MODULE - I

Data mining applications:

1. Classification: Eg: In loan database, to classify an applicant as a prospective or defaulter, given his various personal and demographic features along with previous purchase characteristics.
2. Estimation: Predict the attribute of a data instance. Eg: estimate the percentage of marks of a student, whose previous marks are already known.
3. Prediction: Predictive model predicts a future outcome rather than the current behaviour. Eg: Predict next week's closing price for the Google share price per unit.
4. Market basket analysis(association rule mining)

   Analyses hidden rules called association rule in a large transactional database.

   {pen, pencil-> book} – whenever pen and pencil are purchased together, book is also purchased.

5. Clustering Classification into different classes based on some similarities but the target classes are unknown.
6. Business intelligence
7. Business data analytics
8. Bioinformatics
9. Web mining
10. Text mining
11. Social network data analysis

Data mining Functionalities

1. Class/Concept Description: Characterization and Discrimination: Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query. For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database

.

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries. For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.

2.Mining Frequent Patterns, Associations, and Correlations, Frequent patterns, as the name suggests, are patterns that occur frequently in data.There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures. A *frequent itemset* typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers. A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a *(frequent) sequential pattern.* A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a *(frequent) structured pattern.* Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Association analysis:

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

A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% **support** means that 1% of all the transactions under analysis show that computer and software are purchased together. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as "*computer) software* [1%, 50%]."
Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule.**

3. Classification and Regression for Predictive Analysis:

Classification is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the the class label is unknown.

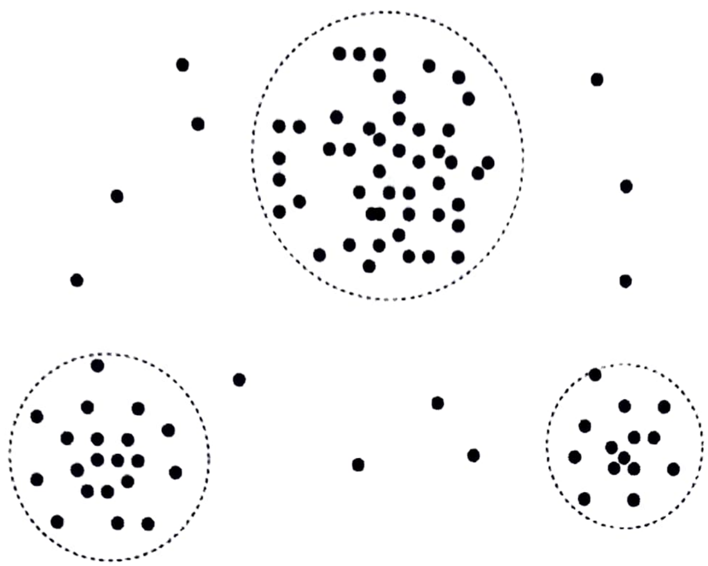
A **decision tree** is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network. A **neural network**,when used for classification, is typically a collection of neuron-like processing unitswithweighted connections between the units. There are many other methods for constructing classificationmodels, such as na¨ıve Bayesian classification, support vector machines, and *k*-nearest-neighbor classification. Whereas classification predicts categorical (discrete, unordered) labels, **regression** models continuous-valued functions. That is, regression is used to predict missing or unavailable *numerical data values* rather than (discrete) class labels. The term *prediction* refers to both numeric prediction and class label prediction.

Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution *trends* based on the available data.

Unlike classification and regression, which analyze class-labeled (training) data sets,

4. Clustering analyzes data objects without consulting class labels. In many cases, classlabeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity.* That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate **taxonomy formation**, that is, the organization of observations into a hierarchy of classes that group similar events together.

