

Semester-5

Data Mining

(According to Purvanchal University Syllabus)

Unit – 1

Introduction

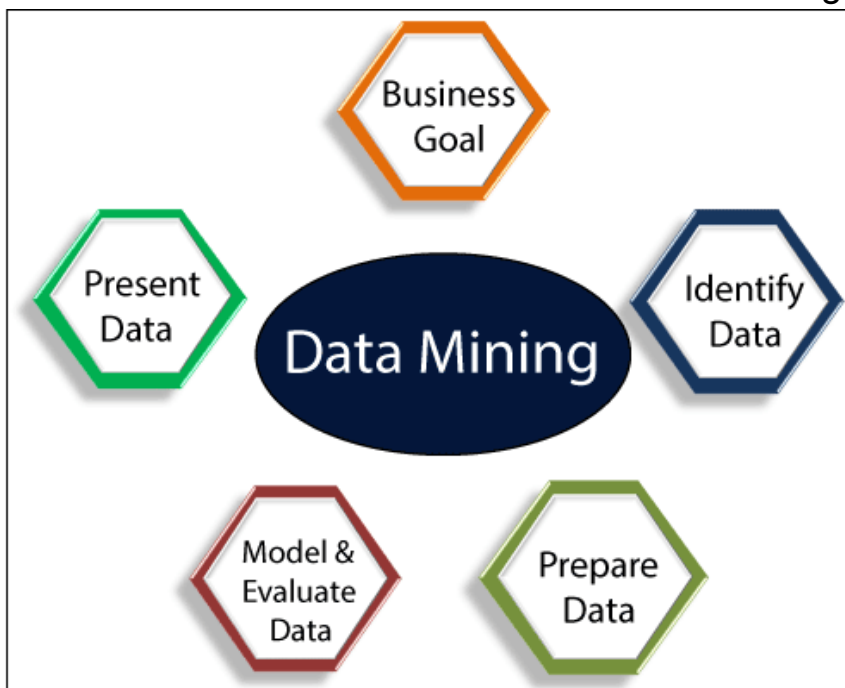
What is Data Mining–

Data Mining is defined as extracting information from huge sets of data.

In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find and find better insight from it.



Types of Data Mining

Data mining can be performed on the following types of data:

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data search ability, reporting, and organization.

Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision-making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed

appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost efficient.
- Data Mining helps the decision-making process of an organization.
- It facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified into two categories: descriptive and predictive.

Descriptive mining tasks characterize the general properties of the data in the database.

Predictive mining tasks perform inference on the current data in order to make predictions.

Concept/Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. For example, in the

Electronics store, classes of items for sale include computers and printers, and concepts of customers include bigSpenders and budgetSpenders.

Data characterization

Data characterization is a summarization of the general characteristics or features of a target class of data.

Data discrimination

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns, are patterns that occur frequently in data. There are many kinds of frequent patterns, including itemsets, subsequences, and substructures.

Association analysis

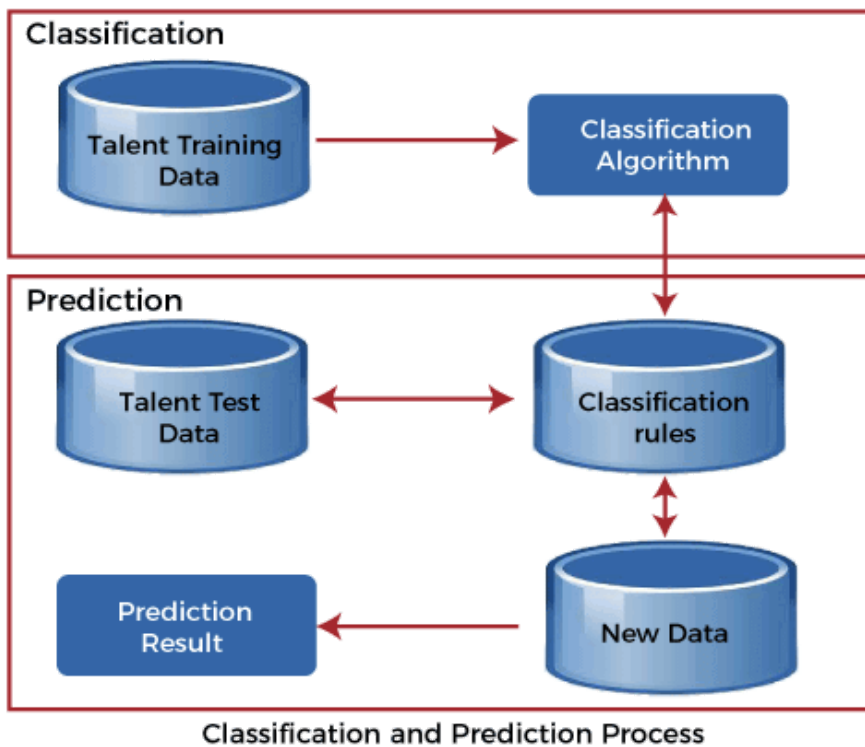
Suppose, as a marketing manager, you would like to determine which items are frequently purchased together within the same transactions.

$\text{buys}(X, \text{"computer"}) = \text{buys}(X, \text{"software"})$ [support=1%, confidence=50%]

where X is a variable representing a customer. Confidence=50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.

Support=1% means that 1% of all of the transactions under analysis showed that computer and software were purchased together.

Classification and Prediction



Classification of Data mining

Following are the examples of cases where the data analysis task is Classification –

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data .and yes or no for marketing data.

How Does Classification Works?

