Objective: To introduce the fundamental principles, algorithms and applications of intelligent data processing and analysis and to provide an in depth understanding of various concepts and popular techniques used in the field of data mining.

A **data center or datacenter (or datacentre),** also called a server farm,1 is a facility used to house computer systems and associated components, such as telecommunications and storage systems. A data center is a physical facility where (usually) multiple companies' computers are located. It is often so that servers can have a larger-bandwidth Internet connection than the company can get in their own facility, as well as having people dedicated to facilities management, including cooling, power, fire prevention, security, etc.

Data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis. A data warehouse is a type of database, or a manner of using a database, to collect large amounts of data.

A **data mart** is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, that may be distributed to support business needs.

So there can be **one or more Data Marts**, that exist **in a Data Warehouse** that is **hosted in a Data Center** that may contain more than one Data Warehouse plus other services.

Data mining is a process of statistical analysis

It will have a total of 80 marks of final exam and 20 marks as internal marking.

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviours:** Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated discovery of previously unknown patterns:** Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can produce the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed.

Databases can be larger in both depth and breadth:

- **More columns:** Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.
- **More rows:** Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

The most commonly used techniques in data mining are:

- **Artificial neural networks:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms:** Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method:** A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.
- **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

Both data mining and data warehousing are business intelligence tools that are used to turn information (or data) into actionable knowledge. The important distinctions between the two tools are the methods and processes each uses to achieve this goal.

Data mining is a process of statistical analysis. Analysts use technical tools to query and sort through terabytes of data looking for patterns. Usually, the analyst will develop a hypothesis, such as customers who buy product X usually buy product Y within six months. Running a query on the relevant data to prove or disprove this theory is data mining. Businesses then use this information to make better business decisions based on how they understand their customers' and suppliers' behaviors.

Data warehousing describes the process of designing how the data is stored in order to improve reporting and analysis. Data warehouse experts consider that the various stores of data are connected and related to each other conceptually as well as physically. A business's data is usually stored across a number of databases. However, to be able to analyze the broadest range of data, each of these databases needs to be connected in some way. This means that the data within them need a way of being related to other relevant data and that the physical databases themselves have a connection so their data can be looked at together for reporting purposes.

So the crux of the relationship between data mining and data warehousing is that if data is properly warehoused, then it is easier to mine. If a data mining query has to run through terabytes of data spread across multiple databases, which sit on different physical networks - - that is not an efficient query and getting results will take a long a time. However, if the data warehouse expert designs a data storage system that closely connects relevant data in different databases, the data miner can now run much more meaningful and efficient queries to improve the business.

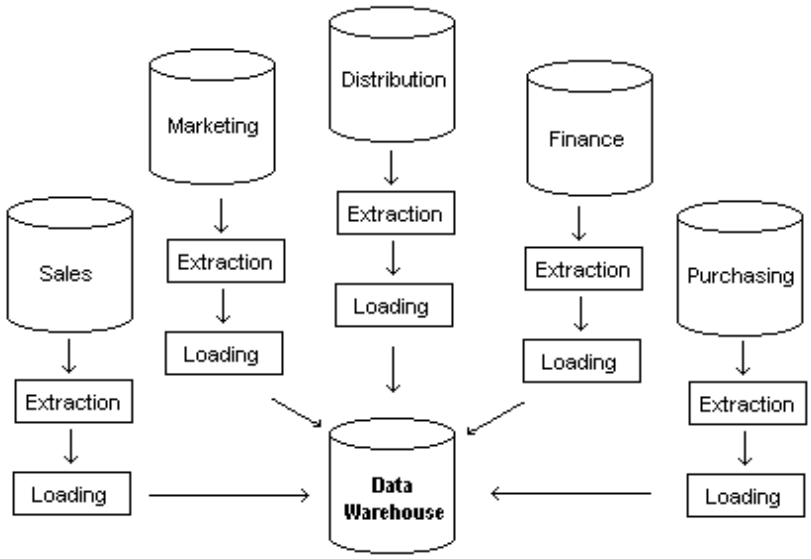


Fig. Data Warehouse

A **data warehouse** is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site. Data warehouses are constructed via a process of **data cleaning**, **data integration**, **data transformation**, **data loading** and **periodic data refreshing**.

