CBCS SCHEME

USN 1 C R 2 3 M C 1 0 7

22MCA252

Second Semester MCA Degree Examination, June/July 2024

Data Mining and Business Intelligence

Time: 3 hrs.

Max. Marks: 100

Note: 1. Answer any FIVE full questions, choosing ONE full question from each module.
2. M : Marks , L: Bloom's level , C: Course outcomes.

		Module – 1	M	L	C
Q.1	a.	Define Data Warehouse. Explain Data Life Cycle Stages in detail.	10	L2	CO1
	b.	Differentiate between Data warehouse and data marts.	10	L2	CO2
		OR			
Q.2	a.	Explain components of Metadata with a neat diagram.	10	L2	CO2
	b.	Describe star and snowflake schema with a neat diagram.	10	L2	CO2
		Module – 2			
Q.3	a.	What is data mining? Explain knowledge Data Discovery process in detail.	10	L2	CO2
	b.	Explain Data Mining task primitives in detail.	10	L2	CO3
		OR			
Q.4	a.	Explain Data Cleaning Process in detail.	10	L2	CO2
	b.	Describe any two data compression methods in detail.	10	L2	CO2
		Module – 3			
Q.5	a.	Define concept description and explain briefly data generalization by Attribute oriented Induction.	10	L2	CO2
	b.	Describe how class comparisons are performed with respect to attribute relevance.	10	L2	CO1
		OR			
Q.6	a.	Explain Market basket analysis basic concepts in detail.	10	L2	CO2
	b.	Explain Apriori Algorithm in detail.	10	L2	CO1
		Module – 4			
Q.7	a.	Define Classification and prediction and enlist the issues regarding classification and prediction.	10	L2	CO2
	b.	Discuss briefly Decision tree and Bayesian classification.	10	L2	CO2
		OR			
Q.8	a.	Explain classification and Regression Tree [CART] in detail.	10	L2	CO2

	b.	Describe in detail Linear and Logistic Regression.	10	L2	CO4
Module – 5					
Q.9	a.	Explain Data Mining for Business Intelligence.	10	L2	CO3
	b.	Explain Big data Business Analytics in detail.	10	L2	CO5
OR					
Q.10		Write short notes on four of the following : i) WEKA TOOL ii) Drill down and roll-up operations iii) Incremental ARM iv) OLAP and OLTP [Online analytical Process and Online Transactional Processing]	20	L2	CO1

* * * * *

VTU Second semester MCA Degree Examination – June/July 2024

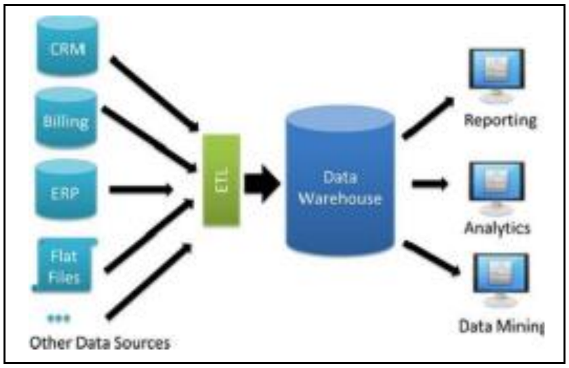
Data Mining and Business Intelligence							Sub Code:	22MCA252
12/11/2024	Duration:	3 hrs	Max Marks:	100	Sem:	II	Branch:	MCA

Note: 1. Answer any FIVE full questions, Choosing ONE full question from each module.

Module-1

Q1. a. Define Data Warehouse. Explain Data Life cycle Stages in detail

Data warehouse is like a relational database designed for analytical needs. It functions on the basis of OLAP (Online Analytical Processing). It is a central location where consolidated data from multiple locations (databases) are stored.



Data Life cycle Stages

1. **Data Cleaning**:
   - **Handling Missing Values**: This involves identifying and addressing missing data, either by removing incomplete records or imputing missing values using statistical methods.
   - **Smoothing Noisy Data**: Techniques such as binning, regression, or clustering are used to remove noise or outliers from the data.
   - **Correcting Inconsistencies**: Ensuring that data values are consistent across different datasets or within a single dataset.
2. **Data Integration**:
   - **Combining Data from Multiple Sources**: This involves merging data from different sources or databases to create a unified dataset. Techniques such as schema integration, entity resolution, and deduplication are used to ensure consistency and avoid redundancy.
3. **Data Transformation**:
   - **Normalization and Scaling**: Transforming data into a suitable format or range, such as scaling numerical attributes to a specific range or normalizing data to ensure consistency across features.

- **Data Discretization**: Converting continuous attributes into discrete intervals or categories.
- **Attribute Construction**: Creating new attributes or features by combining or transforming existing ones.

4. **Data Reduction**:
   - **Dimensionality Reduction**: Reducing the number of attributes in the dataset by using techniques like Principal Component Analysis (PCA) or feature selection methods.
   - **Data Compression**: Reducing the volume of data by encoding techniques or removing redundant information.
   - **Sampling**: Selecting a representative subset of the data to reduce the dataset size while maintaining its analytical value.
5. **Data Discretization and Concept Hierarchy Generation**:
   - **Discretization**: Transforming continuous data into categorical data by creating a set of ranges or intervals.
   - **Concept Hierarchy Generation**: Organizing data attributes into a hierarchy or levels of abstraction, which can simplify the analysis process.

Q1.b. Differentiate between Data Warehouse and Data marts

Sl.No	Data Warehouse	Data Mart
1.	Data warehouse is a Centralised system.	While it is a decentralised system.
2.	In data warehouse, lightly denormalization takes place.	While in Data mart, highly denormalization takes place.
3.	Data warehouse is top-down model.	While it is a bottom-up model.
4.	To built a warehouse is difficult.	While to build a mart is easy.
5.	In data warehouse, Fact constellation schema is used.	While in this, Star schema and snowflake schema are used.
6.	Data Warehouse is flexible.	While it is not flexible.
7.	Data Warehouse is the data-oriented in nature.	While it is the project-oriented in nature.
8.	Data Ware house has long life.	While data-mart has short life than warehouse.
9.	In Data Warehouse, Data are contained in detail form.	While in this, data are contained in summarized form.
10.	Data Warehouse is vast in size.	While data mart is smaller than warehouse.
11.	The Data Warehouse might be somewhere between 100 GB and 1 TB+ in size.	The Size of Data Mart is less than 100 GB.
12.	The time it takes to implement a data warehouse might range from months to years.	The Data Mart deployment procedure is time-limited to a few months.
13.	It uses a lot of data and has comprehensive	Operational data are not present in Data

	operational data.	Mart.
14.	It collects data from various data sources.	It generally stores data from a data warehouse.
15.	Long time for processing the data because of large data.	Less time for processing the data because of handling only a small amount of data.
16.	Complicated design process of creating schemas and views.	Easy design process of creating schemas and views.

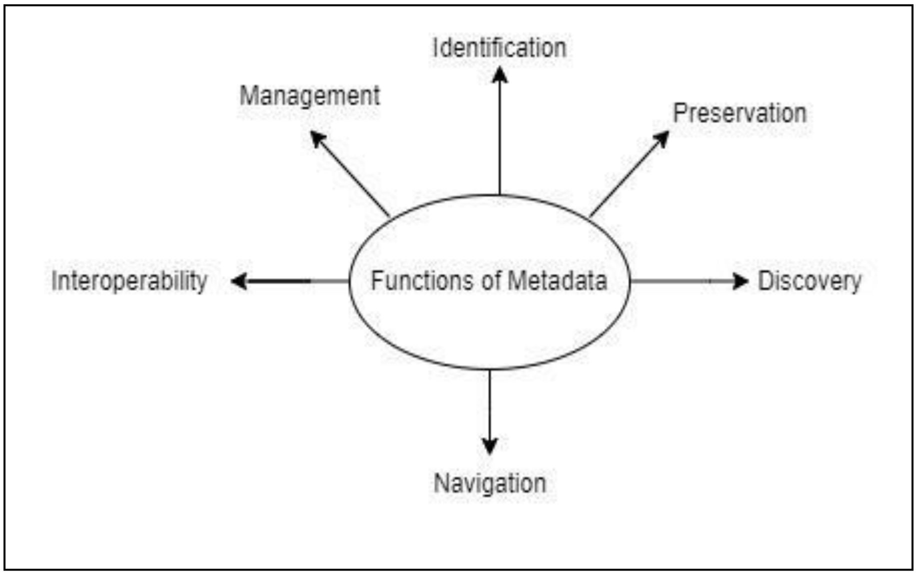
Q2. a. Explain components of Metadata with a neat diagram

Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata.
Metadata can be broadly categorized into three categories −

Business Metadata − It has the data ownership information, business definition, and changing policies.

Technical Metadata − It includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.

Operational Metadata − It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.

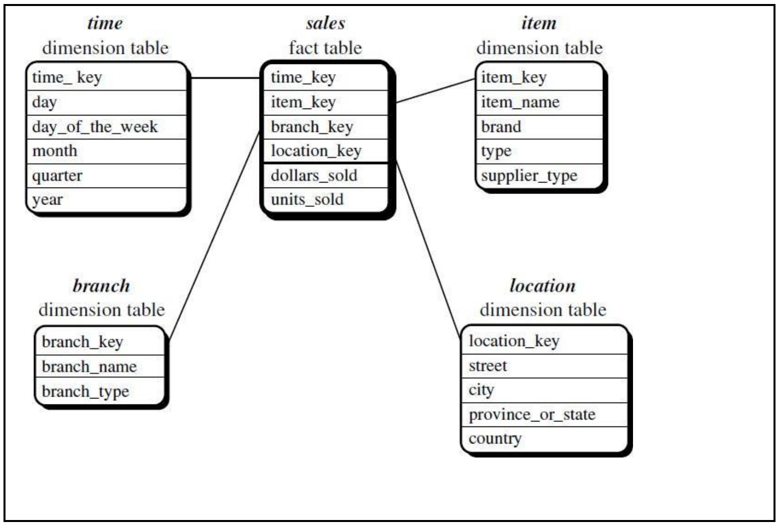


Management: Metadata can supports the content management and content administration entire its lifecycle, with the tasks i.e. control of the version, control of the access, rights of the management. It supports the different management tasks and different tasks and the activities, data resources are properly controlled and maintained with an organisation.

Metadata enabled the data management practices and accessibility of the digital resources of the entire lifecycle.
Identification: Metadata can helps the identify uniquely and shows the differences of individual content or information. Metadata can be assigned and managed the unique identifiers or tags for the each data asset and it enabling the efficient identification and retrieve entire the various platforms. By provide these unique identifiers and establishing the conventions, metadata enabled the perfect identification, retrieves the management of the data assets and supports the wide range of organizational processes.
Preservation: This metadata provides that certain information used for long-term preservation and captured the digital assets and maintained the management of digital assets. Preservation metadata captured the information about the technical, administrative, decision making actions. The capturing and manages the preserving metadata, organization that can be integrity, reliability and the usability of its digital collections, ensuring the accessibility for future generations.
Discovery: In Discovery, descriptive metadata utilize the discovery of content by searching the keywords, searching the subjects or other attributes. Metadata can be served as the discovering the related information or data by provide the attributes, relatable keywords that helped that the user can easily locate the resource and access the resources.
Navigation: Navigation is nothing but structural metadata is understand the structure of difficult content or dataset and organize the complex datasets into easily understandable to the users. It plays a major role to enabling the users to explore and access the data resources with the repositories or collections of the soft copy.
Interoperability: Metadata standards can enable the interoperability between the different platforms and systems, allowing to the exchange and data integration. Metadata can play the main role for promote the interoperability by provides the standard formats, structures and vocabularies for the attributes of the representing data and relationship along with the different environments.

Q2. b. Describe star and snowflake schema with neat diagram

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.
- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.
- Each dimension has only one dimension table and each table holds a set of attributes.
- For example, the location dimension table contains the attribute set {location_key,
- street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.



Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.

- The dimension tables are not normalized. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools.