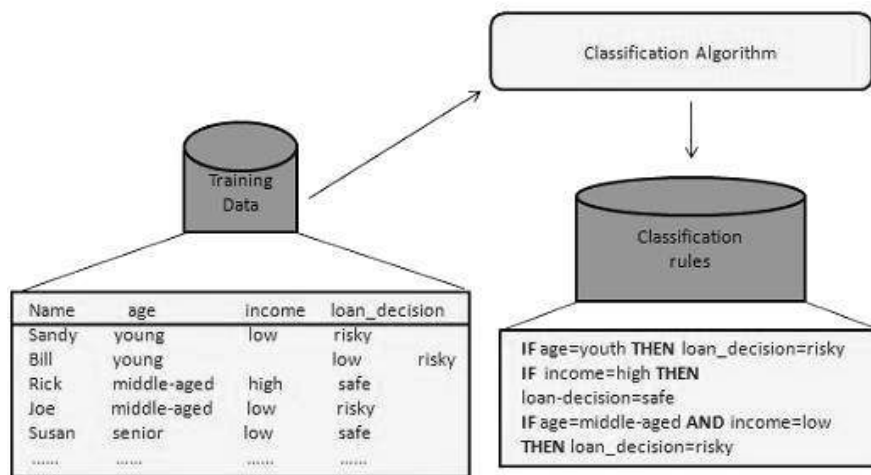
With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification includes two steps –

- Building the Classifier or Model
- Using Classifier for Classification

Building the Classifier or Model

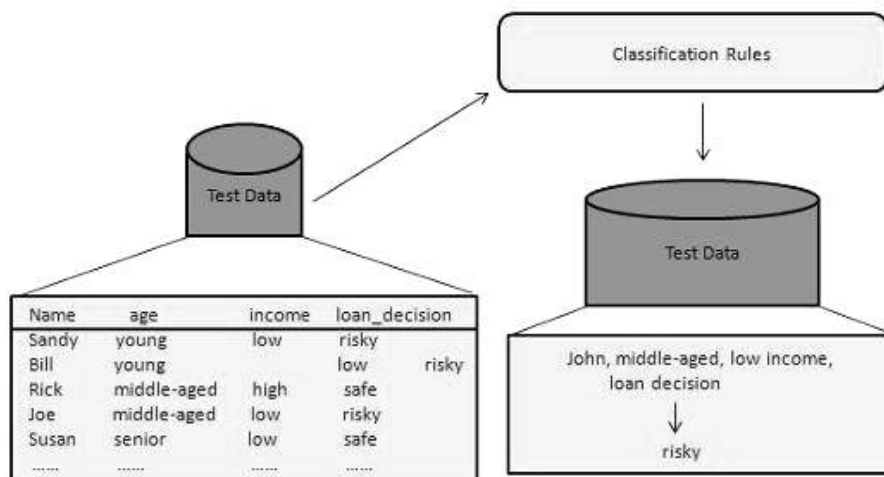
- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a

category or class. These tuples can also be referred to as sample, object or data points.



Using Classifier for Classification

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



Data mining Task

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining –

- Descriptive
- Classification and Prediction

Descriptive Function

The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions –

- Class/Concept Description

- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

Class/Concept Description

Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways –

- **Data Characterization** – This refers to summarizing data of class under study. This class under study is called as Target Class.
- **Data Discrimination** – It refers to the mapping or classification of a class with some predefined group or class.

Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns –

- **Frequent Item Set** – It refers to a set of items that frequently appear together, for example, milk and bread.
- **Frequent Subsequence** – A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- **Frequent Sub Structure** – Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.

For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

Mining of Correlations

It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms –

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

The list of functions involved in these processes are as follows –

- **Classification** – It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
- **Prediction** – It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
- **Outlier Analysis** – Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
- **Evolution Analysis** – Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

Primitives Integration of Data Mining System-

- We can specify a data mining task in the form of a data mining query.
- This query is input to the system.
 - A data mining query is defined in terms of data mining task primitives.

Note – These primitives allow us to communicate in an interactive manner with the data mining system. Here is the list of Data Mining Task Primitives –

- Set of task relevant data to be mined.

- Kind of knowledge to be mined.
- Background knowledge to be used in discovery process.
- Interestingness measures and thresholds for pattern evaluation.
- Representation for visualizing the discovered patterns.

Set of task relevant data to be mined

This is the portion of database in which the user is interested. This portion includes the following –

- Database Attributes
- Data Warehouse dimensions of interest

Kind of knowledge to be mined

It refers to the kind of functions to be performed. These functions are –

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Background knowledge

The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction.

Interestingness measures and thresholds for pattern evaluation

This is used to evaluate the patterns that are discovered by the process of knowledge discovery. There are different interesting measures for different kind of knowledge.

Representation for visualizing the discovered patterns

This refers to the form in which discovered patterns are to be displayed. These representations may include the following. –

- Rules
- Tables
- Charts
- Graphs

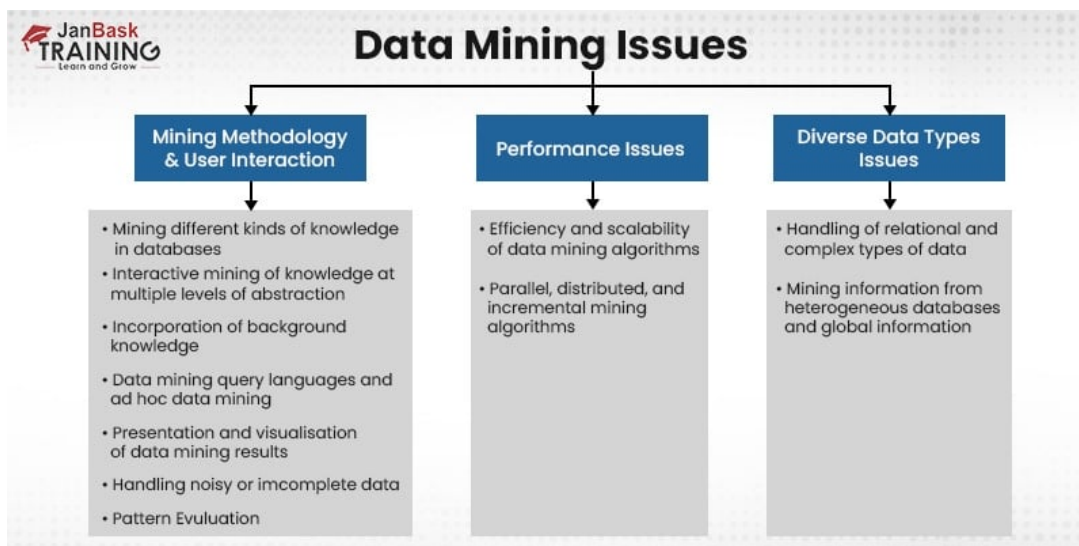
- Decision Trees
- Cubes

Major issues in Data Mining

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding –

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues –

- **Mining different kinds of knowledge in databases** – Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** – The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** – To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** – Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

- **Presentation and visualization of data mining results** – Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** – The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
 - **Pattern evaluation** – The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows –