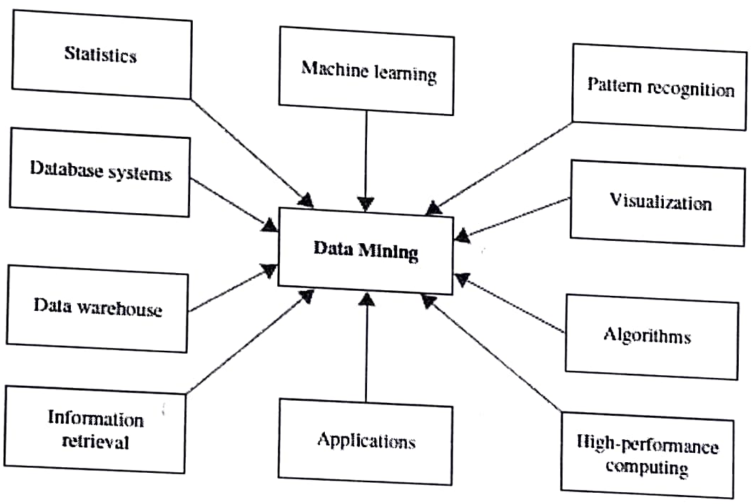


6. Outlier Analysis
A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are **outliers**. Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as **outlier analysis** or **anomaly mining.**
Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.
Are All Patterns Interesting?
A data mining system has the potential to generate thousands or even millions of patterns, or rules.
A pattern is **interesting** if it is (1) *easily understood* by humans, (2) *valid* on new or test data with some degree of *certainty*, (3) potentially *useful*, and (4) *novel.* A pattern is also interesting if it validates a hypothesis that the user *sought to confirm.* An interesting pattern represents **knowledge.**
Several **objective measures of pattern interestingness** exist. These are based on the structure of discovered patterns and the statistics underlying them. An objective

measure for association rules of the form $X \Rightarrow Y$ is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies. This is taken to be the probability $P(X \cup Y)$, where $X \cup Y$ indicates that a transaction contains both X and Y, that is, the union of itemsets X and Y. Another objective measure for association rules is confidence, which assesses the degree of certainty of the detected association. This is taken to be the conditional probability $P(Y|X)$, that is, the probability that a transaction containing X also contains Y. More formally, support and confidence are defined as

$$support(X \Rightarrow Y) = P(X \cup Y),$$
$$confidence(X \Rightarrow Y) = P(Y|X).$$

Technologies for data mining



Statistics:

Statistics studies the collection, analysis, interpretation or explanation, and presentationof data. Data mining has an inherent connection with statistics. A **statistical model** is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes.

In other words, such statistical models can be the outcome of a data mining task. Alternatively, data mining tasks can be built on top of statistical models. For example, we can use statistics to model noise and missing data values. Then, when mining patterns in a large data set, the data mining process can use the model to help identify and handle noisy or missing values in the data. Statistics research develops tools for prediction and forecasting using data and statistical models. Statistical methods can be used to summarize or describe a collection of data.

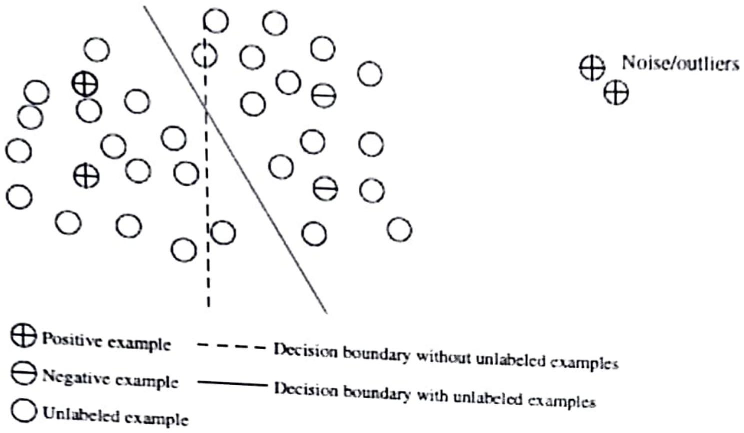
Statistical methods can also be used to verify data mining results. For example, after a classification or prediction model is mined, the model should be verified by statistical hypothesis testing. A **statistical hypothesis test** (sometimes called *confirmatory data analysis*) makes statistical decisions using experimental data. A result is called *statistically significant* if it is unlikely to have occurred by chance. If the classification or prediction model holds true, then the descriptive statistics of the model increases the soundness of the model.