Advantages:

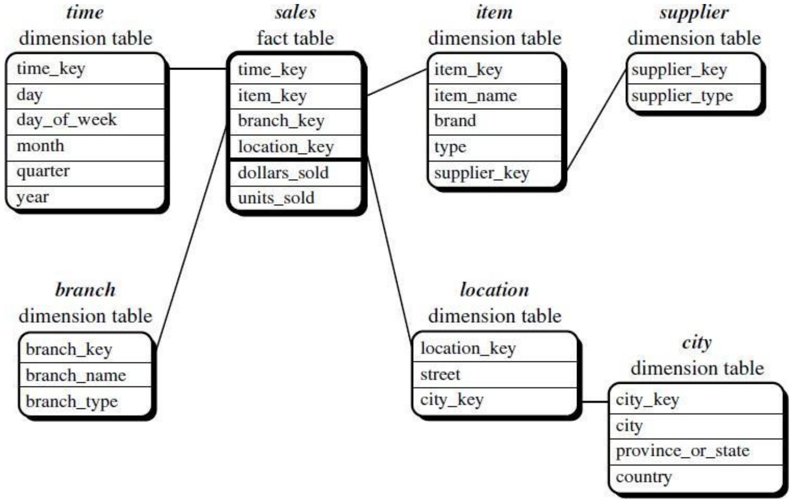
(i) Simplest and Easiest

(ii) It optimizes navigation through database

(iii) Most suitable for Query Processing

Snowflake Schema:

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized.
- For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.



- Now the item dimension table contains the attributes item_key, item_name, type,
- brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.
- A snowflake schemas can have any number of dimension, and each dimension can have any number of levels.
- The following diagram shows a snowflake schema with two dimensions, each having three levels.

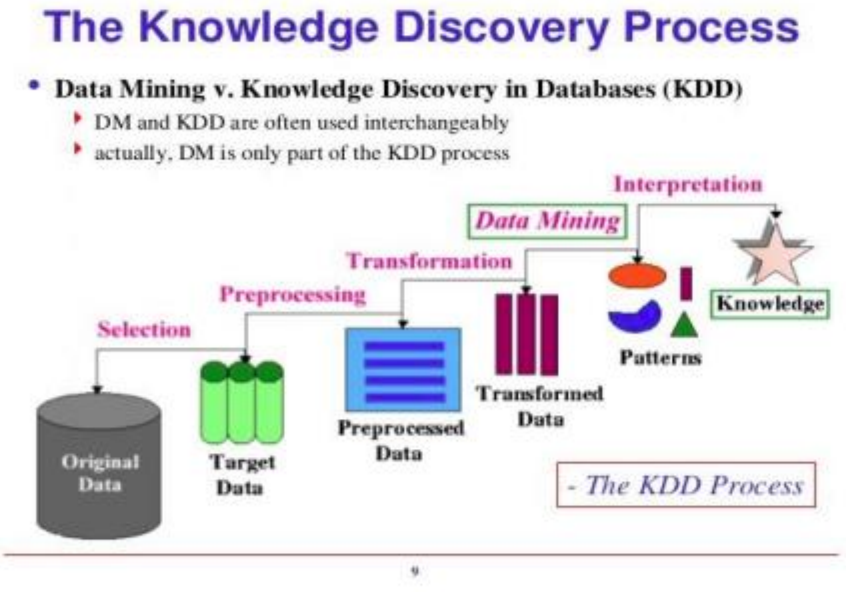
Advantages:
(i) Less redundancies due to normalization Dimension Tables.
(ii) Dimension Tables are easier to update.
Disadvantages:

- It is complex schema when compared to star schema.

Module-2

Q3. a. What is data mining? Explain Knowledge Data Discovery process in detail

Data Mining also known as Knowledge Discovery in Databases refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.



The Knowledge Discovery Process

- **Data Mining v. Knowledge Discovery in Databases (KDD)**
  - DM and KDD are often used interchangeably
  - actually, DM is only part of the KDD process

Figure 1.2 KDD Process

1. **Data Cleaning:** Data cleaning is defined as removal of noisy and irrelevant data from collection.
- ➢ Cleaning in case of Missing values.
- ➢ Cleaning noisy data, where noise is a random or variance error.
- ➢ Cleaning with Data discrepancy detection and Data transformation tools.

2. **Data Integration:** Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).
   - ➢ Data integration using Data Migration tools.
   - ➢ Data integration using Data Synchronization tools.
   - ➢ Data integration using ETL (Extract-Load-Transformation) process
3. **Data Selection**: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   - ➢ Data selection using Neural network.
   - ➢ Data selection using Decision Trees.
   - ➢ Data selection using Naive bayes.
   - ➢ Data selection using Clustering, Regression, etc.
4. **Data Transformation:** Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.
   Data Transformation is a two-step process:
   Data Mapping: Assigning elements from source base to destination to capture• transformations.
   Code generation: Creation of the actual transformation program.•
5. **Data Mining:** Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   - ➢ Transforms task relevant data into patterns
   - ➢ Decides purpose of model using classification or characterization.
6. Pattern Evaluation: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.
   - ➢ Find interestingness score of each pattern.
   - ➢ Uses summarization and Visualization to make data understandable by user.

7. Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
   - ➢ Generate reports.
   - ➢ Generate tables.
   - ➢ Generate discriminant rules, classification rules, characterization rules, etc.

Q3. b. Explain Data mining task primitives in detail

Data mining task primitives are fundamental components that define a data mining task. These primitives specify essential aspects of what the data mining process should achieve, including the type of data to analyze, the types of patterns to uncover, and the criteria for evaluating those patterns. By specifying task primitives, users can tailor the data mining process to their specific goals and needs. Here are the main task primitives in data mining, explained in detail:

1. **Task-Relevant Data Specification**

- **Description**: This primitive defines the data to be mined. It includes selecting the specific subset of data, attributes, and records that are relevant to the mining task.
- **Components**:
  - **Database or Data Warehouse**: Specifies the source of the data.
  - **Attributes or Fields**: Defines the specific fields (e.g., age, income, product type) to be considered.
  - **Data Segmentation or Filtering Criteria**: Specifies conditions or constraints to select a subset of records.
- **Example**: For a customer behavior analysis, the task-relevant data might include the "customer" database, with attributes like age, location, purchase history, and exclude records where "status = inactive."

2. **Kind of Knowledge to be Mined**

- **Description**: This primitive specifies the types of patterns or knowledge the mining task should uncover. It helps determine the data mining techniques to be applied based on the desired output.
- **Common Types**:
  - **Classification and Regression**: Identifying categories or predicting values (e.g., classifying customers into loyalty levels).
  - **Clustering**: Grouping similar data points (e.g., segmenting customers based on purchasing behavior).
  - **Association Rules**: Discovering relationships or associations between attributes (e.g., market basket analysis to find products often bought together).
  - **Sequential Patterns**: Finding patterns or trends over time (e.g., identifying purchase sequences).
  - **Outlier Detection**: Identifying data points that differ significantly from others (e.g., detecting fraudulent transactions).

- **Example**: If the goal is to discover purchasing patterns, the knowledge to be mined might include association rules among purchased items.

3. **Background Knowledge**

- **Description**: Background knowledge consists of domain knowledge, constraints, and hierarchies that can be used to influence the data mining process and refine the results. It helps in interpreting the patterns or applying domain-specific logic.
- **Types**:
    - **Hierarchies**: Allows for concept hierarchy, such as classifying locations by country, state, and city.
    - **Constraints and Business Rules**: Specifies rules that should influence mining, such as minimum support and confidence levels for association rules.
- **Example**: For a retail business, background knowledge might specify that product categories can be grouped into high-level categories like "electronics" and "apparel," and transactions must meet a minimum support threshold of 5%.

4. **Interestingness Measures**

- **Description**: Interestingness measures are criteria that assess the significance, novelty, and usefulness of the patterns found. They help filter out irrelevant patterns, ensuring that only meaningful insights are presented.
- **Types**:
    - **Objective Measures**: Quantitative metrics such as support (frequency of occurrence), confidence (probability of co-occurrence), or lift (measure of association strength) used in association rule mining.
    - **Subjective Measures**: Qualitative criteria based on user-defined expectations or business relevance, such as whether a pattern is surprising or useful in decision-making.
- **Example**: For an association rule in a market basket analysis, a rule like "If a customer buys milk, they also buy bread" might be considered interesting if it has at least 20% support and 80% confidence.

5. **Representation for Visualizing Discovered Patterns**

- **Description**: This primitive defines the format in which the mined patterns or insights should be presented to make them interpretable for the end-user. Effective representation ensures that insights can be easily understood and applied.
- **Common Formats**:
    - **Graphs and Charts**: Visual representations like bar charts, line graphs, or pie charts.
    - **Rules and Decision Trees**: Representation of classification or association rules, and visualization of decisions based on criteria.
    - **Clusters and Multi-dimensional Plots**: Representation of grouped data points or complex data in multiple dimensions.
    - **Textual Summaries**: Natural language summaries of insights or patterns.

- **Example**: In a customer segmentation task, clusters of customers may be visualized using scatter plots, where each point represents a customer and clusters are color-coded based on purchasing behavior.

Summary of Data Mining Task Primitives

Task Primitive	Purpose	Example
Task-Relevant Data Specification	Defines the subset of data, attributes, and records to analyze	Select "Customer" database with relevant fields
Kind of Knowledge to be Mined	Specifies types of patterns (e.g., classification, clustering, association)	Discover association rules in purchase history
Background Knowledge	Provides domain-specific information, such as hierarchies and constraints	Use "location" hierarchy and apply minimum support for patterns
Interestingness Measures	Determines criteria for evaluating and filtering patterns	Use support threshold of 20% and confidence of 80%
Representation for Visualization	Defines the format for presenting results, e.g., graphs, rules, decision trees, or textual forms	Show clusters in a scatter plot for customer segmentation

Q4. a. Explain Data cleaning process in detail

Data cleaning is an essential step in the data mining process. It is crucial to the construction of a model. The step that is required, but frequently overlooked by everyone, is data cleaning. The major problem with quality information management is data quality. Problems with data quality can happen at any place in an information system. Data cleansing offers a solution to these issues.

Data cleaning is the process of correcting or deleting inaccurate, damaged, improperly formatted, duplicated, or insufficient data from a dataset. Even if results and algorithms appear to be correct, they are unreliable if the data is inaccurate. There are numerous ways for data to be duplicated or incorrectly labeled when merging multiple data sources.

In general, data cleaning lowers errors and raises the caliber of the data. Although it might be a time-consuming and laborious operation, fixing data mistakes and removing incorrect information must be done. A crucial method for cleaning up data is data mining. A method for finding useful information in data is data mining. Data quality mining is a novel methodology that uses data mining methods to find and fix data quality issues in sizable databases. Data mining mechanically pulls intrinsic and hidden information from large data sets. Data cleansing can be accomplished using a variety of data mining approaches.

To arrive at a precise final analysis, it is crucial to comprehend and improve the quality of your data. To identify key patterns, the data must be prepared. Exploratory data mining is understood. Before doing business analysis and gaining insights, data cleaning in data mining enables the user to identify erroneous or missing data.

Steps for Cleaning Data
You can follow these fundamental stages to clean your data even if the techniques employed may vary depending on the sorts of data your firm stores:

1. Remove duplicate or irrelevant observations
Remove duplicate or pointless observations as well as undesirable observations from your dataset. The majority of duplicate observations will occur during data gathering. Duplicate data can be produced when you merge data sets from several sources, scrape data, or get data from clients or other departments. One of the most important factors to take into account in this procedure is de-duplication. Those observations are deemed irrelevant when you observe observations that do not pertain to the particular issue you are attempting to analyze.

You might eliminate those useless observations, for instance, if you wish to analyze data on millennial clients but your dataset also includes observations from earlier generations. This can improve the analysis's efficiency, reduce deviance from your main objective, and produce a dataset that is easier to maintain and use.

2. Fix structural errors
When you measure or transfer data and find odd naming practices, typos, or wrong capitalization, such are structural faults. Mislabelled categories or classes may result from these inconsistencies. For instance, "N/A" and "Not Applicable" might be present on any given sheet, but they ought to be analyzed under the same heading.

3. Filter unwanted outliers
There will frequently be isolated findings that, at first glance, do not seem to fit the data you are analyzing. Removing an outlier if you have a good reason to, such as incorrect data entry, will improve the performance of the data you are working with.

However, occasionally the emergence of an outlier will support a theory you are investigating. And just because there is an outlier, that doesn't necessarily indicate it is inaccurate. To determine the reliability of the number, this step is necessary. If an outlier turns out to be incorrect or unimportant for the analysis, you might want to remove it.

4. Handle missing data
Because many algorithms won't tolerate missing values, you can't overlook missing data. There are a few options for handling missing data. While neither is ideal, both can be taken into account, for example:

Although you can remove observations with missing values, doing so will result in the loss of information, so proceed with caution.