- **Efficiency and scalability of data mining algorithms** – In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** – The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** – The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** – The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

Unit – 2

Data Processing

Process the Data Descriptive Data

Summarization –

Data preparation or preprocessing is a big issue for both data warehousing and data mining

Discriptive data summarization is need for quality data preprocessing

- **Data preparation includes**

- o Data cleaning and data integration
- o Data reduction and feature selection
- o Discretization
- o A lot a methods have been developed but data preprocessing still an active area of research

Measuring Central Tendency

The central tendency of data, include:

- Mean
- Weighted mean
- Trimmed mean
- Median
- Mode
- Midrange

Mean: The most common and most effective numerical measure of the “center” of a set of data is the (arithmetic) mean.
mean. (sample vs. population)

Mean

Mean of Grouped Data:

$$\bar{x} = \frac{\sum fx}{n}$$

where: \bar{x} = mean
 f = frequency of each class
 x = mid-interval value of each class
 n = total frequency
 $\sum fx$ = sum of the product of mid – interval values and their corresponding frequency

Weighted mean: Sometimes, each value in a set may be associated with a weight, the weights reflect the significance,

Trimmed mean

- A major problem with the mean is its sensitivity to extreme (e.g., outlier) values.
- Even a small number of extreme values can corrupt the mean.
- the trimmed mean is the mean obtained after cutting off Descriptive Data Summarization values at the high and low extremes.
- For example, we can sort the values and remove the top and bottom 2% before computing the mean.
- We should avoid trimming too large a portion (such as 20%) at both ends as this can result in the loss of valuable information.

Median:

Suppose that a given data set of N distinct values is sorted in numerical order. The median is the middle value if odd number of values, or average of the middle two values otherwise

For skewed (asymmetric) data, a better measure of the center of data is the median.

Mode and Midrange:

- Mode is the another measure of central tendency – The mode for a set of data is the value that occurs most frequently in the set.
- If each data value occurs only once, then there is no mode.
- The midrange can also be used to assess the central tendency of a data set
- It is the average of the largest and smallest values in the set.

Dispersion of Data

The degree to which numerical data tend to spread is called the dispersion, or variance of the data.

The most common measures of data dispersion are:

- Range
- Five-number summary (based on quartiles)
- Interquartile range (IQR)
- Standard deviation

Range: difference between highest and lowest observed values.

Inter-quartile range (IQR): $IQR = Q3 - Q1$ – IQR is a simple measure of spread that gives the range covered by the middle half of

the data.

Five number summary: min, Q 1, Median, Q 3, max – Contain information about the endpoints (e.g., tails) of the data.

Standard deviation: s (or σ) is the square root of variance s^2 (or σ^2)

– σ measures spread about the mean and should be used only when the mean is chosen as the measure of center.

– $\sigma = 0$ only when there is no spread, that is, when all observations have the same value.

Graphic Displays of – Basic Descriptive Data Summaries

There are many types of graphs for the display of data summaries and distributions, such as:

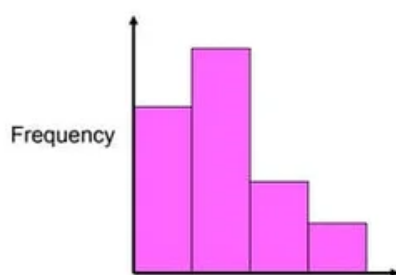
- Bar charts
- Pie charts
- Line graphs
- Boxplot
- Histograms
- Quantile plots
- Scatter plots
- Loess curves

Histogram:

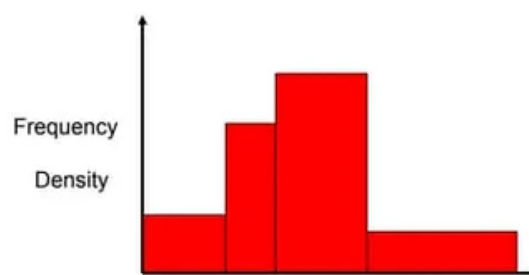
Histograms

Histograms look like frequency diagrams.

The difference is that you get the frequencies from the area of the rectangles and not the height.



Frequency
Diagram

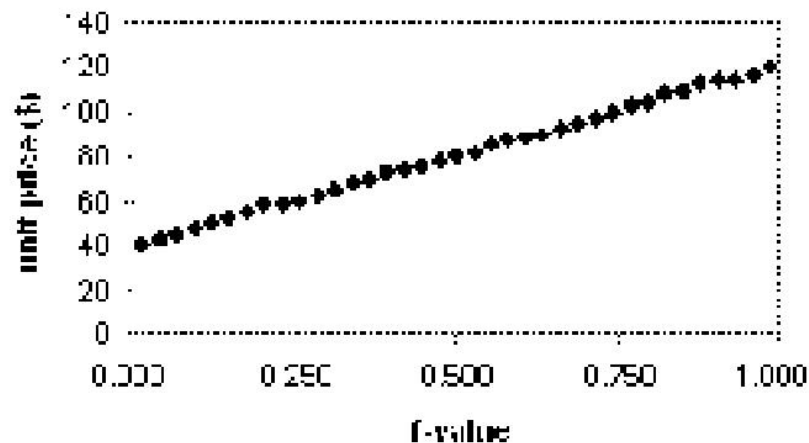


Histogram

Quantile plots:

Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
 - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 $f_i\%$ of the data are below or equal to the value x_i