Machine learning: investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to *automatically* learn to recognize complex patterns and make intelligent decisions based on data. For example, a typical machine learning problem is to program a computer so that it can automatically recognize handwritten postal codes on mail after learning from a set of examples. Machine learning is a fast-growing discipline. Here, we illustrate classic problems in machine learning that are highly related to data mining.

Supervised learning is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.

Unsupervised learning is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labeled. Typically, we may use clustering to discover classes within the data. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits. Suppose that it finds 10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to 9, respectively. However, since the training data are not labeled, the learned model cannot tell us the semantic meaning of the clusters found.

Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model. In one approach, labeled examples are used to learn class models and unlabeled examples are used to refine the boundaries between classes. For a two-class problem, we can think of the set of examples belonging to one class as the *positive examples* and those belonging to the other class as the *negative examples*. In Figure if we do not consider the unlabeled examples, the dashed line is the decision boundary that best partitions the positive examples from the negative examples. Using the unlabeled examples, we can refine the decision boundary to the solid line. Moreover, we can detect that the two positive examples at the top right corner, though labeled, are likely noise or outliers.



Active learning is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an example, which may be from a set of unlabeled examples or synthesized by the learning program. The goal is to optimize the model quality by actively acquiring knowledge from human users, given a constraint on how many examples they can be asked to label.

Major issues in data mining

1. Mining methodology and user interaction issues
   - *Mining different kinds of knowledge in databases:*

Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery task. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

- *Interactive mining of knowledge at multiple levels of abstraction:*

Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

- *Incorporation of background knowledge:*

Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction

2. Mining methodology and user interaction issues
   - *Pattern evaluation—the interestingness problem:*

A data mining system can uncover thousands of patterns. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

3. Performance issues
    - *Efficiency and scalability of data mining algorithms:*

To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. The running time of a data mining algorithm must be predictable and acceptable in large databases.

- *Parallel, distributed, and incremental mining algorithms:*

The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

4. Issues relating to the diversity of database types:
    - *Handling of relational and complex types of data:*

Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. Specific data mining systems should be constructed for mining specific kinds of data.

- *Mining information from heterogeneous databases and global information systems:*

Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining

Data Warehouse: Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse refers to a database that is maintained separately from an organization's operational databases. A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process"

1. **Subject-oriented**: A data warehouse is organized around major subjects, such as customer, supplier, product, and sales.

A data warehouse focuses on the modelling and analysis of data for decision makers(not on day to day transaction).

Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

2. **Integrated**: data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.
3. **Time-variant**: Data are stored to provide information from a historical perspective
   Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
4. **Non-volatile**: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency

control mechanisms. It usually requires only two operations in data accessing: *initial loading of data and access of data.*

<u>Data warehousing</u> is the process of constructing and using data warehouses.

- The construction of a data warehouse requires data cleaning, data integration, and data consolidation.
- The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows "knowledge workers" (e.g., managers, analysts, and executives) to use the warehouse to quickly and conveniently obtain an overview of the data, and to make sound decisions based on information in the warehouse.

Data warehousing is very useful from the point of view of heterogeneous database integration. The traditional database approach to heterogeneous database integration was a 'query- driven' approach data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis.

<u>Difference between Operational Database systems and Data Warehouse</u>

- Operational Database systems
  - Main task is to perform on-line transaction and query processing. These systems are called **on-line transaction processing (OLTP)** systems.
  - They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.
- Data Warehouse
  - serve users or knowledge workers in the role of data analysis and decision making.
  - Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as **on-line analytical processing (OLAP)** systems.

<u>Difference between OLTP and OLAP</u>

- **Users and system orientation:**
  - OLTP system is *customer-oriented and is used for* transaction and query processing by clerks, clients, and information technology professionals.
  - OLAP system is *market-oriented and is used for data analysis by knowledge* workers, including managers, executives, and analysts.
- **Data contents:**
  - OLTP system manages current data
  - OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.
- **Database design:**
  - An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.

- An OLAP system typically adopts either a *star or snowflake model* *and a subjectoriented* database design.