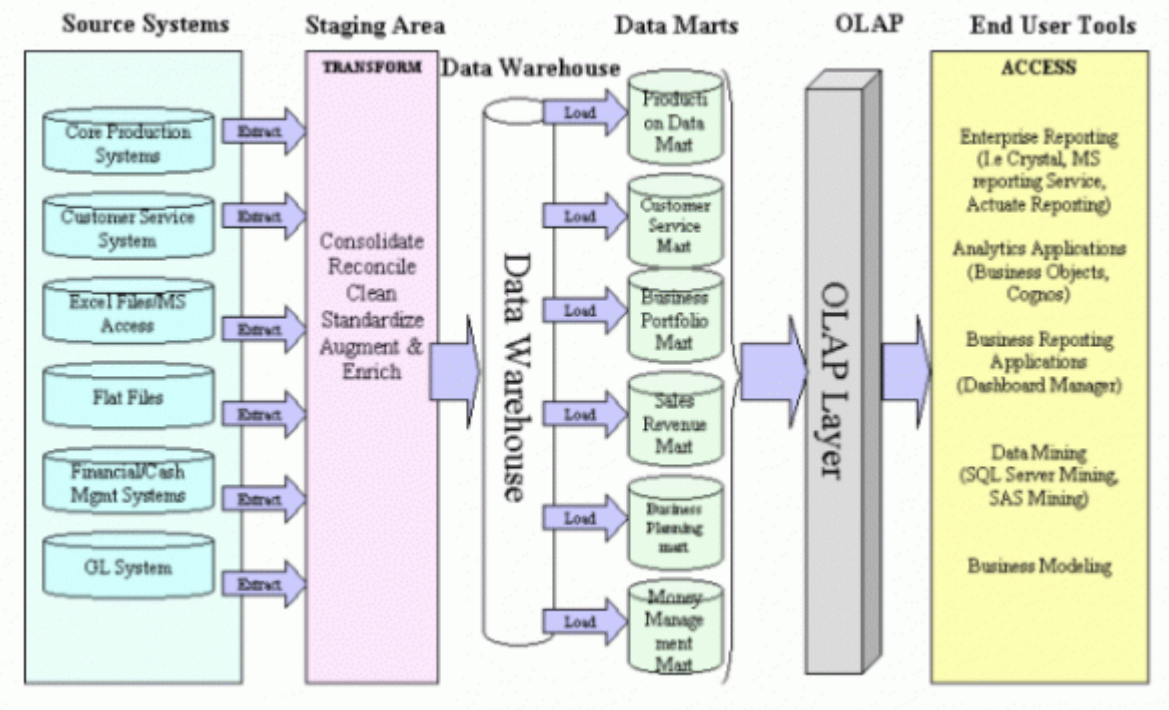


Fig. Typical Framework of data warehouse with data mining, OLAP and other statistical analysis