Again, there is a chance to undermine the integrity of the data since you can be working from assumptions rather than actual observations when you input missing numbers based on other observations.

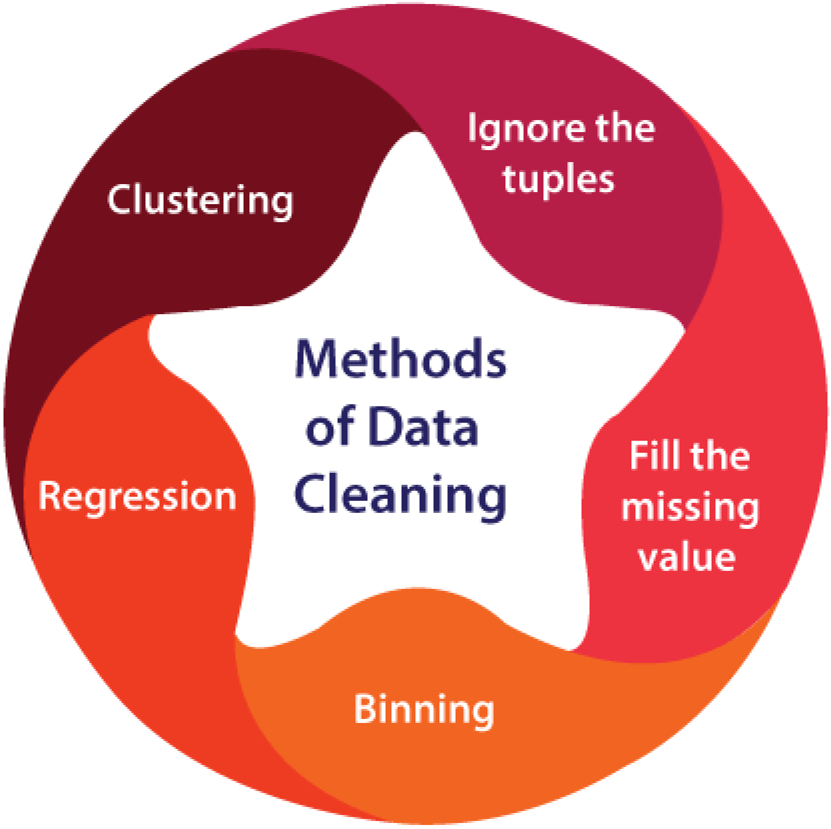
To browse null values efficiently, you may need to change the way the data is used.

5. **Validate and QA**

As part of fundamental validation, you ought to be able to respond to the following queries once the data cleansing procedure is complete:

Techniques for Cleaning Data

The data should be passed through one of the various data-cleaning procedures available. The procedures are explained below:



1. **Ignore the tuples:** This approach is not very practical because it is only useful when a tuple has multiple characteristics and missing values.
2. **Fill in the missing value:** This strategy is also not very practical or effective. Additionally, it could be a time-consuming technique. One must add the missing value to the approach. The most common method for doing this is manually, but other options include using attribute means or the most likely value.
3. **Binning method:** This strategy is fairly easy to comprehend. The values nearby are used to smooth the sorted data. The information is subsequently split into several equal-sized parts. The various techniques are then used to finish the assignment.

4. **Regression:** With the use of the regression function, the data is smoothed out. Regression may be multivariate or linear. Multiple regressions have more independent variables than linear regressions, which only have one.
5. **Clustering:** This technique focuses mostly on the group. Data are grouped using clustering. After that, clustering is used to find the outliers. After that, the comparable values are grouped into a "group" or "cluster".

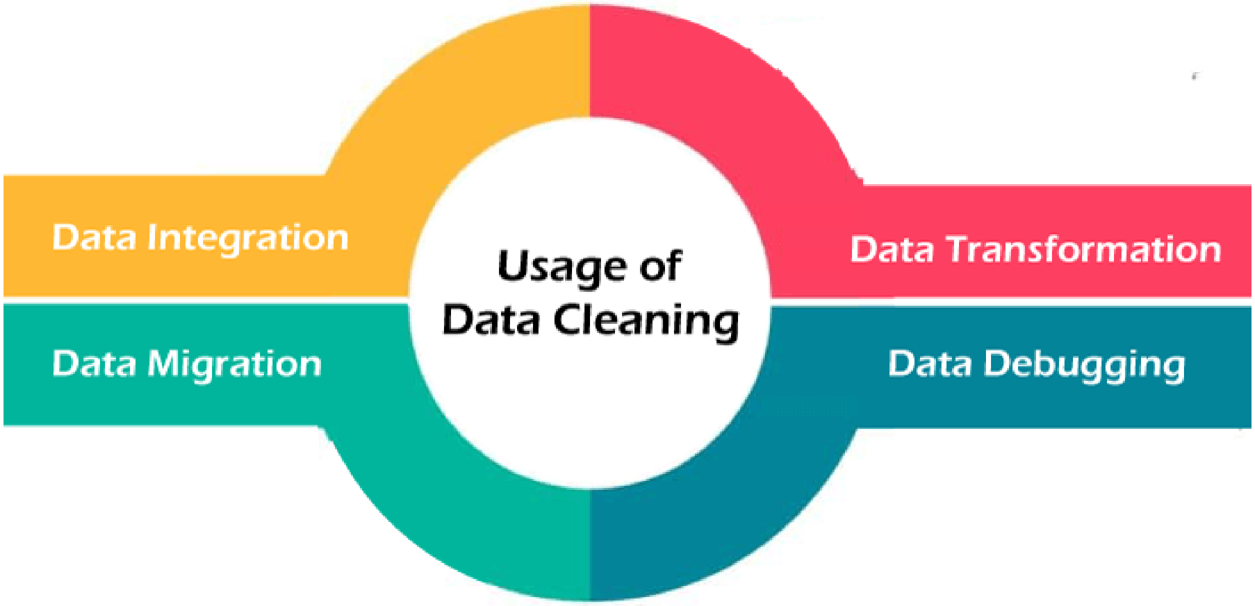
Process of Data Cleaning

The data cleaning method for data mining is demonstrated in the subsequent sections.

1. **Monitoring the errors:** Keep track of the areas where errors seem to occur most frequently. It will be simpler to identify and maintain inaccurate or corrupt information. Information is particularly important when integrating a potential substitute with current management software.
2. **Standardize the mining process:** To help lower the likelihood of duplicity, standardize the place of insertion.
3. **Validate data accuracy:** Analyse the data and spend money on data cleaning software. Artificial intelligence-based tools were utilized to thoroughly check for accuracy.
4. **Scrub for duplicate data:** To save time when analyzing data, find duplicates. By analyzing and investing in independent data-erasing technologies that can analyze imperfect data in quantity and automate the operation, it is possible to avoid again attempting the same data.
5. **Research on data:** Our data needs to be vetted, standardized, and duplicate-checked before this action. There are numerous third-party sources, and these vetted and approved sources can extract data straight from our databases. They assist us in gathering the data and cleaning it up so that it is reliable, accurate, and comprehensive for use in business decisions.
6. **Communicate with the team:** Keeping the group informed will help with client development and strengthening as well as giving more focused information to potential clients.

Usage of Data Cleaning in Data Mining.

The following are some examples of how data cleaning is used in data mining:

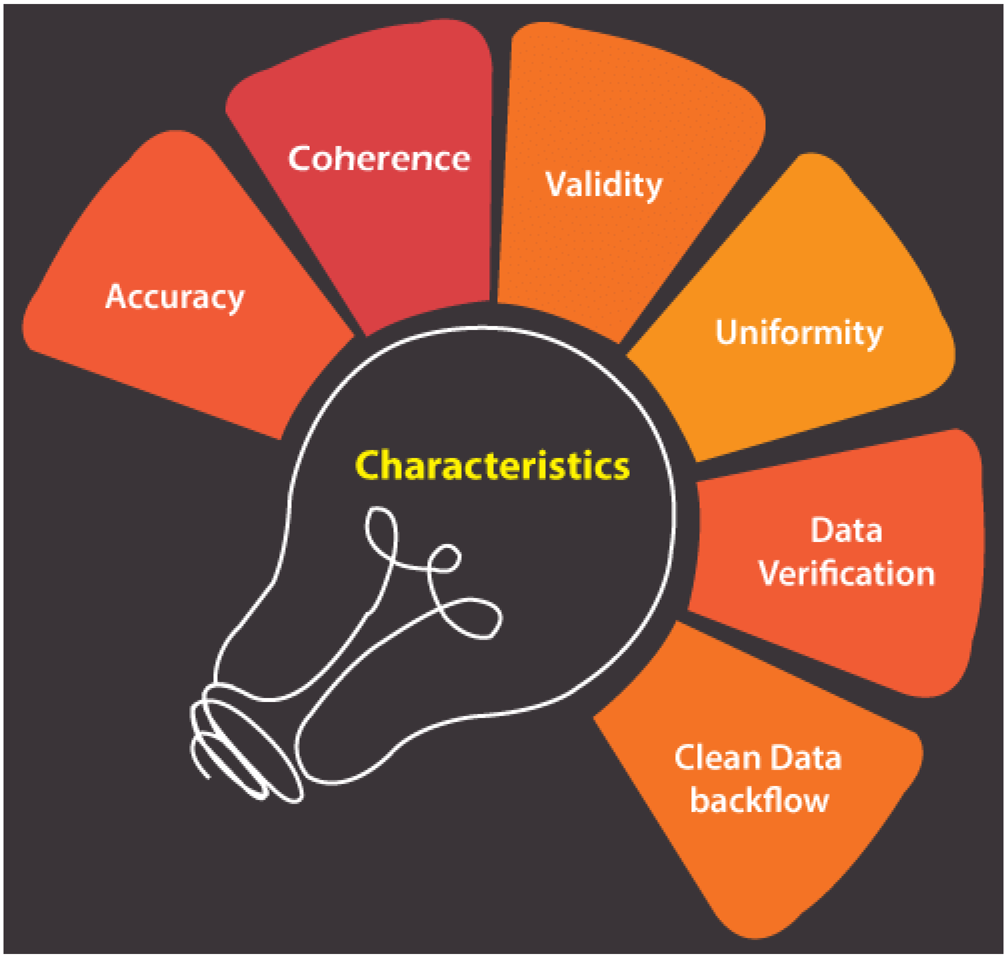


- **Data Integration:** Since it is challenging to guarantee quality with low-quality data, data integration is crucial in resolving this issue. The process of merging information from various data sets into one is known as data integration. Before transferring to the ultimate location, this step makes sure that the embedded data set is standardized and formatted using data cleansing technologies.
- **Data Migration:** The process of transferring a file from one system, format, or application to another is known as data migration. To ensure that the resulting data has the correct format, structure, and consistency without any delicacy at the destination, it is crucial to maintain the data's quality, security, and consistency while it is in transit.
- **Data Transformation:** The data must be changed before being uploaded to a location. Data cleansing, which takes into account system requirements for formatting, organizing, etc., is the only method that can achieve this. Before conducting additional analysis, data transformation techniques typically involve the use of rules and filters. Most data integration and data management methods include data transformation as a necessary step. Utilizing the systems' internal transformations, data cleansing tools assist in cleaning the data.
- **Data Debugging in ETL Processes:** To prepare data for reporting and analysis throughout the extract, transform, and load (ETL) process, data cleansing is essential. Only high-quality data are used for decision-making and analysis thanks to data purification.

Cleaning data is essential. For instance, a retail business could receive inaccurate or duplicate data from different sources, including CRM or ERP systems. A reliable data debugging tool would find and fix data discrepancies. The deleted information will be transformed into a common format and transferred to the intended database.

Characteristics of Data Cleaning

To ensure the correctness, integrity, and security of corporate data, data cleaning is a requirement. These may be of varying quality depending on the properties or attributes of the data. The key components of data cleansing in data mining are as follows:



- **Accuracy:** The business's database must contain only extremely accurate data. Comparing them to other sources is one technique to confirm their veracity. The stored data will also have issues if the source cannot be located or contains errors.
- **Coherence:** To ensure that the information on a person or body is the same throughout all types of storage, the data must be consistent with one another.
- **Validity:** There must be rules or limitations in place for the stored data. The information must also be confirmed to support its veracity.
- **Uniformity:** A database's data must all share the same units or values. Since it doesn't complicate the process, it is a crucial component while doing the Data Cleansing process.
- **Data Verification:** Every step of the process, including its appropriateness and effectiveness, must be checked. The study, design, and validation stages all play a role in the verification process. The disadvantages are frequently obvious after applying the data to a specific number of changes.
- **Clean Data Backflow:** After addressing quality issues, the previously clean data must be replaced with data that is not present in the source so that legacy applications can profit from it and avoid the need for a subsequent data-cleaning program.