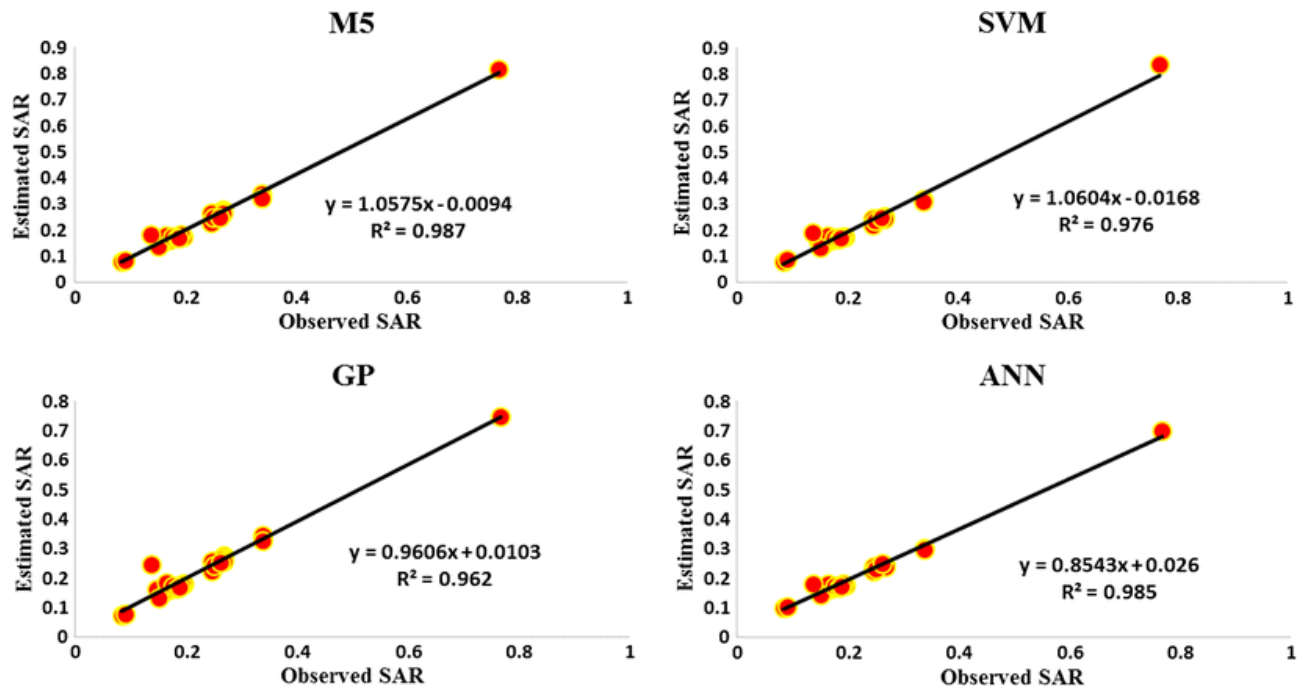


18

Scatter plots:

A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, clusters of points, or outliers between two numerical attributes.

Each pair of values is treated as a pair of coordinates and plotted as points in the plane.



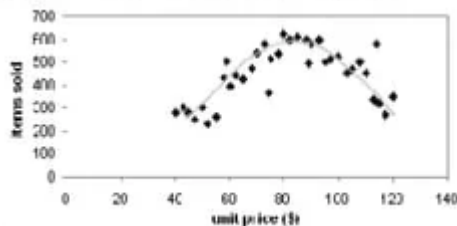
Loess curves:

Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence.

The word loess is short for local regression.

Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



Data Cleaning-

● Importance

- o “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
- o “Data cleaning is the number one problem in data warehousing”—DCI survey Data cleaning tasks

● Fill in missing values

- o Identify outliers and smooth out noisy data
- o Correct inconsistent data
- o Resolve redundancy caused by data integration

Data Integration and Transformation

Data integration:

- Combines data from multiple sources into a coherent store.

Schema integration: e.g., A.cust -id ° B.cust-#

- Integrate metadata from different sources
-

Entity identification problem:

- Identify real world entities from multiple data sources, e.g., e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
 - o For the same real world entity, attribute values from sources are different.
 - o Possible reasons: different representations, different scales, e.g., metric vs. British units metric vs. British units

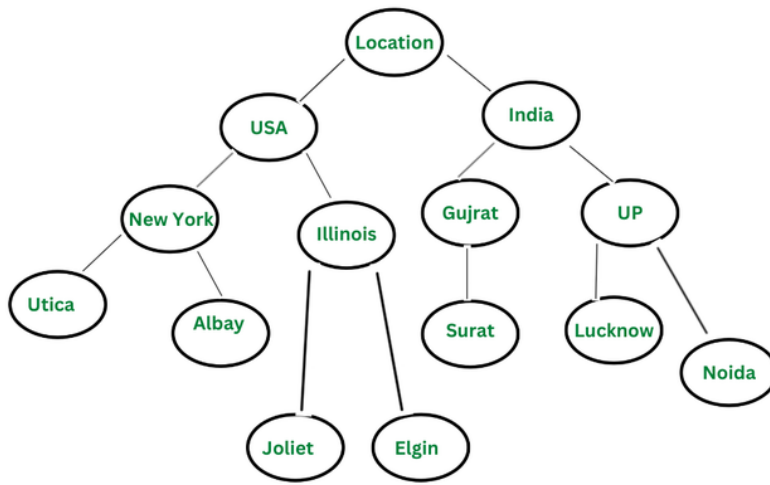
Data Reduction-

Data Reduction Strategies

- **Need for data reduction**
 - A database/data warehouse may store terabytes of data
 - Complex data analysis/mining may take a very long time to run on the complete data set
- **Data reduction**
 - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- **Data reduction strategies**
 - Data cube aggregation
 - Attribute Subset Selection
 - Numerosity reduction — e.g., fit data into models
 - Dimensionality reduction - Data Compression
 - Discretization and concept hierarchy generation

2

Data Discrimination - Concept Hierarchy Generation



Concept Hierarchy for Dimension Location

Unit – 3

Data Warehouse

Data Warehouse is a relational database management system (RDBMS) construct to meet the requirement of transaction processing systems. It can be loosely described as any centralized data repository which can be queried for business benefits. It is a database that stores information oriented to satisfy decision-making requests.

Understanding a Data Warehouse

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.
- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.

OLAP –

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information. This chapter cover the types of OLAP, operations on OLAP, difference between OLAP, and statistical databases and OLTP.

Types of OLAP Servers

We have four types of OLAP servers –

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

Relational OLAP

ROLAP servers are placed between relational back-end server and client front end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following –

- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

Specialized SQL Servers

Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

Data Warehouse Multidimensional Data Model –

Multidimensional OLAP (MOLAP) uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the dataset is sparse. Therefore, many MOLAP servers use two levels of data storage representation to handle dense and sparse datasets.

