix, E and A clusters has minimum distance i.e. 1, so merge them together to form cluster (E, A)



Distance Matrix:

Dist ((E A), C) = MIN (Dist (E, C), Dist (A, C))
= MIN (2, 2) = 2

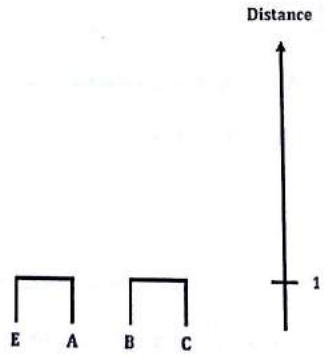
Dist ((E, A), B) = MIN (Dist (E, B), Dist (A, B))
= MIN (2, 5) = 2

Dist ((E, A), D) = MIN (Dist (E, D), Dist (A, D))
= MIN (3, 3) = 3

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

Step - 2:

Consider the distance matrix obtained in step 1. Since B, C distance is minimum, we combine B and C.



Distance Matrix:

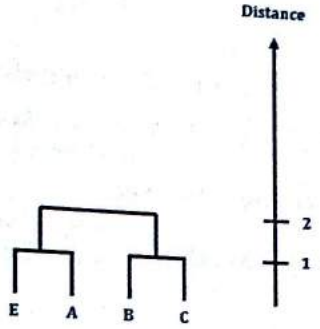
Dist ((B, C), (E, A)) = MIN (Dist (B, E), Dist (B, A), Dist (C, E), Dist (C, A))
= MIN (2, 5, 2, 2) = 2

Dist ((B, C), D) = MIN (Dist (B, D), Dist (C, D))
= MIN (3, 6) = 3

Item	E, A	B, C	D
E, A	0		
B, C	2	0	
D	3	3	0

Step - 3:

Consider the distance matrix obtained in step 2. Since (E, A) and (B, C) distance is minimum, we combine them.



Distance **Matrix:**

Dist ((E, A), (B, C)) = MIN (Dist (E, B), Dist (E, C), Dist (A, B), Dist (A, C)
= MIN (2, 2, 5, 2) = 2

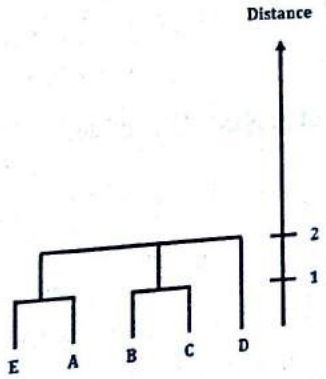
Dist ((B, C), D) = MIN (Dist (B, D), Dist (C, D))
= MIN (3, 6) = 3

Item	E, A, B, C	D
E, A, B, C	0	
D	2	0

Step – 4:

Finally we combine D with (E, A, B, C)

Final **Dendrogram:**



Q1] GIVE FIVE EXAMPLES OF APPLICATIONS THAT CAN BE USE CLUSTERING. DESCRIBE ONE CLUSTERING ALGORITHM WITH THE HELP OF EXAMPLE.

ANS:

CLUSTERING:

Refer Q1 from University Questions (Clustering Part)

APPLICATIONS OF CLUSTERING:

I) **Marketing:** Clustering is used in many marketing applications such as market research, pat recognition, data analysis, and image processing.

II) **Biology:** Clustering can also be used in classifying plants and animals into different classes ba on their features.

III) **Libraries:** Clustering can be used for book ordering in libraries.

IV) **Insurance:** Using clustering, different groups of policy holders can be identified.

V) **City Planning:** Details like geographical locations, house type and groups of house can identified using clustering.

VI) **Astronomy:** Clustering helps to find groups of similar stars and galaxies.

VII) **WWW:** It can be used to find groups of similar access patterns using weblog data.

VIII) **Earthquake Studies:** Clustering is used to identify dangerous zones based on earthqua epicenter.