Tools for Data Cleaning in Data Mining

Data Cleansing Tools can be very helpful if you are not confident of cleaning the data yourself or have no time to clean up all your data sets. You might need to invest in those tools, but it is worth the expenditure. There are many data cleaning tools in the market. Here are some top-ranked data cleaning tools, such as:

1. OpenRefine
2. Trifacta Wrangler
3. Drake
4. Data Ladder
5. Data Cleaner
6. Cloudingo
7. Reifier
8. IBM Infosphere Quality Stage
9. TIBCO Clarity
10. Winpure

Benefits of Data Cleaning

When you have clean data, you can make decisions using the highest-quality information and eventually boost productivity. The following are some important advantages of data cleaning in data mining, including:

- Removal of inaccuracies when several data sources are involved.
- Clients are happier and employees are less annoyed when there are fewer mistakes.
- The capacity to map out the many functions and the planned uses of your data.
- Monitoring mistakes and improving reporting make it easier to resolve inaccurate or damaged data for future applications by allowing users to identify where issues are coming from.
- Making decisions more quickly and with greater efficiency will be possible with the use of data cleansing tools.

Q4. b. Describe any two data compression methods in detail

Data compression is essential in data storage and transmission, as it reduces the amount of storage space needed or the bandwidth required for data transmission. Two commonly used data compression methods are **Lossless Compression** and **Lossy Compression**. These methods are often used in various data applications, depending on the need for data fidelity and the desired compression ratio. Let’s look at these methods in detail.

1. Lossless Compression

- **Description**: Lossless compression is a data compression method that allows the original data to be perfectly reconstructed from the compressed data without any loss of information. This type of compression is crucial when exact replication of data is required, such as with text files, executable files, and some types of image and sound files.
- **Common Techniques**:
  - **Huffman Coding**:
    - **Process**: Huffman Coding is a variable-length coding method that assigns shorter codes to frequently occurring symbols and longer codes to less frequent symbols. This technique uses a binary tree structure where each leaf node represents a symbol, and the path to each leaf node represents the binary code for that symbol.

        - **Example**: If a text file has many occurrences of the letter "e," Huffman coding will assign a shorter binary code (e.g., "1") to it, while less common letters like "z" may receive longer codes.
        - **Advantages**: Huffman coding is highly efficient for data with skewed symbol distributions and works well with data that has repeating elements.
    - **Run-Length Encoding (RLE)**:
        - **Process**: Run-Length Encoding compresses data by identifying and encoding consecutive repeating characters or symbols as a single character followed by the count of repetitions. This technique is effective for data with long sequences of repeated elements, such as binary data or images with large regions of the same color.
        - **Example**: In a sequence like "AAAAABBBBCC," RLE would compress it to "5A4B2C," where each letter is followed by the number of times it repeats.
        - **Advantages**: RLE is easy to implement and works particularly well with data that has many runs or repeating patterns, like in black-and-white images or simple graphics.
- **Applications**: Lossless compression is used in ZIP files, PNG image format, GIF images, and FLAC audio files, where data integrity must be preserved.
- **Advantages**:
    - No data is lost, making it suitable for situations where exact reconstruction of the original data is necessary.
    - Useful for compressing text, executable files, and other data where fidelity is paramount.
- **Disadvantages**:
    - The compression ratio (the extent to which data size is reduced) is usually lower than lossy compression, making it less efficient for certain types of media, such as high-resolution images or audio.

2. Lossy Compression

- **Description**: Lossy compression is a data compression method that reduces file size by permanently removing certain data, especially data deemed less critical for user perception. It is typically used for media files like images, audio, and video, where a certain degree of data loss is acceptable as it is often imperceptible to the human senses.
- **Common Techniques**:
    - **Discrete Cosine Transform (DCT)**:
        - **Process**: DCT is widely used in image and audio compression. It transforms a signal or image from the spatial domain to the frequency domain, where data can be represented as a sum of sinusoids of varying frequencies. DCT then discards less important high-frequency components that the human eye or ear is less likely to notice, thereby achieving compression.
        - **Example**: In JPEG compression, DCT is applied to 8x8 pixel blocks of an image, and high-frequency components are discarded to reduce file size. This results in a compressed image that appears visually similar to the original but with a significantly reduced file size.

        - **Advantages**: DCT achieves high compression ratios and is computationally efficient, making it ideal for image and video compression.
    - **Transform Coding with Quantization**:
        - **Process**: This technique involves converting data into a different representation (e.g., from spatial to frequency domain) and then quantizing the data by rounding off small values to zero, effectively reducing the amount of information stored. Quantization eliminates less critical data by reducing precision, which is usually imperceptible.
        - **Example**: In MP3 audio compression, audio signals are broken down into frequency bands. Lower precision is used in frequency bands that the human ear is less sensitive to, allowing for smaller file sizes with minimal loss in perceived quality.
        - **Advantages**: This technique achieves significant reduction in file sizes, making it very efficient for multimedia compression.
- **Applications**: Lossy compression is used in JPEG images, MP3 audio files, and MPEG video formats, where the user tolerates a degree of quality loss for reduced storage or transmission requirements.
- **Advantages**:
    - Allows for very high compression ratios, resulting in smaller file sizes.
    - Ideal for applications where approximate data quality is acceptable, such as streaming audio and video.
- **Disadvantages**:
    - Some data is permanently lost, which may degrade quality if the compression level is too high.
    - Not suitable for data that requires exact reproduction, like text files or critical data archives.

Summary Table of Lossless vs. Lossy Compression

Feature	**Lossless Compression**	**Lossy Compression**
Data Integrity	No data loss; exact reconstruction possible	Some data is permanently removed
Compression Ratio	Lower compression ratios	Higher compression ratios
Techniques	Huffman Coding, Run-Length Encoding	Discrete Cosine Transform, Quantization
Applications	ZIP files, PNG images, GIF, FLAC audio	JPEG images, MP3 audio, MPEG video
Ideal For	Text files, executable files, precise data	Images, audio, and video where minor loss is tolerable

Module-3

Q5. a. Define concept description and explain briefly data generalization by Attribute oriented induction

The simplest kind of descriptive data mining is concept description. A concept usually refers to a collection of data such as frequent_buyers, graduate_students, and so on. As a data mining task, concept description is not a simple enumeration of the data. Instead, concept description generates descriptions for characterization and comparison of the data. It is some times called class description, when the concept to be described refers to a class of objects. Characterization provides a concise and succinct summarization of the given collection of the data, while concept or class comparison (also known as discrimination) provides discriminations comparing two or more collections of data. Since concept description involves both characterization and comparison, techniques for accomplishing each of these tasks will study. Concept description has close ties with the data generalization. Given the large amount of data stored in database, it is useful to be describe concepts in concise and succinct terms at generalized at multiple levels of abstraction facilities users in examining the general behavior of the data. Given the ABCompany database, for example, instead of examining individual customer transactions, sales managers may prefer to view the data generalized to higher levels, such as summarized to higher levels, such as summarized by customer groups according to geographic regions, frequency of purchases per group, and customer income. Such multiple dimensional, multilevel data generalization is similar to multidimensional data analysis in data warehouses. The fundamental differences between concept description in large databases and online analytical processing involve the following.

Complex data types and aggregation:

Data warehouses and OLAP tools are based on a multidimensional data model that views data in the form of a data cube , consisting of dimensions (or attributes) and measures(aggregate functions). However, the possible data types of the dimensions and measures for most commercial versions of these systems are restricted. Many current OLAP systems confine dimensions to non-numeric data, similarly, measures (such as count (), sum (), average ()) in current OLAP systems apply only to numeric data. In contrast, for concept formation, the database attributes can be of various data types, including numeric, nonnumeric, spatial, text, or image. Furthermore, the aggregation of attributes in a database may include sophisticated data types, such as the collection of nonnumeric data, the merging of spatial region, the composition of images, the integration of texts, and the grouping of object pointers. Therefore, OLAP, with its restrictions on the possible dimension and measure types, represents a simplified model for data analyses. Concept description in databases can handle complex data types of the attributes and their aggregations, as necessary.

User-control versus automation:

On-line analytical processing in data warehouses is a purely user-controlled process. the selection of dimensions and the application of OLAP operations, such as drill-down, roll-up, slicing, and dicing, are directed and controlled by the users, although the control in most OLAP systems is quite user-friendly, users do require a good understanding of the role of each dimension. Furthermore, in order to find a satisfactory description of the data, users may need to specify a long sequence of OLAP operations. In contrast, concept description in data mining strives for a more automated process that helps determine which dimensions (or attributes) should be included in the analyses, and the degree to which the giver data set should be generalized in order to produce an interesting summarization of the data. Recently, data warehousing and OLAP technology has been evolving towards handling more complex types of data and embedding more knowledge discovery mechanisms. As this technology continues to develop , it is expected that additional descriptive data mining features will be integrated into future OLAP systems. Methods for concept description, including multilevel generalization, summarization, characterization, and comparison are outlined below. Such methods set the foundation for implementation of two major functional modules in data mining: multiple-level characterization and comparison. In addition, you will also examine techniques for the presentation of concept a description in multiple forms, including tables, charts, graphs, and rules.

Data Generalization and Summarization-Based Characterization

Data and objects in databases often contain detailed information at primitive concept levels. .For example, the item relation in sales database may contain attributes describing low-level item information such s item _ID , name , brand, category, supplier, place_made, and price. It is useful to be able to summarize a large set or data and present it at a high conceptual level.. For example, summarizing a large set of items relating to Christmas season sales provides a general description of such data , which can be very helpful for sales and marketing managers. This requires an important functionality in data mining: data generalization. Data generalization is a process that abstracts a large set of task-relevant data in a database from a relatively low conceptual level to higher conceptual levels. Methods for the efficient and flexible generalization of large data sets can be categorized according to two approaches :(1) the data cube (or OLAP) approach and (2) the attribute –oriented induction approach .In this section, we describe the attribute-oriented induction approach.

Attribute-Oriented Induction

The attribute-oriented induction (AOI)) approach to data generalization and summarization-based characterization was first proposed in 1989,a few years prior to the introduction of the data cube approach. The data cube approach can be considered as a data warehouse-based, pre-computationoriented, materialized-view approach. It performs off-line aggregation before an OLAP or data mining query is submitted for processing. On the other hand, the attribute-oriented induction approach, at least in its initial proposal, is a relational database query –oriented, generalization –based, on-line data analysis technique. However, there is no inherent barrier distinguishing the two approaches based on on-line aggregation versus off-line pre computation. Some aggregations in the data cube can be computed on-line, while off-line while off-line pre -computation of multidimensional space can speed up attribute –oriented induction as well.

Q5. b. Describe how class comparisons are performed with respect to attribute relevance

Class comparisons with respect to **attribute relevance** involve analyzing the significance of each attribute (or feature) in distinguishing between classes in a dataset. This process identifies which attributes contribute the most to classifying data accurately, helping to improve the quality of models by focusing on relevant features.

Steps for Class Comparison Using Attribute Relevance:

1. **Identify Relevant Attributes**:
    - Calculate statistical measures, such as **information gain**, **gain ratio**, **chi-square** scores, or **mutual information**, to quantify the importance of each attribute in predicting class labels.
    - These metrics help rank attributes by their contribution to distinguishing between classes.
2. **Attribute Selection**:
    - Select attributes with high relevance scores, indicating they are significant in differentiating between classes.
    - Irrelevant or redundant attributes, which contribute little to class separation, may be dropped to streamline the model.
3. **Class Distribution Comparison**:
    - For each class, analyze the distribution of attribute values to determine how well each attribute separates one class from another.

    - Techniques like **box plots**, **histograms**, or **mean comparisons** can help visualize differences in attribute values across classes.
4. **Correlation Analysis**:
    - Check for correlation between attributes and class labels. Highly correlated attributes are more relevant for class separation.
    - **Correlation coefficients** (e.g., Pearson or Spearman) are used to assess these relationships.
5. **Dimensionality Reduction (Optional)**:
    - Use dimensionality reduction techniques like **Principal Component Analysis (PCA)** or **Linear Discriminant Analysis (LDA)** to reduce data to a smaller set of relevant features.
    - This step improves interpretability and computational efficiency without sacrificing classification accuracy.