Points to Remember –

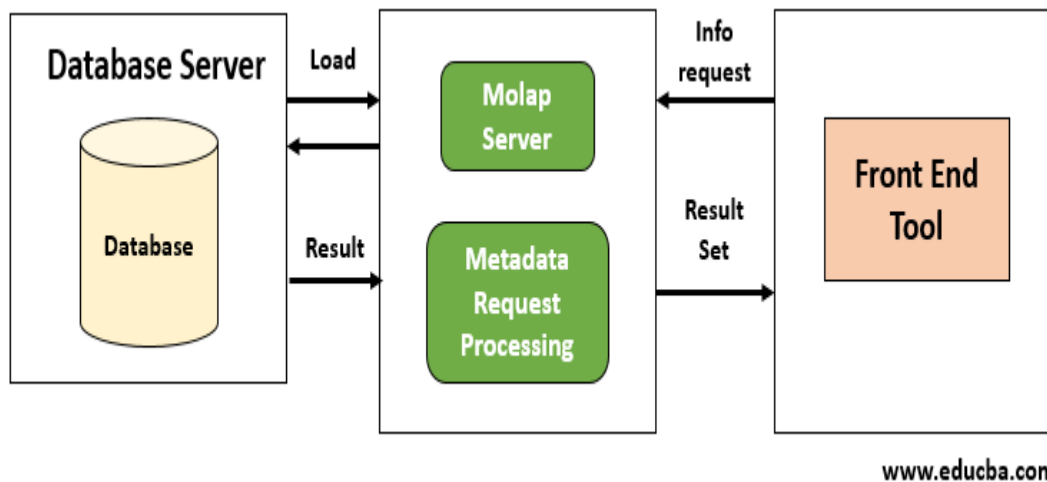
- MOLAP tools process information with consistent response time regardless of level of summarizing or calculations selected.
- MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis.
- MOLAP tools need fastest possible performance.
- MOLAP server adopts two level of storage representation to handle dense and sparse data sets.

- Denser sub-cubes are identified and stored as array structure.
- Sparse sub-cubes employ compression technology.

MOLAP Architecture

MOLAP includes the following components –

- Database server.
- MOLAP server.
- Front-end tool.



Advantages



- MOLAP allows fastest indexing to the pre-computed summarized data.
- Helps the users connected to a network who need to analyze larger, less-defined data.
- Easier to use, therefore MOLAP is suitable for inexperienced users.

Disadvantages

- MOLAP are not capable of containing detailed data.
- The storage utilization may be low if the data set is sparse.

MOLAP vs ROLAP

Comparing ROLAP & MOLAP

• MOLAP 	• ROLAP 
<ul style="list-style-type: none"> • Data Stored as Aggregates • Data is Hierarchical • Data is current only up to recent update • Pre-defined queries, functions • Data per transaction : Large 	<ul style="list-style-type: none"> • Data Stored as Records • Data is Relational • Data is current • Supports ad-hoc queries • Data per transaction : Smaller

Data Warehouse Architecture

A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise. Each data warehouse is different, but all are characterized by standard vital components.

Production applications such as payroll accounts payable product purchasing and inventory control are designed for online transaction processing (OLTP). Such applications gather detailed data from day to day operations.

Data Warehouse applications are designed to support the user ad-hoc data requirements, an activity recently dubbed online analytical processing (OLAP). These include applications such as forecasting, profiling, summary reporting, and trend analysis.

Data warehouses and their architectures vary depending upon the elements of an organization's situation.

Three common architectures are:

- Data Warehouse Architecture: Basic
- Data Warehouse Architecture: With Staging Area
- Data Warehouse Architecture: With Staging Area and Data

Data Warehouse Architecture: Basic

Operational System

An operational system is a method used in data warehousing to refer to a system that is used to process the day-to-day transactions of an organization.

Flat Files

A Flat file system is a system of files in which transactional data is stored, and every file in the system must have a different name.

Meta Data

A set of data that defines and gives information about other data.

Meta Data used in Data Warehouse for a variety of purpose, including:

Meta Data summarizes necessary information about data, which can make finding and work with particular instances of data more accessible. For example, author, data build, and data changed, and file size are examples of very basic document metadata.

Metadata is used to direct a query to the most appropriate data source.

Lightly and highly summarized data

The area of the data warehouse saves all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.

The goals of the summarized information are to speed up query performance. The summarized record is updated continuously as new information is loaded into the warehouse.

End-User access Tools

The principal purpose of a data warehouse is to provide information to the business managers for strategic decision-making. These customers interact with the warehouse using end-client access tools.

The examples of some of the end-user access tools can be:

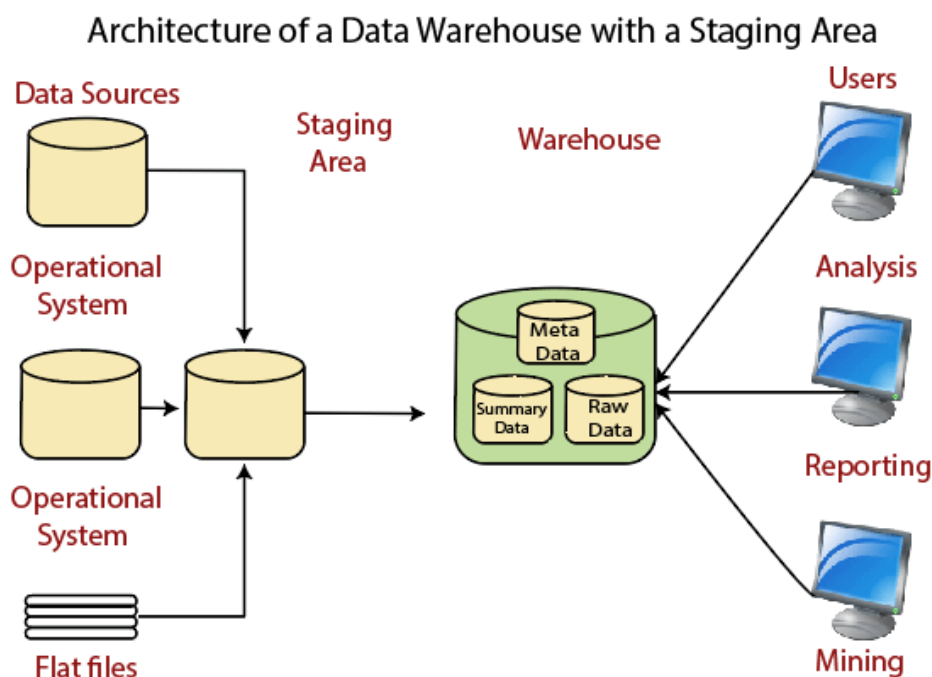
- Reporting and Query Tools
- Application Development Tools
- Executive Information Systems Tools
- Online Analytical Processing Tools
- Data Mining Tools