CLUSTERING ALGORITHM:

Refer Q1 from University Questions (K-Means Clustering Part)

Q2] EXPLAIN HIERARCHICAL CLUSTERING METHODS.

ANS:

HIERARCHICAL METHOD:

1. This method creates a hierarchical decomposition of the given dataset of objects.
2. There are two approaches – Agglomerative approach and Divisive approach.

I) **Agglomerative Hierarchical Clustering:**

Refer Q2 Agglomerative Algorithm Part.

II) **Divisive Hierarchical Clustering:**

- It is just the reverse of Agglomerative Hierarchical approach.
- This approach is also known as the **top-down approach.**
- In this, we start with all of the objects in the same cluster.
- In the continuous iteration, a cluster is split up into smaller clusters.
- Process is repeated until each object in one cluster is split up or the termination condition holds.
- This method is **rigid**, i.e., once a merging or splitting is done, it can never be undone.

AGGLOMERATIVE V/S DIVISIVE HIERARCHICAL CLUSTERING:

Agglomerative and Divisive Hierarchical Clustering on data objects {1, 2, 3, 4, 5} is shown in figure 9.3.

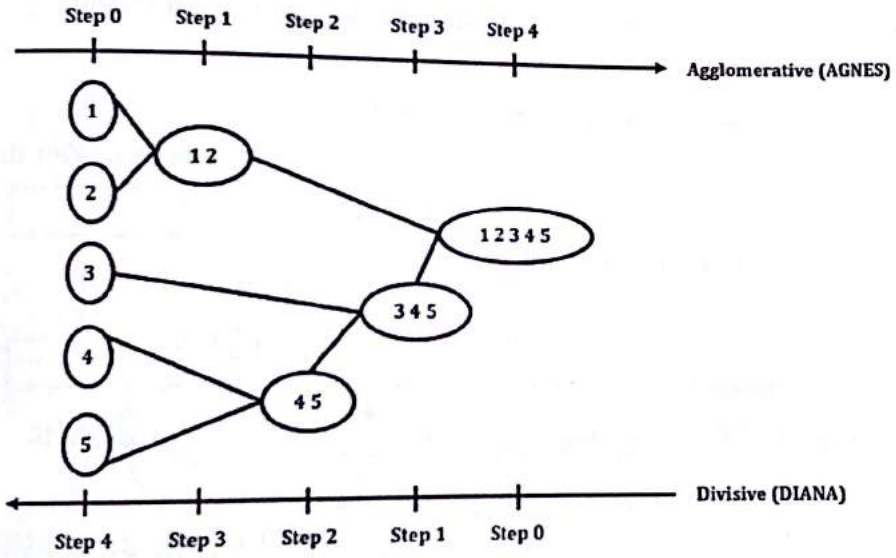


Figure 9.3: Agglomerative v/s Divisive Hierarchical Clustering.

CHAPTER - 10: MINING FREQUENT PATTERN AND ASSOCIATION RULE

[5M – MAY16]

Q1] FP TREE.

ANS:

FP TREE:

1. FP Tree Stands for **Frequent Pattern Tree.**
2. An FP Tree is a tree structure which consists of one root labelled as "null" and Set of item-prefix sub trees.
3. Each node in the item-prefix sub tree consists of three fields:
   a. Item-name.
   b. Count.
   c. Node-link.
4. FP Tree is a compact structure that stores **quantitative information** about frequent patterns in a database.
5. The size of FP Tree is bounded by size of database.
6. But due to frequent pattern sharing, the size of the tree is usually much smaller than its original database.
7. Figure 10.1 shows the example of an FP Tree.