Benefits of Attribute Relevance Analysis in Class Comparison:

- **Improved Model Accuracy**: Models trained on relevant features often have better predictive accuracy and generalization.
- **Reduced Complexity**: By focusing on a subset of meaningful attributes, the model becomes simpler, faster, and less prone to overfitting.
- **Enhanced Interpretability**: Attribute relevance clarifies which features are most important for decision-making, making the model's logic easier to understand.

Q6. a. Explain Market basket analysis basic concepts in detail

This process analyzes customer buying habits by finding associations between the different items That customers place in their shopping baskets. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequentlypurchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket. Such information can lead to increased sales by helping retailers dos elective marketing and plan their shelf space.

Example:

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.

Q6. b. Explain Apriori algorithm in detail

The **Apriori principle** is a foundational concept in association rule mining that states:
If an itemset is frequent, then all its subsets must also be frequent.
Conversely:
If an itemset is infrequent, then all its supersets will also be infrequent.
This principle allows for efficient pruning of candidate itemsets when searching for frequent itemsets in large datasets. By eliminating infrequent itemsets early, the Apriori algorithm reduces the computational complexity of finding frequent itemsets.

Apriori Algorithm for Frequent Itemset Generation

The **Apriori algorithm** is a widely used technique for finding frequent itemsets in large databases. It leverages the Apriori principle to generate and prune candidate itemsets efficiently. The algorithm works iteratively, increasing the size of itemsets (from 1-itemset, 2-itemset, etc.) until no more frequent itemsets can be found.

Steps in Apriori Algorithm:

1. **Set Minimum Support Threshold**:
   - Before starting, a minimum support threshold is set to filter out infrequent itemsets.
2. **Generate 1-Itemsets (C1)**:
   - The algorithm first scans the dataset and counts the frequency (support) of each item. This generates a list of 1-itemsets.
3. **Prune Infrequent 1-Itemsets**:
   - Itemsets whose support is below the minimum support threshold are discarded. The remaining itemsets are called frequent 1-itemsets (L1).
4. **Generate Candidate 2-Itemsets (C2)**:
   - From the frequent 1-itemsets (L1), pairs of items (2-itemsets) are generated. These are called candidate 2-itemsets.
5. **Prune Infrequent 2-Itemsets**:
   - Again, the support of each 2-itemset is calculated, and those with support below the minimum threshold are removed, leaving frequent 2-itemsets (L2).
6. **Repeat Process for Larger Itemsets**:
   - The algorithm continues iteratively, generating candidate 3-itemsets (C3) from frequent 2-itemsets (L2), then candidate 4-itemsets (C4), and so on.
   - At each step, the infrequent itemsets are pruned based on the support threshold.
7. **Terminate**:
   - The algorithm stops when no further frequent itemsets can be generated (i.e., no more candidates pass the support threshold).
8. **Generate Association Rules**:
   - After generating all frequent itemsets, the algorithm can derive **association rules** (like X → Y) from these itemsets, which must satisfy the minimum **confidence** threshold.

Example of Apriori Algorithm:

Consider the following transactions:

Transaction ID	Items Bought
T1	Bread, Butter, Milk
T2	Bread, Butter
T3	Milk, Eggs

Transaction ID	Items Bought
T4	Bread, Butter, Eggs
T5	Butter, Eggs

Step 1: Set minimum support = 2 transactions (40%)

Step 2: Generate 1-itemsets and prune:

Item	Support Count	Pruned?
Bread	3	No
Butter	4	No
Milk	2	No
Eggs	3	No

All 1-itemsets are frequent.

Step 3: Generate 2-itemsets from frequent 1-itemsets:

2-Itemset	Support Count	Pruned?
{Bread, Butter}	3	No
{Bread, Milk}	1	Yes
{Bread, Eggs}	1	Yes
{Butter, Milk}	1	Yes
{Butter, Eggs}	2	No
{Milk, Eggs}	1	Yes

After pruning, only {Bread, Butter} and {Butter, Eggs} remain as frequent 2-itemsets.

Step 4: Generate 3-itemsets:

3-Itemset	Support Count	Pruned?
{Bread, Butter, Eggs}	1	Yes

There are no frequent 3-itemsets, so the algorithm terminates.

Efficiency of the Apriori Algorithm

The Apriori algorithm is efficient due to its ability to **prune the search space**. Instead of generating all possible itemsets, it focuses only on the frequent itemsets, significantly reducing the number of calculations.

However, the algorithm may still have limitations with very large datasets because of multiple passes over the data, which can be time-consuming.

Module-4

Q7. a. Define classification and prediction and enlist the issues regarding classification and prediction

Classification and Prediction:

- Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.
- Classification predicts categorical (discrete, unordered) labels, prediction models
- Continuous valued functions.
- For example, we can build a classification model to categorize bankloan applications as
- either safe or risky, or a prediction model to predict the expendituresof potential customers on computer equipment given their income and occupation.
- A predictor is constructed that predicts a continuous-valued function, or ordered value, as
- opposed to a categorical label.
- Regression analysis is a statistical methodology that is most often used for numeric
- prediction.
- Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics.
- Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data.

Issues Regarding Classification and Prediction:
1**.Preparing the Data for Classification and Prediction:**
The following preprocessing steps may be applied to the data to help improve the accuracy, efficiency, and scalability of the classification or prediction process.
(i)Data cleaning:
This refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques) and the treatment of missingvalues (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics).
Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning.
(ii)Relevance analysis:
Many of the attributes in the data may be redundant.
Correlation analysis can be used to identify whether any two given attributes are statisticallyrelated.
For example, a strong correlation between attributes A1 and A2 would suggest that one of the two could be removed from further analysis.
A database may also contain irrelevant attributes. Attribute subset selection can be used in these cases to find a reduced set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.
Hence, relevance analysis, in the form of correlation analysis and attribute subset

selection, can be used to detect attributes that do not contribute to the classification or prediction task.

Such analysis can help improve classification efficiency and scalability.

(iii)Data Transformation And Reduction

The data may be transformed by normalization, particularly when neural networks or methods involving distance measurements are used in the learning step.

Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as -1 to +1 or 0 to 1.

The data can also be transformed by generalizing it to higher-level concepts. Concept hierarchies may be used for this purpose. This is particularly useful for continuous valuedattributes.

For example, numeric values for the attribute income can be generalized to discrete ranges, such as low, medium, and high. Similarly, categorical attributes, like street, can be generalized to higher-level concepts, like city.

Data can also be reduced by applying many other methods, ranging from wavelet transformation and principle components analysis to discretization techniques, such as binning, histogram analysis, and clustering.

Q7. b. Discuss briefly Decision tree and Bayesian classfication

Algorithm for Decision Tree Induction

Input:

- Training dataset `D` with `n` attributes and `m` instances.
- Target attribute (class label).

Output:

- A decision tree.

Steps:

1. **Start**
   - If all instances in `D` belong to the same class, return a single-node tree with that class label.
2. **Attribute Selection**
   - For each attribute `A`, compute the **information gain** (or **Gini index**, **Gain ratio**, etc., depending on the criterion) to determine how well `A` classifies the instances.
   - Select the attribute `A` that has the highest information gain as the **splitting attribute**.
3. **Create Node**
   - Create a decision node in the tree corresponding to the selected attribute `A`.
4. **Partition the Dataset**

  - Partition the dataset `D` into subsets `D1, D2, ..., Dk`, based on the values of the selected attribute `A`.
  - For each subset `Di` corresponding to a value `vi` of attribute `A`:
    - If `Di` is empty, create a leaf node with the most common class label from the parent set.
    - If `Di` contains instances with more than one class, repeat the process recursively using `Di` as the new dataset.

5. **Stop Condition**
   - The recursion stops when:
     - All instances in a subset belong to the same class.
     - There are no more attributes to split on.
     - The dataset is empty.
6. **Return Tree**
   - Once all subsets are processed, return the complete decision tree.

Algorithm: DecisionTreeInduction(D, attributes)

```
 Input:
   D - Dataset
   attributes - List of attributes

 Output:
   Decision tree

 if all instances in D have the same class label then
   return a leaf node with that class label
 else if attributes is empty then
   return a leaf node with the most common class label in D
 else
   A = Attribute with highest information gain from attributes
   Create a decision node with A as the splitting attribute
   for each value vi of attribute A do
     Di = Subset of D where A = vi
     if Di is empty then
       Add a leaf node with the most common class label in D
     else
       Add the subtree DecisionTreeInduction(Di, attributes - A) to the current node
   return the decision node
```

Naïve Bayes Classifier

The **Naïve Bayes Classifier** is a probabilistic machine learning model based on **Bayes' Theorem**. It is widely used for classification tasks due to its simplicity, efficiency, and effectiveness, especially with large datasets. Despite its simplicity, it performs surprisingly well for various real-world applications like spam filtering, text classification, sentiment analysis, and medical diagnosis.

Key Assumption: Conditional Independence

The "naïve" part of Naïve Bayes comes from the assumption that all the features (or attributes) are **independent** of each other, given the class label. This means that the presence or absence of a particular feature in a class is assumed to be independent of the presence or absence of any other feature. While this assumption rarely holds in real-world situations, Naïve Bayes still performs well in many cases, even when the assumption is violated.

Bayes' Theorem

Naïve Bayes is based on **Bayes' Theorem**, which relates the probability of a class given a set of features to the probability of the features given the class. Bayes' Theorem is stated as:

P(C|X)=P(X|C)·P(C)P(X)P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}P(C|X)=P(X)P(X|C)·P(C)

Where:

- P(C|X)P(C|X)P(C|X) is the **posterior probability**: the probability of class C given the feature vector X.
- P(X|C)P(X|C)P(X|C) is the **likelihood**: the probability of the feature vector X given the class C.
- P(C)P(C)P(C) is the **prior probability**: the overall probability of class C.
- P(X)P(X)P(X) is the **evidence**: the overall probability of the feature vector X.

Working of Naïve Bayes Classifier

1. **Training Phase**:
   - Compute the prior probability for each class.
   - For each feature, compute the conditional probability of the feature value given the class label.
2. **Prediction Phase**:
   - For a new data point, calculate the posterior probability for each class using Bayes' Theorem.
   - Assign the class label with the highest posterior probability.

Advantages

- **Efficient**: It is computationally efficient and works well with large datasets.
- **Simple to Implement**: Easy to understand and implement, even with basic mathematical knowledge.
- **Works Well with High-Dimensional Data**: It performs well in text classification tasks where the data has many features.

Limitations

- **Strong Assumption of Independence**: The assumption that all features are independent is often unrealistic in real-world data.
- **Zero Probability Problem**: If a particular feature value is missing in the training dataset for a class, the model will assign a zero probability to it. This is often mitigated by using techniques like Laplace smoothing.

Applications

- **Spam Filtering**: Classifies emails as spam or non-spam based on the frequency of words.
- **Text Classification**: Categorizes documents into predefined categories, such as news articles.
- **Sentiment Analysis**: Determines the sentiment (positive/negative) of text, such as customer reviews.

Q8. a. Explain classification and Regression Tree[CART] in detail

➢ A decision tree is a tree-like model that is used for making decisions. It consists of nodes that represent decision points, and branches that represent the outcomes of those decisions. The decision points are based on the values of the input variables, and the outcomes are the possible classifications or predictions.

▸ A decision tree is constructed by recursively partitioning the input data into subsets based on the values of the input variables. Each partition corresponds to a node in the tree, and the partitions are chosen so as to minimize the impurity of the resulting subsets.

The algorithm works by recursively partitioning the training data into smaller subsets using binary splits. The tree starts at the root node, which contains all the training data, and recursively splits the data into smaller subsets until a stopping criterion is met.

▸ At each node of the tree, the algorithm selects a feature and a threshold that best separates the
training data into two groups, based on the values of that feature. This is done by choosing the feature and threshold that maximizes the information gain or the Gini impurity, which are measures of how well a split separates the data.

▸ The process continues recursively, with each node in the tree splitting the data into two smaller
subsets, until a stopping criterion is met. The stopping criterion could be a maximum depth for the tree, a minimum number of instances in each leaf node, or other criteria.

▸ Once the tree is built, it can be used to make predictions by traversing the tree from the root node to a leaf node that corresponds to the input data. For regression problems, the prediction is the average of the target values in the leaf node. For classification problems, the prediction is the majority class in the leaf node.