- **View:**
  - An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations.
  - An OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization.
  - OLAP systems also deal with information that originates from different organizations.
  - OLAP data are stored on multiple storage media.
- **Access patterns:**
  - The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms.

Accesses to OLAP systems are mostly read-only operations although many could be complex queries.

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

DIFFERENCES BETWEEN OLTP AND **DWH**

OLTP	DWH
Designed to support business transactions support	Designed to support decision making process
Data is volatile	Data is non volatile
It holds current data	It holds historical data (5-10 years)
Detailed data	Summarized data
Normalized data	Denormalized data
Designed for running the business	Designed for analyzing the business
Clerical/End user access	Managerial access
E-R Modeling	Dimensional modeling
Transaction processing	Query processing
10MB-100GB Database size	100GB-2TB Database size

MINDMAJIX

<u>Need for Data Warehousing</u>

1. The data ware house market supports such diverse industries as manufacturing, retail, telecommunications, and health care. Think of a personnel database for a company that is continually modified as personnel are added and deleted... If management wishes determine if there is a problem with too many employees quitting. To analyze this problem, they would need to know which employees have left, when they left, why they left, and other information about their employment. For management to make these types of high-level business analyses, more historical data not just the current snapshot are required.

   A data warehouse is a data repository used to support decision support systems
2. The basic motivation is to increase business profitability. Traditional data processing applications support the day-to-day clerical and administrative decisions, while data warehousing supports long-term strategic decisions.

3. For increasing customer focus, which includes the analysis of customer buying patterns (such as buying preference, buying time, budget cycles, and appetites for spending)

4. For repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions in order to fine tune production strategies; analyzing operations and looking for sources of profit.

5. For managing the customer relationships, making environmental corrections, and managing the cost of corporate assets.

6. The below figure shows a simple view of a data warehouse. The basic components of a data warehousing system include data migration, the warehouse, and access tools. The data are extracted from operational systems, but must be reformatted, cleansed, integrated, and summarized before being placed in the warehouse.

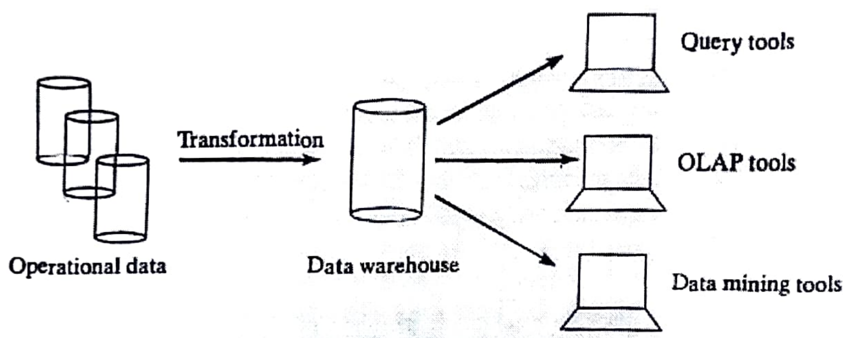


FIGURE 2.14: Data warehouse.

Challenges for Data Warehousing

1. Unwanted data must be removed.
2. Converting heterogeneous sources into one common schema. This problem is the same as that found when accessing data from multiple heterogeneous sources. Each operational database may contain the same data with different attribute names. For example, one system may use "Employee ID," while another uses "EID" for the same attribute. In addition, there may be multiple data types for the same attribute.
3. As the operational data is probably a snapshot of the data, multiple snapshots may need to be merged to create the historical view.
4. Summarizing data is performed to provide a higher level view of the data. This summarization may be done at multiple granularities and for different dimensions.
5. New derived data (e.g., using age rather than birth date) may be added to better facilitate decision support functions.
6. Handling missing and erroneous data must be performed. This could entail replacing them with predicted or default values or simply removing these entries. The portion of the transformation that

deals with ensuring valid and consistent data is sometimes referred to as data scrubbing or data staging.