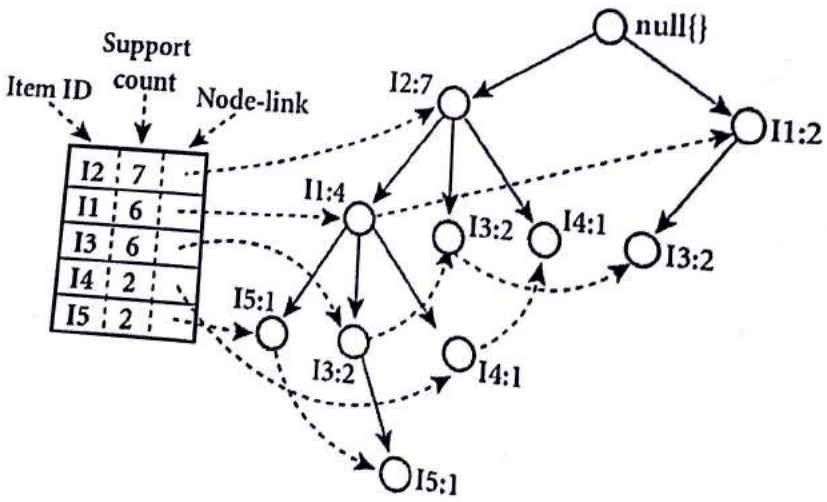


Figure 10.1: Example of an FP Tree.

Advantages:

- FP Tree is much faster than Apriori.
- It provides compresses data set.

Disadvantages:

- FP Tree may not fit in memory.
- It is expensive to build.

Q2] MULTILEVEL & MULTIDIMENSIONAL ASSOCIATION RULE.

ANS:

[5M – MAY16]

MULTILEVEL ASSOCIATION RULE:

- Rules which combine association with hierarchy of concepts are called as **Multilevel Association Rules.**
- In multilevel association rule, items are always in the form of **hierarchy.**
- Items which are placed at leaf nodes has lower support.
- An item can be generalized or specialized as per the described hierarchy of that item.
- Figure 10.2 shows the example of multilevel association rule.

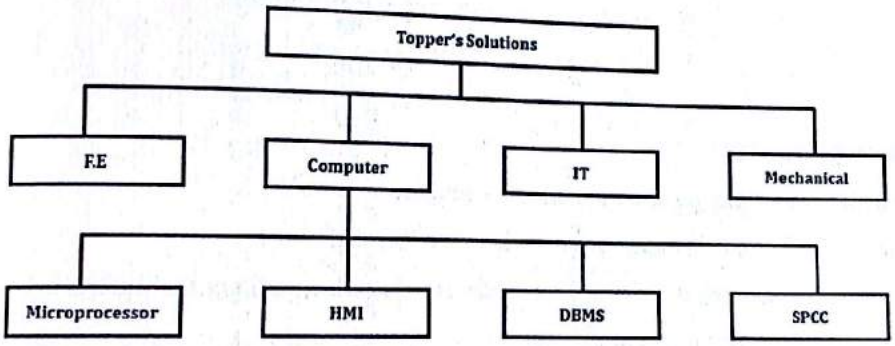


Figure 10.2: Example of multilevel association rule.

MULTIDIMENSIONAL ASSOCIATION RULE:

- Rules which combine association with multiple dimensions are called as **Multidimensional Association Rules.**
- In this, Rule contains two or more dimensions or predicates.
- There are two types; Inter dimension association rules and hybrid dimension association rules.
  - **Inter dimension association rules:** This rule does not have any repeated predicate. For Example:

    Gender (X, "Male") ^ Salary (X, "High") → Buys (X, "Computer")
  - **Hybrid dimension association rules:** This rule have many occurrences of same predicate i.e. buys.