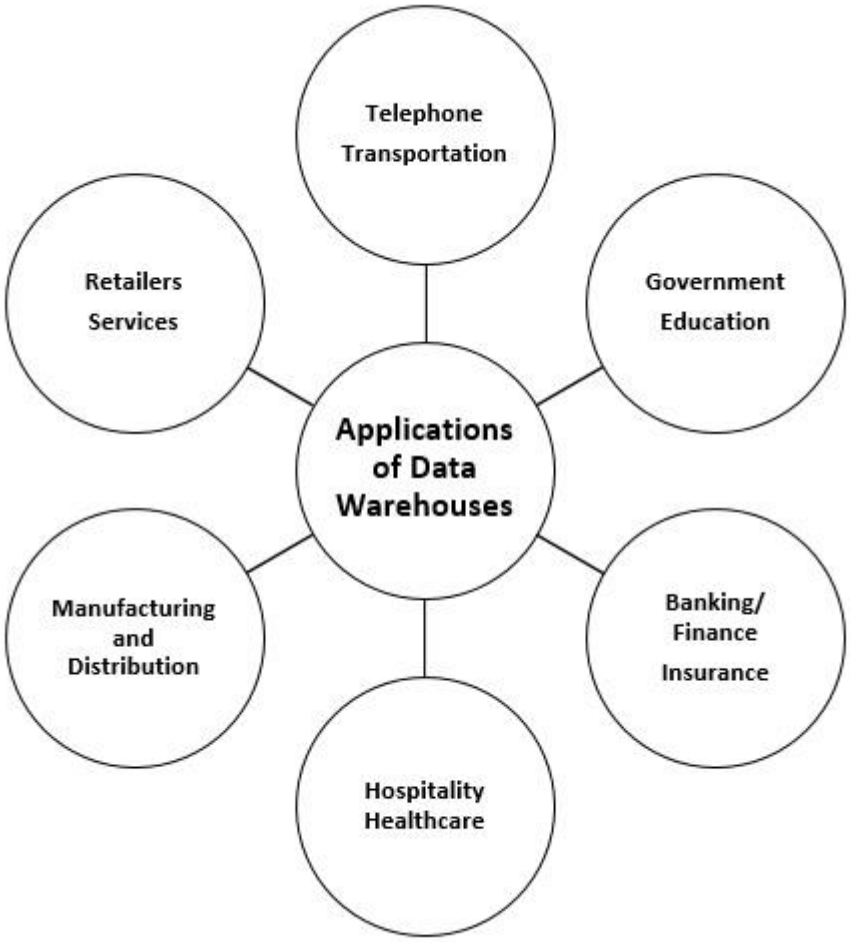
In **data warehouse** data are stored to provide information from a historical perspective (such as from the past 5-10 years) and are typically summarized. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or summarized to a high level, for each sales region.

A data warehouse is usually modeled by a **multidimensional database structure**, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as *count* or *sales_amount*. The actual physical structure of a data warehouse may be relational data store or a **multidimensional data cube**. A data cure provides a multidimensional view of data and allows the pre-computation and fast accessing of summarized data.

A **data mart** is a department subset of a data warehouse. It focuses on selected subjects and its scope is department-wide, where data warehouse span an entire organization and its scope is department-wide.
Data warehouse tools help support data analysis, additional tools for **data mining** are required to allow more in depth and automated analysis.

12 Applications of Data Warehouse

12 Applications of Data Warehouse: Data Warehouses owing to their potential have deep-rooted applications in every industry which use historical data for prediction, statistical analysis, and decision making.



Banking Industry

In the banking industry, concentration is given to risk management and policy reversal as well analysing consumer data, market trends, government regulations and reports, and more importantly financial decision making.

Most banks also use warehouses to manage the resources available on deck in an effective manner. 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

Similar to the applications seen in banking, mainly 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. Apart from these, information is exchanged business partners and clientele.

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 analyse tax records, health policy records and their respective providers, and also their entire 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. The entire financial department of most universities depends on data warehouses, inclusive of the Financial Aid department.

Healthcare

One of the most important sector which utilizes data warehouses is the Healthcare sector. All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyse 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

As the saying goes in the insurance services sector, "Insurance can never be bought, it can only be sold", the warehouses are primarily used to analyse data patterns and customer trends, apart from maintaining records of already existing participants.

Manufacturing and Distribution Industry

This industry is one of the most important sources of income for any state. 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, analyse 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, analyse previous data and customer feedback to evaluate the weaker product lines and eliminate them.

The Retailers

Retailers serve as middlemen between producers and consumers. It is important for them to maintain records of both the parties to ensure their existence in the market.

They use warehouses to track items, their advertising promotions, and the consumers buying trends. They also analyse 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.

Apart from those operations, 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.

The internal environment of the industry uses them to analyse customer feedback, performance, manage crews on board as well as analyse customer financial reports for pricing strategies.

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 or rows are enlarged 10 times. It takes no more than 10 times to execute a query.
    - **Column (Dimension) Salability** − 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.

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.
- SStandardization 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.