7. Data warehouse queries are often complex. They involve the computation of large groups of data at summarized levels, and may require the use of special data organization, access, and implementation methods based on multidimensional views.
8. **Data Quality** – In a data warehouse, data is coming from many disparate sources from all facets of an organization. When a data warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges. Poor data quality results in faulty reporting and analytics necessary for optimal decision making.
9. **Understanding Analytics** – When building a data warehouse, analytics and reporting will have to be taken into design considerations. In order to do this, the business user will need to know exactly what analysis will be performed.
10. **Quality Assurance** – The end user of a data warehouse is using Big Data reporting and analytics to make the best decisions possible. Consequently, the data must be 100 percent accurate or a credit union leader could make ill-advised decisions that are detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue that will require a lot of resources to ensure the information provided is accurate.
11. **Performance** – Building a data warehouse is similar to building a car. A car must be carefully designed from the beginning to meet the purposes for which it is intended. Yet, there are options each buyer must consider to make the vehicle truly meet individual performance needs. A data warehouse must also be carefully designed to meet overall performance requirements. While the final product can be customized to fit the performance needs of the organization, the initial overall design must be carefully thought out to provide a stable foundation from which to start.
12. **Designing the Data Warehouse** – People generally don't want to "waste" their time defining the requirements necessary for proper data warehouse design. Usually, there is a high level perception of what they want out of a data warehouse. However, they don't fully understand all the implications of these perceptions and, therefore, have a difficult time adequately defining them. This results in miscommunication between the business users and the technicians building the data warehouse. The typical end result is a data warehouse which does not deliver the results expected by the user. Since the data warehouse is inadequate for the end user, there is a need for fixes and improvements immediately after initial delivery.
13. **User Acceptance** – People are not keen to changing their daily routine especially if the new process is not intuitive. There are many challenges to overcome to make a data warehouse that is quickly adopted by an organization.
14. **Cost** – A frequent misconception among credit unions is that they can build data warehouse in-house to save money.. The harsh reality is an effective do-it-yourself effort is very costly.

Applications of DWH

There are three kinds of data warehouse applications: information processing, analytical processing, and data mining.

1) Information processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs. A current trend in data warehouse information processing is to construct low-cost web-based accessing tools that are then integrated with web browsers.

2) Analytical processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historic data in both summarized and detailed forms. The major strength of online analytical processing over information processing is the multidimensional data analysis of data warehouse data.

3) Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

Different areas are:

- **Banking Industry**
  - In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.
  - Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.
  - Analysis of card holder's transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity.

- **Finance Industry**
  - Revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

- **Consumer Goods Industry**
  - They are used for prediction of consumer trends, inventory management, market and advertising research.
  - In-depth analysis of sales and production is also carried out.

- **Government and Education**
  - The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.
  - The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers.
  - Criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.
  - Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management.

- **Healthcare**

- All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.

- **Hospitality Industry**
  - A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services.
  - They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.
- **Insurance**
  - The warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants.
  - The design of tailor-made customer offers and promotions is also possible through warehouses.
- **Manufacturing and Distribution Industry**
  - A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyze current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.
  - They also use them for product shipment records, records of product portfolios, identify profitable product lines, analyze previous data and customer feedback to evaluate the weaker product lines and eliminate them.
  - For the distributions, the supply chain management of products operates through data warehouses.
- **The Retailers**
  - Retailers serve as middlemen between producers and consumers.
  - They use warehouses to track items, their advertising promotions, and the consumers buying trends.
  - They also analyze sales to determine fast selling and slow selling product lines and determine their shelf space through a process of elimination.
- **Services Sector**

Data warehouses find themselves to be of use in the service sector for maintenance of financial records, revenue patterns, customer profiling, resource management, and human resources

- **Telephone Industry**
  - The telephone industry operates over both offline and online data burdening them with a lot of historical data which has to be consolidated and integrated.

- Analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.

- **Transportation Industry**

    - In the transportation industry, data warehouses record customer data enabling traders to experiment with target marketing where the marketing campaigns are designed by keeping customer requirements in mind.

    - To analyze customer feedback, performance, manage crews on board as well as analyze customer financial reports for pricing strategies.