    Gender (X, "Male") ^ Buys (X, "TV") → Buys (X, "DVD")

Q3] DISCUSS ASSOCIATION RULE MINING AND APRIORI ALGORITHM. APPLY AR MINING TO FIND ALL FREQUENT ITEM SETS AND ASSOCIATION RULES FOR THE FOLLOWING DATASET:

Minimum Support Count = 2

Minimum Confidence = 70%

Transaction_ID	Items
100	1, 2, 5
200	2, 4
300	2, 3
400	1, 2, 4
500	1, 3
600	1, 3
700	1, 3, 2, 5
800	1, 3
900	1, 2, 3

ANS:

[10M – MAY16]

ASSOCIATION RULE MINING:

1. Association rule mining is the **data mining process.**
2. It is used to find the **rules that govern the associations.**
3. Association rule mining is a procedure which is meant to find frequent patterns, correlations and associations in various kinds of databases.
4. Databases can be relational databases, transactional databases, and other forms of data repositories.
5. Association Rule Mining are of two types; Multilevel association rule and Multidimensional association rule.

I) **Multilevel Association Rule:**

Refer Q2.

II) **Multidimensional Association Rule:**

Refer Q2.

APRIORI ALGORITHM:

1. Apriori Algorithm is one of **frequent Itemset mining method.**
2. It is used to solve the frequent item set problem.
3. Apriori Algorithm uses a **"Bottom Up" Approach.**

- Apriori Algorithm analyzes a data set to determine which combinations of items can occur together frequently.

Advantages:

- Easy to implement.
- Apriori Algorithm can be easily parallelized.

Disadvantages:

- Performance is low.
- It requires many database scans.

EXAMPLE:

Given:

Minimum Support Count = 2

Minimum Confidence = 70%

Transaction_ID	Items
100	1, 2, 5
200	2, 4
300	2, 3
400	1, 2, 4
500	1, 3
600	1, 3
700	1, 3, 2, 5
800	1, 3
900	1, 2, 3

Solution:

Step-1:

Scan the transaction database and find the count of Items.

Candidate List = {1, 2, 3, 4, 5}

C_1 =

Itemset	Support Count
1	7
2	6
3	6
4	2
5	2

Step-2:

Check whether each candidate item is present in at least two transactions because the given support count is 2.

L_1 =

Itemset	Support Count
1	7
2	6
3	6
4	2
5	2

Step-3:

Now generate Candidate C_2 from L_1 and find the support count for items.

C_2 =

Itemset	Support Count
1,2	4
1,3	5
1,4	1
1,5	2
2,3	3
2,4	2
2,5	2
3,4	0
3,5	1
4,5	0

Step-4:

Now we compare Candidate C_2 generated in step 3 with the minimum support count and prune those Itemsets which do not satisfy the minimum support count.

L_2 =

Itemset	Support Count
1,2	4
1,3	5
1,5	2
2,3	3
2,4	2
2,5	2

Step-5

Now generate Candidate C_3 from L_2 and find the support count for items.

C_3 =

Itemset	Support Count
1,2,3	2
1,2,5	2
1,3,5	1
2,3,4	0
2,3,5	1
2,4,5	0

Step-6

Now we compare

Itemsets which do

Step-7:

Frequent Itemse

Let consider the

Following are th

Association Ru
1 ^ 2 => 5
1 ^ 5 => 2
2 ^ 5 => 1
1 => 2 ^ 5
2 => 1 ^ 5
5 => 1 ^ 2

Minimum Co

strong rules.

Q4] A D

80%

TRA

ANS:

Step-6:

Now we compare Candidate C_3 generated in step 5 with the minimum support count and prune those itemsets which do not satisfy the minimum support count.

L_3 =

Itemset	Support Count
1, 2, 3	2
1, 2, 5	2

Step-7:

Frequent Itemset are {1, 2, 3} and {1, 2, 5}

Let consider the frequent Itemset {1, 2, 5}

Following are the association rules that can be generated shown below with the support and confidence.

Association Rule	Support	Confidence	Confidence %
1 ^ 2 => 5	2	2/4	50
1 ^ 5 => 2	2	2/2	100
2 ^ 5 => 1	2	2/2	100
1 => 2 ^ 5	2	2/7	29
2 => 1 ^ 5	2	2/6	33
5 => 1 ^ 2	2	2/2	100

Minimum Confidence threshold is 70 %. So the following rules are considered as output, as they are strong rules.