Data Warehouse Architecture: With Staging Area

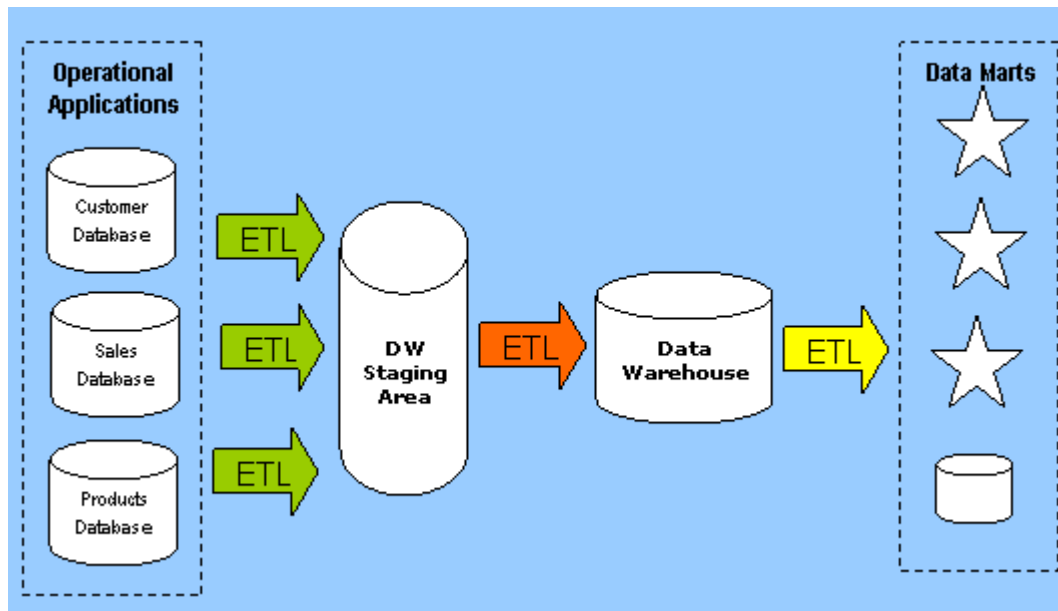
We must clean and process your operational information before put it into the warehouse.

We can do this programmatically, although data warehouses uses a staging area (A place where data is processed before entering the warehouse).

A staging area simplifies data cleansing and consolidation for operational method coming from multiple source systems, especially for enterprise data warehouses where all relevant data of an enterprise is consolidated.



Data Warehouse Staging Area is a temporary location where a record from source systems is copied.



Data Warehouse Architecture: With Staging Area and Data Marts

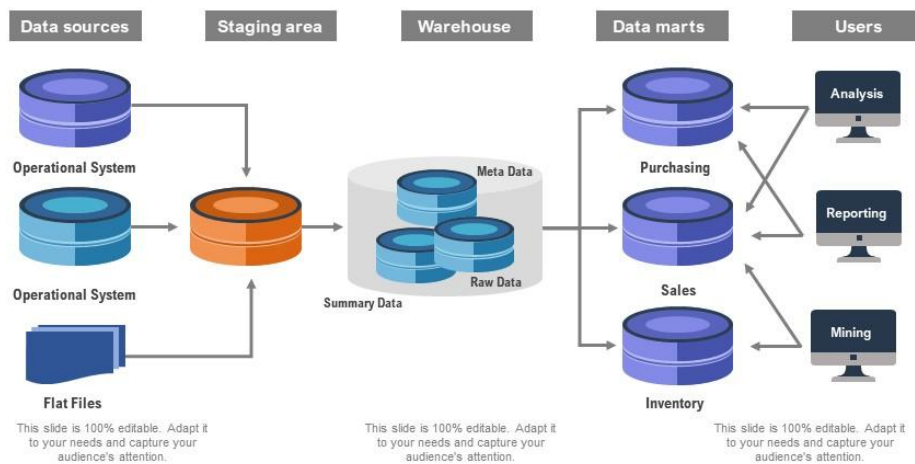
We may want to customize our warehouse's architecture for multiple groups within our organization.

We can do this by adding data marts. A data mart is a segment of a data warehouses that can provided information for reporting and analysis on a section, unit, department or operation in the company, e.g., sales, payroll, production, etc.

The figure illustrates an example where purchasing, sales, and stocks are separated. In this example, a financial analyst wants to analyze historical data for purchases and sales or mine historical information to make predictions about customer behavior.

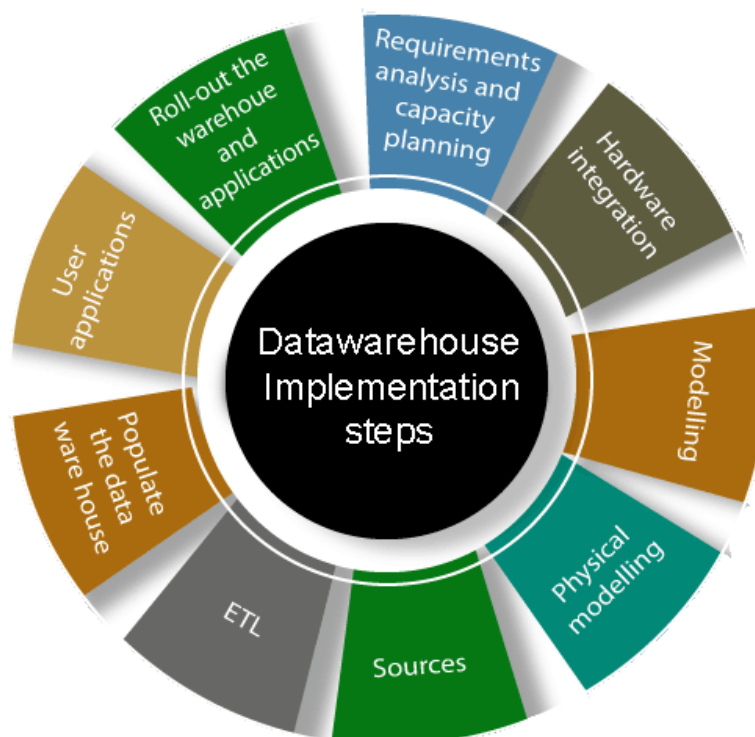
Data Warehouse Model with Staging Area and End Users

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.