Calculate the Gini impurity for the entire dataset. This is the impurity of the root node.

▸ For each input variable, calculate the Gini impurity for all possible split points. The split point that results in the minimum Gini impurity is chosen.

- The data is split into two subsets based on the chosen split point, and a new node is created for each subset.
- Steps 2 and 3 are repeated for each new node, until a stopping criterion is met. This stopping criterion could be a maximum tree depth, a minimum number of data points in a leaf node, or a minimum reduction in impurity.
- The resulting tree is the decision tree.

To illustrate how the Gini impurity is calculated, consider the following example. Suppose we have a binary classification problem, where we want to predict whether a person will buy a particular product based on their age and income.

Age	**Income**	**Buy Product?**
30	20000	Yes
40	50000	Yes
20	30000	No
50	60000	No
60	80000	Yes

- We want to build a decision tree to predict whether a person will buy the product based on their age and income. Suppose we want to split the data based on age, with a threshold of 35.

The left child node will contain the data where age is less than or equal to 35, and the right child node will contain the data where age is greater than 35.

The Gini impurity for the left child node can be calculated as follows:

Gleft = 1 − (12)2 − (12)2 = 0.5

There are two data points in the left child node, one of which buys the product, and one of which does not.

- Therefore, the probability of a randomly chosen element being labeled as "Yes" is 0.5, and the probability of it being labeled as "No" is also 0.5.
- The Gini impurity for the right child node can be calculated as follows:

Gright = 1 − (23)2 − (13)2 ≈ 0.444

There are three data points in the right child node, two of which buy the product, and one of which does not.

- Therefore, the probability of a randomly chosen element being labeled as "Yes" is 2/3, and the probability of it being labeled as "No" is 1/3.

The overall Gini impurity for the split can be calculated as a weighted average of the Gini impurities for the child nodes, where the weights are proportional to the number of data points in each node.

Gsplit =25Gleft +35Gright ≈ 0.48

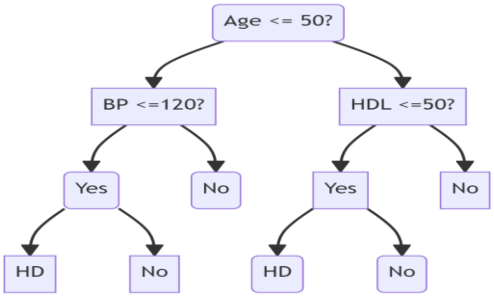
- The decision tree algorithm will try different thresholds and different features to find the split that minimizes the Gini impurity.

Let’s consider an example of using the CART algorithm for a binary classification problem.

- Suppose we have a dataset of patients with information about their age, gender, blood pressure, and cholesterol level, and whether or not they have heart disease. We want to build a decision tree to predict whether a new patient will have heart disease based on their age,gender, blood pressure, and cholesterol level.

- To start, we calculate the Gini impurity of the entire dataset. Suppose there are 500 patients in the dataset, and 200 of them have heart disease, and 300 of them do not. The Gini impurity is:
G = 1 − (200/500)2 − (300/500)2 = 0.48
- Next, we consider each input variable and each possible split point. Suppose we choose age as the first split variable. We consider all possible split points and calculate the Gini impurity for each split. Suppose the split point that results in the minimum Gini impurity is 50 years.
We split the data into two subsets: patients who are 50 years old or younger, and patients who are older than 50. We create two new nodes for these subsets and calculate the Gini impurity for each node.
- Suppose the first node contains 300 patients, of which 100 have heart disease and 200 do not.
The Gini impurity of this node is:
G1 = 1 − (100/300)2 − (200/300)2 = 0.44
- Suppose the second node contains 200 patients, of which 100 have heart disease and 100 do not.
The Gini impurity of this node is:
G2 = 1 − (100/200)2 − (100200)2 = 0.5
- We choose the split that results in the minimum Gini impurity, which is the split on age at 50.
- We create two new nodes for the subsets and continue the process until we meet a stopping criterion.

- Suppose we set the stopping criterion to be a maximum tree depth of 2. The resulting decision tree would look like this:



- This decision tree can be used to predict whether a new patient will have heart disease based on their age, blood pressure, and cholesterol level.

Advantages

- It is a simple and intuitive algorithm that is easy to understand and interpret.

- It can handle both numerical and categorical data.
- It can handle missing values by imputing them with surrogate splits.
- It can handle multi-class classification problems by using an extension called the multi-class CART.

Disadvantages

- It tends to overfit the data, especially if the tree is allowed to grow too deep.
- It is a greedy algorithm that may not find the optimal tree.
- It may be biased towards predictors with many categories or high cardinality.
- It may produce unstable results if the data is sensitive to small changes or noise.

Q8. b. Describe in detail Linear and Logistic Regression

Linear Regression:

Straight-line regression analysis involves a response variable, y, and a single predictor variable x.

It is the simplest form of regression, and models y as a linear function of x.

That is,y = b+wx

where the variance of y is assumed to be constant

band w are regression coefficientsspecifying the Y-intercept and slope of the line.

The regression coefficients, w and b, can also be thought of as weights, so that we can equivalently write, y = w0+w1x

These coefficients can be solved for by the method of least squares, which estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line.

Let D be a training set consisting of values of predictor variable, x, for some population and their associated values for response variable, y. The training set contains |D| data points of the form(x1, y1), (x2, y2), ... , (x|D|, y|D|).

The regression coefficients can be estimated using this method with the following equations:

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1\bar{x}$$

where x is the mean value of x1, x2, ... , x|D| , and y is the mean value of y1, y2,..., y|D|
.

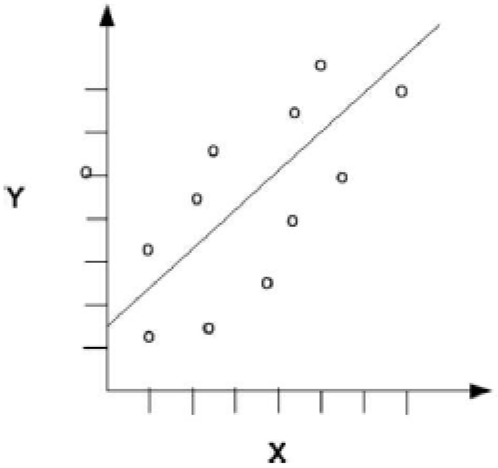
The coefficients w0 and w1 often provide good approximations to otherwise complicated regression equations.

Multiple Linear Regression:

- It is an extension of straight-line regression so as to involve more than one predictor variable.
- It allows response variable y to be modeled as a linear function of, say, n predictor variables or attributes, A1, A2, ..., An, describing a tuple, X.

- An example of a multiple linear regression model basedon two predictor attributes or
- variables, A1 and A2, isy = w0+w1x1+w2x2 where x1 and x2 are the values of attributes A1 and A2, respectively, in X.
- Multiple regression problemsare instead commonly solved with the use of statistical software packages, such as SAS,SPSS, and S-Plus.

Figure 4-1 Linear Relationship Between x and y



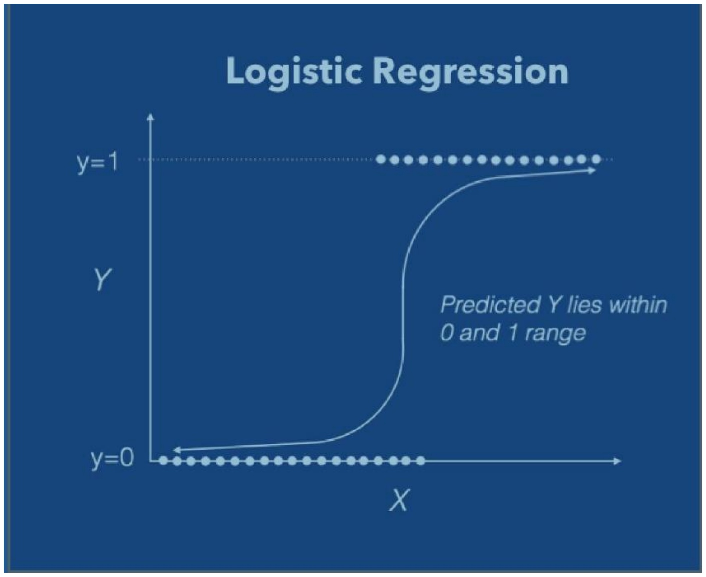
Logistic Regression:

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,

➢To predict whether an email is spam (1) or (0)

➢whether the tumor is malignant (1) or not (0)



Module-5

Q9. a. Explain data mining for Business Intelligence

Business intelligence (BI) is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining.

Business intelligence applications can be:

- Mission-critical and integral to an enterprise's operations or occasional to meet a special requirement
- Enterprise-wide or local to one division, department, or project
- Centrally initiated or driven by user demand.

Applications:

a) Balanced Scorecard

1. The Balance Scorecard (BSC) is a framework for managing business performance.

2. BSC provides a framework for designing a set of measures for business activities as being the key drivers of the business or Key Perfomance Indicators(KPIs).

3. KPIs are collected from CRM,ERP, Accounting , Personnel, Inventory.

4. BSC provides executives and managers with a method for reporting and analyzing key performance indicators to determine if operational activities are aligned with the company's overall strategy and vision.

5. The BSC methodology is a management technique for structuring these scorecards and displays financial, internal process, customer and learning/growth data.

6. The balance scorecard takes four perspectives:

- Financial: Satisfying the stakeholders in the company – owners, employees, suppliers. For e.g. the objectives of this perspective would be to achieve a certain level of profitability, or growth.

- Customer or Market: Satisfying the customers such that they buy product and services to support the Financial perspective, e.g. increase customer satisfaction, introduce a new product.

- Internal Business Processes: Supporting the Financial and Customer perspectives through having appropriate and well operated processes or procedures e.g. the sales process, the product implementation process.

- Learning, Innovation and Growth: Supporting the Financial, Customer and Internal Business Process perspectives through having the ability to change, improve and innovate through the acquisition of new knowledge, skills and technology.

7. Example of BSC:

- Knowledge-enhanced Predictive Reports(KPRs) can improve business visibility harnessing BSC with predictive modeling and business logic using expert systems.
- KPRs can analyze changes in business drivers and co-inference them automatically to detect hidden patterns underneath complex numbers.
- KPRs incorporate predictive modeling with rule-based expert systems into report writing and charting systems.

- Predictive analytics can be used to detect patterns and trends in business drivers automatically from hidden numbers, and to predict future directions.

- Rule-based expert systems can be used to leverage complexity of various business drivers and indicators. Expert systems based on business logic can take this task as an expert, making balanced scorecards friendlier and easier to understand.

- Web-based reporting and charting engines are essential in generating balanced scorecards in a timely real-time fashion so that executives and business users can recognize developing situation in real-time.

- Incorporation of predictive analytics and rule-based expert systems into BSC provides a number of advantages:

  - It will make BSC much easier to comprehend potential problems and successes.
  - Trends developing can be detected early so that actions can be taken quickly
  - Complexity in interpreting KPIs is removed in real-time by embedded knowledge of business experts

b) Fraud Detection

Fraud Detection for Telecommunications Industry
1. The telecommunications industry has expanded dramatically in the last few years with the development of affordable mobile phone technology.

2. With the increasing number of mobile phone users, global mobile phone fraud is also set to rise.

3. There are many different types of telecom fraud and these can occur at various levels.

4. The two most prevalent types are subscription fraud and superimposed or surfing.

5. Subscription fraud occurs when the fraudster obtains a subscription to a service, often with false identity details, with no intention of paying. This is thus at the level of a phone number – all transactions from this number will be fraudulent.

6. Superimposed fraud is the use of a service without having the necessary authority and is usually detected by the appearance of phantom calls on a bill.

7. There are several ways to carry out superimpose fraud, including mobile phone cloning and obtaining calling card authorization details.

8. Superimposed fraud will generally occur at the level of individual calls – the fraudulent calls will be mixed in with the legitimate ones.

9. Subscription fraud will generally be detected at some point through the billing process – although the aim is to detect it well before that, since large costs can quickly be run up.

10. Superimpose fraud can remain undetected for a long time.

11. Telecommunications networks generate vast quantities of data, sometimes on the order of several GBs per day, so that data mining techniques are of particular importance.

12. At a low level, simple rule-based detection systems use rules such as the apparent use of the same phone in two very distant geographical locations in quick succession, calls which appear to overlap in time, and very high value and long calls.