Rules	Confidence
1 ^ 5 => 2	100 %
2 ^ 5 => 1	100 %
5 => 1 ^ 2	100 %

Q4] A DATABASE HAS FIVE TRANSACTIONS. LET MIN-SUPPORT = 60% AND MIN-CONFIDENCE = 80%. FIND ALL FIND FREQUENT ITEM SETS BY USING APRIORI ALGORITHM. T_ID IS THE TRANSACTION ID

T_ID	Items Bought
T-1000	M, O, N, K, E, Y
T-1001	D, O, N, K, E, Y
T-1002	M, A, K, E
T-1003	M, U, C, K, Y
T-1004	C, O, O, K, E

[10M – DEC16]

ANS:

Given:

Minimum Support = 60 %

Minimum Confidence = 80%

T_ID	Items Bought
T-1000	M, O, N, K, E, Y
T-1001	D, O, N, K, E, Y
T-1002	M, A, K, E
T-1003	M, U, C, K, Y
T-1004	C, O, O, K, E

Solution:

Step-1:

Scan the transaction database and find the count of Items.

Candidate List = {A, C, D, E, K, M, N, O, U, Y}

C_1 =

Itemset	Support Count
A	1
C	2
D	1
E	4
K	5
M	3
N	2
O	4
U	1
Y	3

Step-2:

Now compare candidate support count with minimum support count (i.e. 60%)

L_1 =

Itemset	Support Count
E	4
K	5
M	3
O	4
Y	3

Step-3:

Now generate Candidate C_2 from L_1 and find the support count for items.

C_2 =

Itemset	Support Count
E, K	4
E, M	2
E, O	3
E, Y	2
K, M	3
K, O	3
K, Y	3
M, O	1
M, Y	2
O, Y	2

Step-4:

Now we compare Candidate C_2 generated in step 3 with the minimum support count and prune those itemsets which do not satisfy the minimum support count (i.e. 60 %).

L_2 =

Itemset	Support Count
E, K	4
E, O	3
K, M	3
K, O	3
K, Y	3

Step-5

Now generate Candidate C_3 from L_2 and find the support count for items.

C_3 =

Itemset	Support Count
E, K, M	2
E, K, O	3
E, K, Y	2
E, O, Y	2
K, M, O	1
K, M, Y	1

Step-6:

Now we compare Candidate C_3 generated in step 5 with the minimum support count and prune those Itemsets which do not satisfy the minimum support count (i.e. 60 %).

L_3 =

Itemset	Support Count
E, K, M	2
E, K, O	3
E, K, Y	2
E, O, Y	2

Step-7:

Frequent Itemset are {E, K, M}, {E, K, O}, {E, K, Y} and {E, O, Y}

Let consider the frequent Itemset {E, K, O}

Following are the association rules that can be generated shown below with the support and confidence.

Association Rule	Support	Confidence	Confidence %
E ^ K => O	3	3/4	75
E ^ O => K	3	3/3	100
K ^ O => E	3	3/3	100
E => K ^ O	3	3/4	75
K => E ^ O	3	3/5	60
O => E ^ K	3	3/4	75

Minimum Confidence threshold is 80 %. So the following rules are considered as output, as they are strong rules.

Rules	Confidence
E ^ O => K	100 %
K ^ O => E	100 %

MISCELLANEOUS

Q1] PROVIDE A SYSTEMATIC DESIGN ANALYSIS FOR MUNICIPAL CORPORATION'S MOBILE APP; THAT PROVIDES INFORMATION ABOUT THE WARDS, THEIR WARD OFFICE, CORPORATES IN THE WARD, SCHOOLS HOSPITALS IN THE WARD AND OTHER INFORMATION OF THE MUNICIPAL OFFICE, YOUR ANALYSIS SHOULD CONSIST OF ALL NECESSARY INTERFACE GUIDELINES.