Data Warehouse Implementation

There are various implementations in data warehouses which are as follows:



1. Requirements analysis and capacity planning: The first process in data warehousing involves defining enterprise needs, defining architectures, carrying out capacity planning, and selecting the hardware and software tools. This step will contain be consulting senior management as well as the different stakeholder.

2. Hardware integration: Once the hardware and software has been selected, they require to be put by integrating the servers, the storage methods, and the user software tools.

3. Modeling: Modelling is a significant stage that involves designing the warehouse schema and views. This may contain using a modeling tool if the data warehouses are sophisticated.

4. Physical modeling: For the data warehouses to perform efficiently, physical modeling is needed. This contains designing the physical data warehouse organization, data placement, data partitioning, deciding on access techniques, and indexing.

5. Sources: The information for the data warehouse is likely to come from several data sources. This step contains identifying and connecting the sources using the gateway, ODBC drives, or another wrapper.

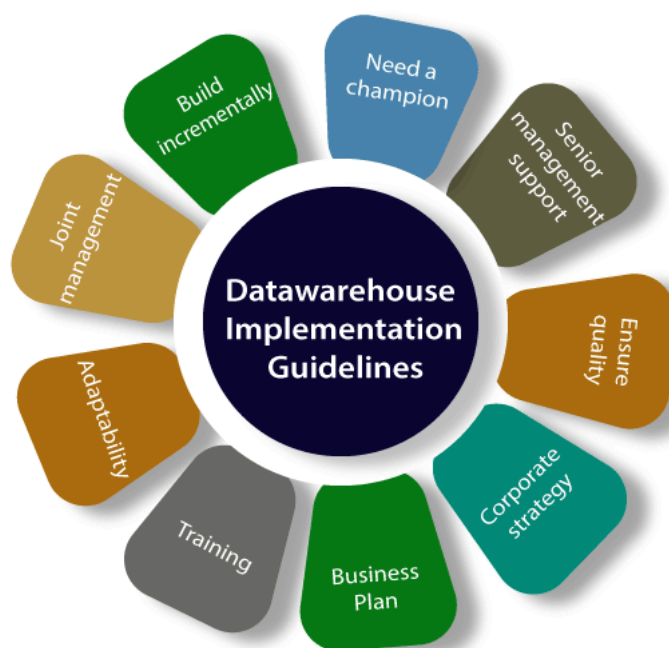
6. ETL: The data from the source system will require to go through an ETL phase. The process of designing and implementing the ETL phase may contain defining a suitable ETL tool vendors and purchasing and implementing the tools. This may contains customize the tool to suit the need of the enterprises.

7. Populate the data warehouses: Once the ETL tools have been agreed upon, testing the tools will be needed, perhaps using a staging area. Once everything is working adequately, the ETL tools may be used in populating the warehouses given the schema and view definition.

8. User applications: For the data warehouses to be helpful, there must be end-user applications. This step contains designing and implementing applications required by the end-users.

9. Roll-out the warehouses and applications: Once the data warehouse has been populated and the end-client applications tested, the warehouse system and the operations may be rolled out for the user's community to use.

Implementation Guidelines



1. Build incrementally: Data warehouses must be built incrementally. Generally, it is recommended that a data marts may be created with one particular project in mind, and once it is implemented, several other sections of the enterprise may also want to implement similar systems. An enterprise data warehouses can then be implemented in an iterative manner allowing all data marts to extract information from the data warehouse.

2. Need a champion: A data warehouses project must have a champion who is active to carry out considerable researches into expected price and benefit of the project. Data warehousing projects requires inputs from many units in an enterprise and therefore needs to be driven by someone who is needed for interacting with people in the enterprises and can actively persuade colleagues.

3. Senior management support: A data warehouses project must be fully supported by senior management. Given the resource-intensive feature of such project and the time they can take to implement, a warehouse project signal for a sustained commitment from senior management.

4. Ensure quality: The only record that has been cleaned and is of a quality that is implicit by the organizations should be loaded in the data warehouses.

5. Corporate strategy: A data warehouse project must be suitable for corporate strategies and business goals. The purpose of the project must be defined before the beginning of the projects.

6. Business plan: The financial costs (hardware, software, and peopleware), expected advantage, and a project plan for a data warehouses project must be clearly outlined and understood by all stakeholders. Without such understanding, rumors about expenditure and benefits can become the only sources of data, subversion the projects.

7. Training: Data warehouses projects must not overlook data warehouses training requirements. For a data warehouses project to be successful, the customers must be trained to use the warehouses and to understand its capabilities.

8. Adaptability: The project should build in flexibility so that changes may be made to the data warehouses if and when required. Like any system, a data warehouse will require to change, as the needs of an enterprise change.

9. Joint management: The project must be handled by both IT and business professionals in the enterprise. To ensure that proper communication with the stakeholder and which the project is the target for assisting the enterprise's business, the business professional must be involved in the project along with technical professionals.

Unit – 4

Mining Frequent Patterns Associations Correlations

Frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern.

Basic Concepts

Frequent Pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a dataset

Goal: finding inherent regularities in data

- What products were often purchased together?— Beer and diapers?!
- What are the subsequent purchases after buying a PC?
- What kinds of DNA are sensitive to this new drug?
- Can we automatically classify Web documents

Applications: Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Basic Concepts: Frequent Patterns

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- **itemset:** A set of one or more items
- **k-itemset** $X = \{x_1, \dots, x_k\}$
- **(absolute) support**, or, **support count** of X : Frequency or occurrence of an itemset X
- **(relative) support**, s , is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is **frequent** if X 's support is no less than a **minsup** threshold

Basic Concepts: Association Rules

id	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Find all the rules $X \rightarrow Y$ with minimum support and confidence

- support, s** , probability that a transaction contains $X \cup Y$
- confidence, c** , conditional probability that a transaction having X also contains Y

Let $\text{minsup} = 50\%$, $\text{minconf} = 50\%$

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
 - $\text{Beer} \rightarrow \text{Diaper}$ (60%, 100%)
 - $\text{Diaper} \rightarrow \text{Beer}$ (60%, 75%)

Frequent Itemsets, Closed Itemsets, and Association Rules

A set of items is referred to as an **itemset**.