13. At a higher level, statistical summaries of call distributions are compare with thresholds determined either by experts or by application of supervised learning methods to known fraud/non-fraud cases.

c) Clickstream Mining

1. The approach which is used by most of the people for surfing information on Websites is difficult to analyze and understand.

2. Quantitative data can lack information about what a user actually intends to do, while qualitative data tends to be localized and is impractical to gather for large samples.
3. Once a website is made public, the user is in ultimate control of their own navigation, often employing a variety of different strategies for browsing.

4. These strategies also vary over time depending, not only on the user's goals, but also on factors such as expertise, familiarity with the site, time pressures and perceive cost of information.

5. Given this continually shifting nature of browsing strategies, the question arises how can these strategies be identified in the use made of an existing Website.

6. One solution is to use the clickstream logs, which contain the address of each page visited, the date and time of the visit and the referring page and are potentially rich source of data on Internet user activity.

7. Clickstream logs can be generate either by software hosted by the client application or directly from the server logs.

8. Collection and Restoration of Clickstream Data:

- A common tool for collecting data on the pages visited by Website users is the use of server-side clickstream data.

- This identifies the pages delivered by a server in response to a client's request. However, these clickstream data logs are often large and unwieldy and present an incomplete picture of activity.

- For example, server-side logs do not record activities that involve browser caching, network caching, or the navigation of pages that are internal to the site but are held on another server.
- Despite these server-side limitations, there are some aspects of user behavior, such as use of the back button or the opening of new/additional windows within the same Website, that can be captured by such techniques such as the Pattern Restore Method(PRM) algorithm.

9. Visualization and Categorization of Clickstream Data:

- Once the clickstream data have been processed, a technique for analyzing and categorizing these data into usage patterns is required.
- The visualization techniques facilitate this by producing ‘Footstep’ graphs.

- These are based on the use of a 2-D x-y plot, where x-axis represents the browsing time between two Web pages and the y-axis the Web page in the users browsing route.
- Thus, the distance travelled on the x-axis represents the time the user has spent browsing and a change in the y-axis represents a transition from one Web page to another

d) Market Segmentation

1. Market Segmentation is a process that segments a market into smaller sub-markets, called segments.

2. Segments are to be homogeneous or have similar attributes.

3. Purchasing patterns and trends can appear prominently in certain segments.

4. Good market segmentation is to create segments where prominent patters can emerge.

5. Market segmentation may be use to analyze the followings:

Market responsiveness analysis: Useful in direct marketing since market responsiveness of product offerings can be readily available.

Market trend Analysis: Analyzing segment-by-segment changes of sales revenues can reveal market trends. Trending information is vital in preparing for ever-changing markets.

It may use one of the following attributes to generate market segments:

- Geographical Regions: Regions, countries, states, zip-codes, countries , etc.
- Demographics: gender , age, income, education etc
- Psychographics: Life style classification
- Sales channels, branches and departments
- Sales representatives

- Product and service types (or product categories)
- Products
- Offer types

6. Segmentation provides opportunities for trend analysis. Trends and patterns embedded in changes of sales revenues can be useful indicators for market shifts. Trend analysis may analyse the following types of segment trend information:

- What are the projected sales revenues for the next three months?
- Which segments are having the highest growth and which segments are having the highest revenue decline?
- Which segments are having the highest growth rates in percentage terms?

7. Sales Trend Analysis:

Timely identification of newly emerging trends is very important to businesses.

Sales patterns of customer segments indicate market trends. Upward and downwards trends in sales signify new market trends. Time-series predictive modeling can be used to identify trends embedded in changes of sales revenues. Understanding of sales trends is important for marketing as well as for customer retention. Typical sales trend analysis includes:

- Which customer segments are having highest growth and highest revenue decline ?
- Which customer segments are having highest growth rates in percentage terms?

8. Trends may be categorized as:

- Short term trends capture rapidly emerging trends
- Mid-term trends capture trends developing in between
- Long term trends capture trends developing over long periods.

e) Retail Industry

1. The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service. The quantity of data collected continues to expand rapidly.

2. Retail data mining can help :

- identify customer buying behaviors,
- discover customer shopping patterns and trends,
- improve the quality of customer service,

- achieve better customer retention and satisfaction,
- enhance goods consumption ratios,
- design more effective goods transportation and distribution policies,
- reduce the cost of business.

3. Design and construction of data warehouses based on the benefits of data mining:

- There can be many ways to design a data warehouse for this industry.
- The levels of detail to include may also vary substantially.
- The outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures.
- This involves deciding which dimensions and levels to include and what preprocessing to performing order to facilitate effective data mining.

4. Multidimensional analysis of sales, customers, products, time, and region:

- The retail industry requires timely information regarding customer needs, product sales, trends and fashions, as well as the quality, cost, profit, and service of commodities.
- It is therefore important to provide powerful multidimensional analysis and visualization tools, including the construction of sophisticated data cubes according to the needs of data analysis.

5. Analysis of the effectiveness of sales campaigns:

- The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers.
- Careful analysis of the effectiveness of sales campaigns can help improve company profits.
- Association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

6. Customer retention—analysis of customer loyalty:

- Customer loyalty and purchase trends can be analyzed systematically. Goods purchased at different periods by the same customers can be grouped into sequences.
- Sequential pattern mining can then be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new ones.

7. Product recommendation and cross-referencing of items:

- By mining associations from sales records, one may discover that a customer who buys a digital camera is likely to buy another set of items.
- Such information can be used to form product recommendations.
- Product recommendations can also be advertised on sales receipts, in weekly flyers, or on the Web to help improve customer service, aid customers in selecting items, and increase sales. Also,
- information such as “hot items this week” or attractive deals can be displayed together with the associative information in order to promote sales.

f) Telecommunications industry

1. The telecommunication market is rapidly expanding and highly competitive.

2. This creates a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service.

3. The following are a few scenarios for which data mining may improve telecommunication services:

4. Multidimensional analysis of telecommunication data:

- Telecommunication data are intrinsically multidimensional, with dimensions such as calling-time, duration, location of caller, location of callee, and type of call.
- The multidimensional analysis of such data can be used to identify and compare the data traffic, system workload, resource usage, user group behavior, and profit.
- Therefore, it is often useful to consolidate telecommunication data into large data warehouses and routinely perform multidimensional analysis using OLAP and visualization tools.

5. Fraudulent pattern analysis and the identification of unusual patterns:

- It is important to :

  (1) identify potentially fraudulent users and their atypical usage patterns;
  (2) detect attempts to gain fraudulent entry to customer accounts; and
  (3) discover unusual patterns that may need special attention, such as busy-hour frustrated all attempts, switch and route congestion patterns.

- Many of these patterns can be discovered by multidimensional analysis, cluster analysis, and outlier analysis.

6. Multidimensional association and sequential pattern analysis:

- It can be used to promote telecommunication services.
- For example, suppose you would like to find usage patterns for a set of communication services by customer group, by month, and by time of day.

7. Mobile telecommunication services:

- Mobile telecommunication, Web and information services, and mobile computing are becoming increasingly integrated and common in our work and life.

- One important feature of mobile telecommunication data is its association with spatiotemporal information. Spatiotemporal data mining may become essential for finding certain patterns.
- Data mining will likely play a major role in the design of adaptive solutions enabling users to obtain useful information with relatively few keystrokes.

8. Use of visualization tools in telecommunication data analysis:

- Tools for OLAP visualization, linkage visualization, association visualization, clustering, and outlier visualization have been shown to be very useful for telecommunication data analysis.

g) Banking & Finance

1. Most banks and financial institutions offer a wide variety of banking services (such as checking and savings accounts), credit (such as business, mortgage, and automobile loans), and investment services.

2. Financial data collected in the banking and financial industry are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining.

3. Here we present a few typical cases:

Design and construction of data warehouses for multidimensional data analysis and data mining:

- Data warehouses need to be constructed for banking and financial data.
- Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, one may like to view the debt and revenue changes by month, by region.

- Data warehouses, data cubes, multi feature and discovery-driven data cubes, characterization and class comparisons, and outlier analysis all play important roles in financial data analysis and mining.

Loan payment prediction and customer credit policy analysis:

- This analysis is critical to the business of a bank.
- Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones.
- For example, factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio, payment to-income ratio, customer income level, education level, residence region, and credit history. Analysis of the customer payment history may find that, say, payment-to income ratio is a dominant factor, while education level and debt ratio are not.
- The bank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profiles show relatively low risks according to the critical factor analysis.

Classification and clustering of customers for targeted marketing:

- Classification and clustering methods can be used for customer group identification and targeted marketing.
- For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking.
- Customers with similar behaviors regarding loan payments may be identified by multidimensional clustering techniques. These can help identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.

Detection of money laundering and other financial crimes:

- To detect money laundering and other financial crimes, it is important to integrate information from multiple databases, as long as they are potentially related to the study.
- Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers.

- Useful tools include:
- Data visualization tools(to display transaction activities using graphs by time and by groups of customers),
  - linkage analysis tools(to identify links among different customers and activities),
- classification tools (to filter unrelated attributes and rank the highly related ones),
  - clustering tools (to group different cases),
  - outlier analysis tools (to detect unusual amounts of fund transfers or other activities),
  - sequential pattern analysis tools (to characterize unusual access sequences).

- These tools may identify important relationships and patterns of activities and help investigators focus on suspicious cases for further detailed examination.

h) CRM

- Customer Relationship Management(CRM) emerged in the last decade to reflect the central role of the customer for the strategic positioning of a company.
- It encompasses all measures for understanding the customers and for exploiting this knowledge to design and implement marketing activities, align production and coordinate the supply chain.
- CRM puts emphasis on the coordination of such measures, also implying the integration of customer-related data, meta-data and knowledge and the centralized planning and evaluation of measures to increase customer lifetime value.
- CRM gains in importance for companies that serve multiple groups of customers and exploit different interaction channels for them.
- CRM is a broadly used term, and covers a wide variety of functions.
- These functions include:
  - marketing automation (e.g. campaign management, cross and up-sell, customer segmentation, customer retention),
  - sales force automation (e.g. contact management , lead generation , sales analytics, generation of quotes, product configuration) and
  - contact centre management(e.g. call management , integration of multiple contact channels, problem escalation and resolution, metrics and monitoring , logging interactions and auditing), among others.

- Data mining helps marketing professionals improve their understanding of customer behavior.

- In turn, this better understanding allows them to target marketing campaigns more accurately and to align campaigns more closely with needs, wants and attitudes of customers and prospects.

Either can be Starting point Action by Organization
Action by customer.Increase in one caused by increase in another Customer understanding (by organization)Measurement and Evaluation.

Q9. b. Explain Big data Business Analytics in detail

Big Data Business Analytics refers to the use of advanced analytical techniques and tools to extract valuable insights, trends, and patterns from large, complex, and varied datasets. As organizations generate vast amounts of data from sources like social media, customer transactions, sensors, and IoT devices, big data analytics becomes essential for gaining actionable intelligence that can enhance decision-making, improve operations, and create competitive advantages.

Key Characteristics of Big Data (The 5 Vs)

Big Data is typically characterized by the following "5 Vs," which distinguish it from traditional data:

1. **Volume**: Refers to the sheer quantity of data generated from diverse sources, which is often too large to process with traditional database systems.
2. **Velocity**: Describes the speed at which data is generated, captured, and processed. Many business applications require real-time or near-real-time data analytics to stay responsive to dynamic conditions.
3. **Variety**: Big Data includes structured, semi-structured, and unstructured data, such as text, images, audio, video, and sensor data, which requires specialized tools to analyze.
4. **Veracity**: The reliability or accuracy of data is critical, as data from multiple sources may contain inconsistencies or inaccuracies. Managing data veracity is essential for trustworthy analytics.
5. **Value**: The ultimate goal of big data analytics is to derive meaningful insights that deliver business value, helping organizations make informed decisions and capitalize on data-driven opportunities.

Key Components of Big Data Business Analytics

1. **Data Sources**:
   - Big data analytics uses a wide variety of data sources, including:
     - **Transaction Data**: Customer purchases, order records, and financial data.
     - **Social Media Data**: Comments, likes, shares, and sentiments from social platforms.