ANS:

DATA CUBE COMPUTATION: [10M – DEC16]

1. Data cube computation is an essential task in data warehouse implementation.
2. The pre-computation of all or part of a data cube can greatly reduce the response time and enhance the performance of OLAP.
3. However, such computation is challenging because it may require substantial computational time and storage space.

DATA CUBE COMPUTATION METHODS:

I) **Multi-way Array Aggregation:**

- The Multi-way Array Aggregation method computes a full data cube by using a **multidimensional array.**
- It is array based **bottom up algorithm.**
- It is a typical MOLAP approach that uses direct array addressing.
- It uses multi-dimensional chunks.
- Figure 1 shows Multi-way Array Aggregation exploration for a 3-D data cube computation.

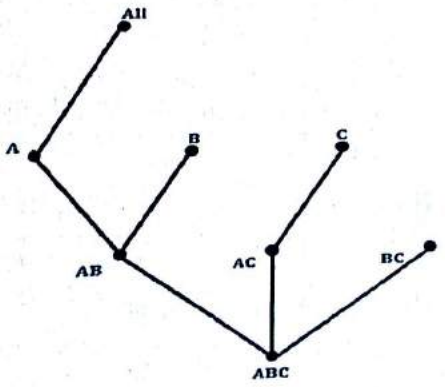


Figure 1: Multi-way Array Aggregation exploration for a 3-D data cube computation.

Limitations:

- It can compute well only for a small number of dimensions.

Q1]

II) **BUC:**

- BUC Stands for **Bottom Up Cube Computation.**
- BUC is an algorithm for the **computation of sparse and iceberg cubes.**
- BUC divides dimensions into partitions and facilitates iceberg pruning.
  - If a partition does not satisfy min_sup, its descendants can be pruned.
  - If min_sup = 1 → compute full CUBE.
- In BUC, No simultaneous aggregation is allowed.
- Figure 2 shows BUC exploration for a 3-D data cube computation.

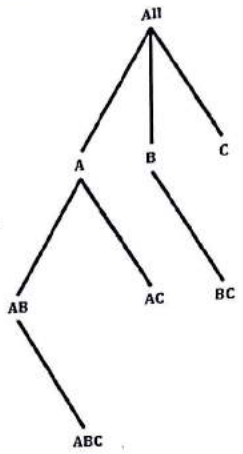


Figure 2: BUC exploration for a 3-D data cube computation.

Q2]

III) **Star Cubing:**

- Star-Cubing combines the **strengths of the Multi-way array aggregation and BUC.**
- It integrates top-down and bottom-up cube computation.
- It explores both multidimensional aggregation (similar to Multi-Way) and Apriori-like pruning (similar to BUC).
- It operates from a data structure called a **star-tree.**

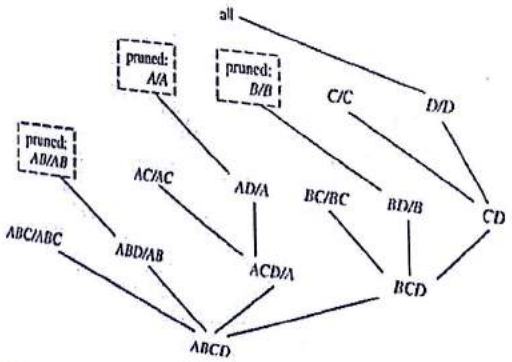


Figure 3: Star-Cubing bottom-up computation with top-down expansion of shared dimensions.

Advantage:

- Reduce the computation time and memory requirements.