An itemset that contains k items is a **k-itemset**.

The set {computer, antivirus software} is a **2-itemset**.

The occurrence frequency of an itemset is the number of transactions that contain the itemset.

This is also known, simply, as the frequency, support count, or count of the itemset.

Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min conf) are called Strong Association Rules.

In general, association rule mining can be viewed as a two-step process:

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min_sup .
2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

Association Rule: Basic Concepts

Given: (1) database of transactions, (2) each transaction is a list of items (purchased by a customer in a visit)

Find: all rules that correlate the presence of one set of items with that of another set of items E.g., 98% of people who purchase tires and auto accessories also get automotive services done

Applications

Maintenance Agreement (What the store should do to boost Maintenance Agreement sales) Home Electronics * (What other products should the store stocks up?) Attached mailing in direct marketing Detecting —ping-pongling of patients, faulty —collisions

Rule Measures: Support and Confidence

Find all the rules $X \rightarrow Y \rightarrow Z$ with minimum confidence and support
Support, s , probability that a transaction contains $\{X \cup Y \cup Z\}$

Confidence, c , conditional probability that a transaction having $\{X \cup Y\}$ also contains Z Let minimum support 50%, and minimum confidence 50%, we have

$A \rightarrow C$ (50%, 66.6%)

$C \rightarrow A$ (50%, 100%)

Association Rule Mining: A Road Map

Boolean vs. quantitative associations (Based on the types of values handled)

$\text{buys}(x, \text{—SQLServerII}) \wedge \text{buys}(x, \text{—DMBookII}) \rightarrow \text{buys}(x, \text{—DBMinerII})$
[0.2%, 60%] \vee $\text{age}(x, \text{—30..39II}) \wedge \text{income}(x, \text{—42..48KII}) \rightarrow \text{buys}(x, \text{—PCII})$ [1%, 75%]

Single dimension vs. multiple dimensional associations (see ex.

Above) \rightarrow Single level vs. multiple-level analysis

What brands of beers are associated with what brands of
diapers? \rightarrow Various extensions

Correlation, causality analysis

Association does not necessarily imply correlation or causality \vee
Maxpatterns and closed itemsets

Unit – 4

Application Trends

Data mining Applications

Here is the list of areas where data mining is widely used –

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

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows –

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry –

- Design and Construction of data warehouses based on the benefits of data mining.

- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services –

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis –

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications –

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection –

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Data Mining System Products-

- There are many data mining system products and domain specific data mining applications. The new data mining systems and applications are being added to the previous systems. Also, efforts are being made to standardize data mining languages.

Choosing a Data Mining System

The selection of a data mining system depends on the following features –

- **Data Types** – The data mining system may handle formatted text, record-based data, and relational data. The data could also be in ASCII text, relational database data or data warehouse data. Therefore, we should check what exact format the data mining

system can handle.

- **System Issues** – We must consider the compatibility of a data mining system with different operating systems. One data mining system may run on only one operating system or on several. There are also data mining systems that provide web-based user interfaces and allow XML data as input.
- **Data Sources** – Data sources refer to the data formats in which data mining system will operate. Some data mining system may work only on ASCII text files while others on multiple relational sources. Data mining system should also support ODBC connections or OLE DB for ODBC connections.
- **Data Mining functions and methodologies** – There are some data mining systems that provide only one data mining function such as classification while some provides multiple data mining functions such as concept description, discovery-driven OLAP analysis, association mining, linkage analysis, statistical analysis, classification, prediction, clustering, outlier analysis, similarity search, etc.
- **Coupling data mining with databases or data warehouse systems** – Data mining systems need to be coupled with a database or a data warehouse system. The coupled components are integrated into a uniform information processing environment. Here are the types of coupling listed below –
 - No coupling
 - Loose Coupling
 - Semi tight Coupling
 - Tight Coupling
- **Scalability** – There are two scalability issues in data mining –
 - **Row (Database size) Scalability** – A data mining system is considered as row scalable when the number of rows are enlarged 10 times. It takes no more than 10 times to execute a query.
 - **Column (Dimension) Scalability** – A data mining system is considered as column scalable if the mining query execution time increases linearly with the number of columns.
- **Visualization Tools** – Visualization in data mining can be categorized as follows –
 - Data Visualization
 - Mining Results Visualization
 - Mining process visualization
 - Visual data mining
- **Data Mining query language and graphical user interface** – An

easy-to use graphical user interface is important to promote user-guided, interactive data mining. Unlike relational database systems, data mining systems do not share underlying data mining query language.

Data Mining Themes

The theoretical foundations of data mining includes the following concepts –

- **Data Reduction** – The basic idea of this theory is to reduce the data representation which trades accuracy for speed in response to the need to obtain quick approximate answers to queries on very large databases. Some of the data reduction techniques are as follows –
 - Singular value Decomposition
 - Wavelets
 - Regression
 - Log-linear models
 - Histograms
 - Clustering
 - Sampling
 - Construction of Index Trees
- **Data Compression** – The basic idea of this theory is to compress the given data by encoding in terms of the following –
 - Bits
 - Association Rules
 - Decision Trees
 - Clusters
- **Pattern Discovery** – The basic idea of this theory is to discover patterns occurring in a database. Following are the areas that contribute to this theory –
 - Machine Learning
 - Neural Network
 - Association Mining
 - Sequential Pattern Matching
 - Clustering
- **Probability Theory** – This theory is based on statistical theory. The basic idea behind this theory is to discover joint probability distributions of random variables.

- **Probability Theory** – According to this theory, data mining finds the patterns that are interesting only to the extent that they can be used in the decision-making process of some enterprise.
- **Microeconomic View** – As per this theory, a database schema consists of data and patterns that are stored in a database. Therefore, data mining is the task of performing induction on databases.
- **Inductive databases** – Apart from the database-oriented techniques, there are statistical techniques available for data analysis. These techniques can be applied to scientific data and data from economic and social sciences as well.

Trends in Data Mining

Data mining concepts are still evolving and here are the latest trends that we get to see in this field –

- Application Exploration.
- Scalable and interactive data mining methods.
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language.
- Visual data mining.
- New methods for mining complex types of data.
- Biological data mining.
- Data mining and software engineering.
- Web mining.
- Distributed data mining.
- Real time data mining.
- Multi database data mining.
- Privacy protection and information security in data mining.