        - **Sensor Data**: IoT devices and smart sensors produce real-time data on environmental factors, machinery conditions, and more.
        - **Machine Logs**: System logs from applications and servers, often used for monitoring and cybersecurity.
        - **External Data**: Economic indicators, market data, and weather data.
2. **Data Collection and Storage**:
    - **Data Collection**: Big data tools like Apache Flume, Kafka, and NiFi are often used to collect and transport data in real time.
    - **Data Storage**: Big data requires scalable storage solutions to handle massive datasets. Distributed storage systems like Hadoop Distributed File System (HDFS), cloud storage, and data lakes are commonly used to store raw and processed data.
3. **Data Processing and Preparation**:
    - Before data can be analyzed, it must be pre-processed, which includes:
        - **Data Cleaning**: Removing inconsistencies, duplicates, and errors from raw data.
        - **Data Transformation**: Converting data into a compatible format for analysis (e.g., encoding, normalization).
        - **Data Integration**: Merging data from multiple sources to create a comprehensive dataset.
    - Big data frameworks like **Apache Hadoop** and **Apache Spark** support distributed data processing, making it possible to process large volumes of data efficiently.
4. **Data Analysis**:
    - **Descriptive Analytics**: Summarizes historical data to understand past performance, trends, and patterns. Commonly used for reporting and data visualization.
    - **Diagnostic Analytics**: Focuses on understanding the reasons behind certain trends or patterns, using techniques such as data mining, clustering, and association analysis.
    - **Predictive Analytics**: Uses historical data to make predictions about future events. Techniques include machine learning models like regression, time series analysis, and classification.
    - **Prescriptive Analytics**: Suggests actions based on predictive insights. Optimization algorithms and simulations are used to provide recommendations on the best course of action.
    - **Streaming Analytics**: Processes data in real time as it is generated, allowing organizations to act immediately on insights. Technologies like Apache Kafka and Spark Streaming support real-time analytics.
5. **Data Visualization**:
    - Visualization tools like **Tableau**, **Power BI**, **D3.js**, and **Qlik** enable users to interact with data through dashboards, graphs, and charts. Visualizations help users interpret complex data more easily, facilitating data-driven decisions.
    - Real-time visualization is particularly useful for monitoring KPIs, identifying trends, and quickly responding to business changes.

Big Data Analytics Techniques

1. **Machine Learning**:
   - Machine learning algorithms identify patterns and make predictions or decisions. Techniques like supervised learning, unsupervised learning, and reinforcement learning are widely used in predictive and prescriptive analytics.
   - **Example**: Retailers use machine learning to predict customer purchasing behavior based on past interactions and transaction history.
2. **Natural Language Processing (NLP)**:
   - NLP allows computers to interpret and process human language. In business, NLP is often used for sentiment analysis, chatbot applications, and analyzing customer feedback.
   - **Example**: Social media sentiment analysis can reveal public opinion on a new product or campaign.
3. **Data Mining**:
   - Data mining extracts patterns and relationships from large datasets. Techniques include association rule mining, clustering, and classification.
   - **Example**: Market basket analysis in retail can identify products frequently bought together, which helps in product bundling and store layout optimization.
4. **Time Series Analysis**:
   - Time series analysis focuses on data points collected over time. It is particularly useful for forecasting, trend analysis, and anomaly detection.
   - **Example**: Time series forecasting can predict future sales based on historical trends, aiding in inventory planning.
5. **Graph Analytics**:
   - Graph analytics examines relationships between entities, such as in social network analysis and recommendation engines.
   - **Example**: A telecom company can use graph analytics to identify influencers in social networks to optimize marketing efforts.

Applications of Big Data Business Analytics

1. **Customer Insights and Personalization**:
   - Analyzing customer behavior helps companies personalize marketing and improve customer experiences. Big data can segment customers, predict churn, and tailor recommendations.
   - **Example**: E-commerce platforms use recommendation algorithms based on customer browsing and purchase history.
2. **Predictive Maintenance**:
   - Big data analytics can predict equipment failures by analyzing sensor data, helping organizations perform maintenance before breakdowns occur, thus reducing downtime and repair costs.
   - **Example**: Manufacturers use predictive maintenance on machinery, reducing unexpected shutdowns.
3. **Fraud Detection and Risk Management**:

    - Big data analytics can detect unusual patterns that may indicate fraud, such as irregular transaction behavior or login attempts.
    - **Example**: Banks use predictive models to flag potentially fraudulent transactions, preventing unauthorized activities.
4. **Supply Chain Optimization**:
    - Big data analytics improves supply chain visibility, optimizes inventory management, and enables better demand forecasting.
    - **Example**: Retailers use demand forecasting to ensure optimal stock levels, minimizing both shortages and overstock.
5. **Product Development**:
    - Analytics provides insights into customer preferences, helping businesses innovate and develop products that meet market demand.
    - **Example**: Automotive companies use customer feedback and market trends to design vehicles that cater to evolving consumer needs.
6. **Healthcare Analytics**:
    - Big data analytics in healthcare helps in disease prediction, patient outcome optimization, and operational efficiency.
    - **Example**: Hospitals use predictive models to identify high-risk patients and personalize treatment plans.

Challenges of Big Data Business Analytics

1. **Data Privacy and Security**: Ensuring the privacy and security of large volumes of sensitive data is challenging, especially with regulatory requirements like GDPR.
2. **Data Quality**: Big data often includes inconsistent and unstructured data, requiring substantial pre-processing to ensure accuracy and reliability.
3. **Scalability**: Processing and analyzing massive datasets require scalable infrastructure, leading to high storage and computational costs.
4. **Complexity and Skill Requirements**: Effective big data analytics requires specialized skills in data engineering, data science, and domain expertise, which may be challenging to acquire and retain.

Q10. a. Write short notes on four of the following

- **i)** WEKA Tool
- **ii)** Drill down and roll-up operations
- **iii)** Incremental ARM
- **iv)** OLAP and OLTP

i) WEKA Tool:

Weka (Waikato Environment for Knowledge Analysis) is an open-source tool for data mining and machine learning developed by the University of Waikato, New Zealand. It offers a wide range of algorithms for classification, regression, clustering, and association rule mining, as well as tools for data preprocessing, visualization, and evaluation.

Key Features

1. **User-Friendly GUI**: Simplifies data analysis for users with no coding experience.
2. **Extensive Algorithm Library**: Includes popular algorithms like decision trees, SVMs, and neural networks.
3. **Data Preprocessing**: Tools for cleaning, transforming, and normalizing data.
4. **Visualization**: Allows for data and model result visualization, aiding interpretation.
5. **Experimenter**: Compares algorithms across datasets to find optimal models.
6. **Knowledge Flow**: Provides a workflow interface for building analysis pipelines.
7. **Scripting Support**: Enables batch processing and automation via CLI.

Common Tasks

- **Classification**: Train models like Naïve Bayes or decision trees to predict classes.
- **Clustering**: Use K-Means for grouping similar data points.
- **Association Mining**: Discover relationships with Apriori or FP-Growth.
- **Feature Selection**: Identify and keep important attributes.
- **Evaluation**: Use metrics like accuracy and confusion matrices for performance analysis.

Limitations

Weka is primarily for smaller datasets and may struggle with very large data due to memory constraints, lacking native integration with big data frameworks like Hadoop or Spark.

Use Case Example: A retailer can use Weka for sentiment analysis on customer reviews by loading, preprocessing text data, selecting features, classifying with Naïve Bayes, and visualizing results to gauge customer sentiment.

ii) Drill down and roll-up operations

1) Roll up

The roll-up operation (also called drill-up or aggregation operation) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by climbing down a concept hierarchy, i.e. dimension reduction. In the following example given at figure 3, it is shown a multidimensional cube containing the products of a Home appliances home appliances like laptop, furniture, mobile and kitchen appliances. If the manager wants to view the sales of all the products quarterly, the Roll-up operation can be performed on the categories.In this aggregation process, data is category hierarchy moves up from mobile to the Kitchen store. In the roll-up process at least one or more dimensions get reduced like category here.

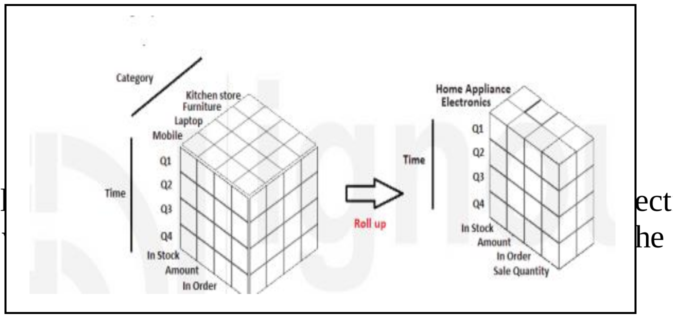


Figure 1: Rol ectronics)
It is also kno he data along the dimension.

2) Drill-down:

The drill down operation (also called roll-down) is the reverse of roll up. It navigates from less detailed data to more detailed data. It can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

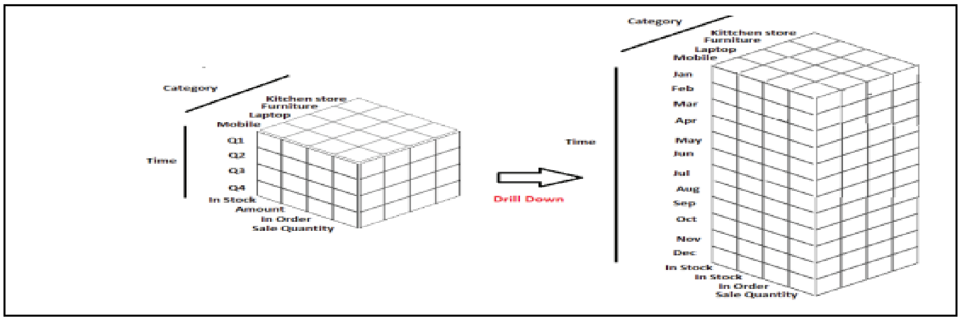


Figure 2: Drill down from Time to Months

You will observe in the above example given at figure 2 a multidimensional cube containing products and time. The Time dimension has been expanded from Quarter →Months to observe the sales month-wise. This is called in Drill down.

iii) Incremental ARM

Incremental Association Rule Mining (Incremental ARM) is a technique used in data mining to update association rules as new data becomes available, without reprocessing the entire dataset. This is essential in environments where data is continuously changing, such as e-commerce transactions, social media interactions, or streaming services.

Key Features of Incremental ARM:

1. **Efficiency**: Instead of recalculating rules from scratch, Incremental ARM updates existing rules based on the new data. This reduces computation time and resource usage.
2. **Adaptability**: Incremental ARM adapts to data changes in real-time, making it useful for dynamic and evolving datasets.
3. **Rule Maintenance**: As new data may invalidate or strengthen existing rules, Incremental ARM helps maintain an accurate set of association rules.
4. **Scalability**: Suitable for large-scale applications where traditional ARM would be too slow or resource-intensive for continuous updates.

Use Cases:

- **Market Basket Analysis**: Continuously updating product associations based on recent purchase patterns.
- **Real-time Recommendations**: Updating user recommendations on streaming or shopping platforms as user behavior changes.

Incremental ARM is valuable for systems that rely on timely and accurate association rules, helping organizations make data-driven decisions without repeated, costly data reprocessing.

iv) OLAP and OLTP

OLAP (Online Analytical Processing) and **OLTP (Online Transaction Processing)** are two distinct systems in data management, each serving different purposes in business operations and analysis.

OLAP (Online Analytical Processing)

- **Purpose**: Primarily used for data analysis and decision-making.
- **Data Type**: Historical and aggregated data, optimized for complex queries.
- **Operations**: Supports operations like slicing, dicing, pivoting, and drill-downs to analyze data from multiple perspectives.
- **Data Structure**: Often uses multi-dimensional data models (data cubes) for better insights.
- **Speed**: Prioritizes query speed for read-heavy operations.
- **Examples**: Business intelligence, data warehousing, and reporting systems.

OLTP (Online Transaction Processing)

- **Purpose**: Designed for handling day-to-day transactions.
- **Data Type**: Real-time, transactional data, such as order processing and inventory management.
- **Operations**: Simple, short queries focused on insert, update, and delete (CRUD operations).
- **Data Structure**: Uses a highly normalized relational database structure to reduce redundancy.
- **Speed**: Optimized for quick transaction processing and maintaining data integrity.
- **Examples**: Banking systems, e-commerce, and customer order management systems.

Key Difference: OLAP is analytical and read-heavy for insights, while OLTP is transactional and write-heavy for managing real-time business operations.