Data Warehouse & Mining – May 2016

(a) For a Super market chain, consider the following dimensions namely product, store, time & promotion. The schema contains a central fact table for sales.

i. Design star schema for the above application.

ii. Calculate the maximum number of base fact tables records for warehouse with the following values given below: [10]

- Time period – 5 Years.
- Store – 300 stores reporting daily sales.
- Product – 40,000 products in each store (about 4000 sell in each store daily)

Ans: [Chapter - 2]

(b) Discuss:

i. The steps in KDD Process.

ii. The architecture of a typical DM System. [10]

Ans: [Chapter - 5]

Q2] (a) We would like to view sales data of a company with respect to three dimensions namely Location, Item and Time. Represent the sales data in the form of a 3-D data cube for the above and perform Roll up, Drill down, Slice and Dice OLAP operations on the above data cube and Illustrate. [10]

Ans: [Chapter - 4]

(b) A simple example from the stock market involving only discrete ranges has profit as categorical attribute, with values {Up, Down} and the training data set is given below. [10]

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Apply decision tree algorithm and show the generated rules.

Ans: [Chapter - 8]

Q3] (a) Illustrate the architecture of a typical DW system. Differentiate DW and Data Mart. [10]

Ans: [Chapter - 1]

(b) Discuss different steps involved in Data Preprocessing. [10]

Ans: [Chapter - 7]

Q4] (a) Discuss various OLAP Models. [10]

Ans: [Chapter - 4]

(b) Explain K-Means clustering algorithm? Apply K-Means Algorithms for the following data set with two clusters. Data Set = {1, 2, 6, 7, 8, 10, 15, 17, 20} [10]

Ans: [Chapter - 9]

Q5] (a) Describe the steps of ETL Process. [10]

Ans: [Chapter - 3]

(b) Discuss Association Rule Mining and Apriori Algorithm. Apply AR Mining to find all frequent item sets and association rules for the following dataset: [10]

Minimum Support Count = 2

Minimum Confidence = 70%

Transaction_ID	Items
100	1, 2, 5
200	2, 4
300	2, 3
400	1, 2, 4
500	1, 3
600	1, 3
700	1, 3, 2, 5
800	1, 3
900	1, 2, 3

Ans: [Chapter - 10]

Q6] Write short notes on any four of the following: [20]

(a) Updates to Dimension tables.

Ans: [Chapter - 2]

(b) Metrics for Evaluating Classifier Performance.

Ans: [Chapter - 8]

(c) FP Tree.

Ans: [Chapter - 10]

(d) Multilevel & Multidimensional Association Rule.

Ans: [Chapter - 10]

(e) Operational Vs. Decisional Support System.

Ans: [Chapter - 1]

Data Warehouse & Mining – Dec 2016

Q1] (a) Consider following dimensions for a Hypermarket chain: Product, Store, Time and Promotion. With respect to this business scenario, answer the following questions. Clearly state any reasonable assumptions you make. Design a star schema. Whether the star schema can be converted to snowflake schema? Justify your answer and draw snowflake schema for the data warehouse (clearly mention the Fact table(s), Dimension table(s), their attributes and measures) [10]

Ans: [Chapter - 2]

(b) Define linear, non-linear and multiple regressions. Plan a regression model for Disease development with respect to change in weather parameters. [10]

Ans: [Chapter - 8]

Q2] (a) What is meant by metadata in the context of a Data warehouse? Explain the different types of Meta data stored in a data warehouse. Illustrate with a suitable example. [10]

Ans: [Chapter - 1]

(b) Describe the various functionalities of Data Mining as a step in the process of knowledge discovery. [10]

Ans: [Chapter - 5]

Q3] (a) In what way ETL cycle can be used in typical data warehouse, explain with suitable instance. [10]

Ans: [Chapter - 3]

(b) What is Clustering Techniques? Discuss the Agglomerative algorithm with the following data and plot a Dendrogram using single link approach. The table below comprises 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

Ans: [Chapter - 9]

Q4] (a) Discuss how computations can be performed efficiently on data cubes. [10]

Ans: [Chapter - Miscellaneous]

(b) A database has five transactions. Let min-support = 60% and min-confidence = 80%. Find all find frequent item sets by using Apriori Algorithm. T_ID is the transaction ID [10]

T_ID	Items Bought
T-1000	M, O, N, K, E, Y

T-1001	D, O, N, K, E, Y
T-1002	M, A, K, E
T-1003	M, U, C, K, Y
T-1004	C, O, O, K, E

Ans: [Chapter - 10]

Q5] (a) Differentiate **[10]**

i. OLTP Vs. OLAP.

ii. Data Warehouse Vs. Data Mart.

Ans: [Chapter - 4]

(b) Why Naïve Bayesian Classification is called "naive"? Briefly outline the major ideas of native Bayesian Classification **[10]**

Ans: [Chapter - 8]

Q6] Write short notes on any four of the following: **[20]**

(a) Application of Data Mining to Financial Analysis.

Ans: [Chapter - 5]

(b) Fact Less Fact Table.

Ans: [Chapter - 2]

(c) Indexing OLAP Data.

Ans: [Chapter - 4]

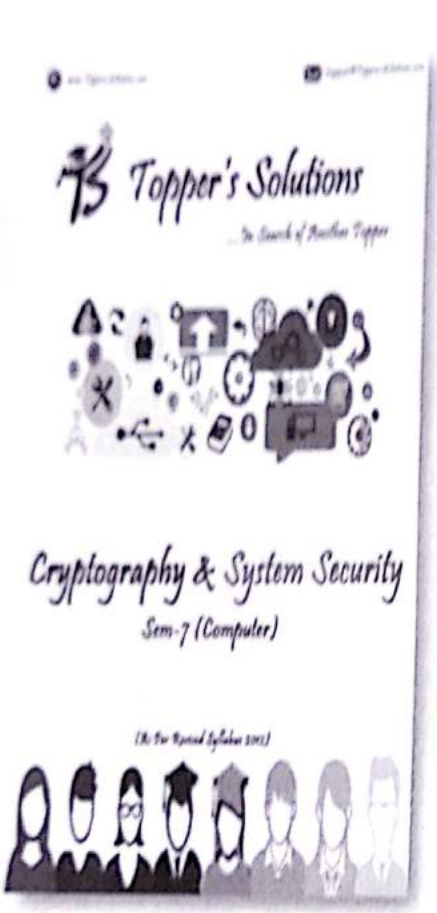
(d) Data Quality.

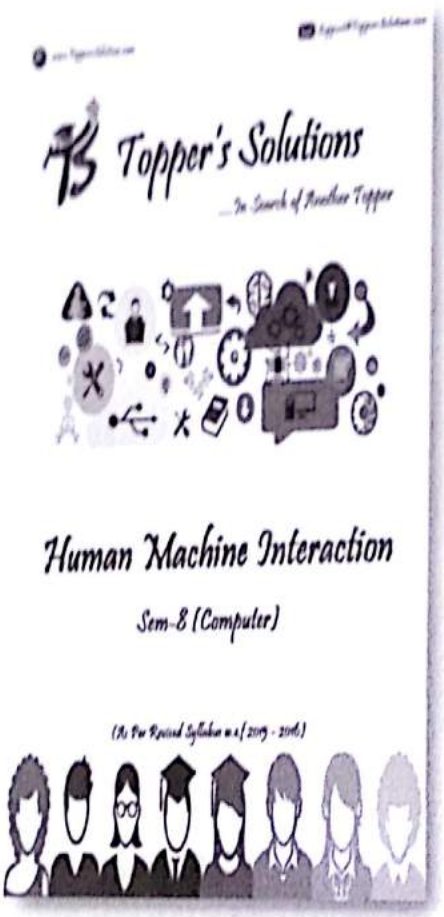
Ans: [Chapter - 3]

(e) Decision Tree based Classification Approach.

Ans: [Chapter - 8]

Other Subjects





** Final Year Projects are also Available @ Topper's Solutions **

We hope you like Topper's Solutions & find them useful. For any suggestions, comments or feedback do contact us at Support@TopperSolutions.com. We are happy to hear you about Topper's Solutions.

Wishing you Best Luck,
Topper's Solutions Team.

Price: Rs. 50

+917507331198

Support@TopperSolutions.com