1. Data Mining Origin
2. Data Mining & Data Warehousing basics

- **Data** : Raw piece of information that is capable of being moved and store.

- **Data Mining :**

"The process of discovering meaningful patterns and trends often previously unknown by using some mathematical algorithm on huge amount of stored data"

"Extraction of interesting, non-trivial(*describing any task that is not quick and easy to accomplish.),* implicit(*information that is not provided intentionally but gathered from available data streams),* previously unknown and potentially useful information or patterns from data in large database."

- Data mining is basically concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.

Comparison of explicit and implicit data that can be gathered from a Facebook update:

Jill posts to her Facebook page, "Jill is going to lunch early with her best friend Megan at Iron Pit BBQ!." From this Facebook status the explicit data tells us that Jill and Megan are going to eat barbeque for lunch and where they're going to eat.
However, from this same status one can derive a vast amount of implicit data such as:

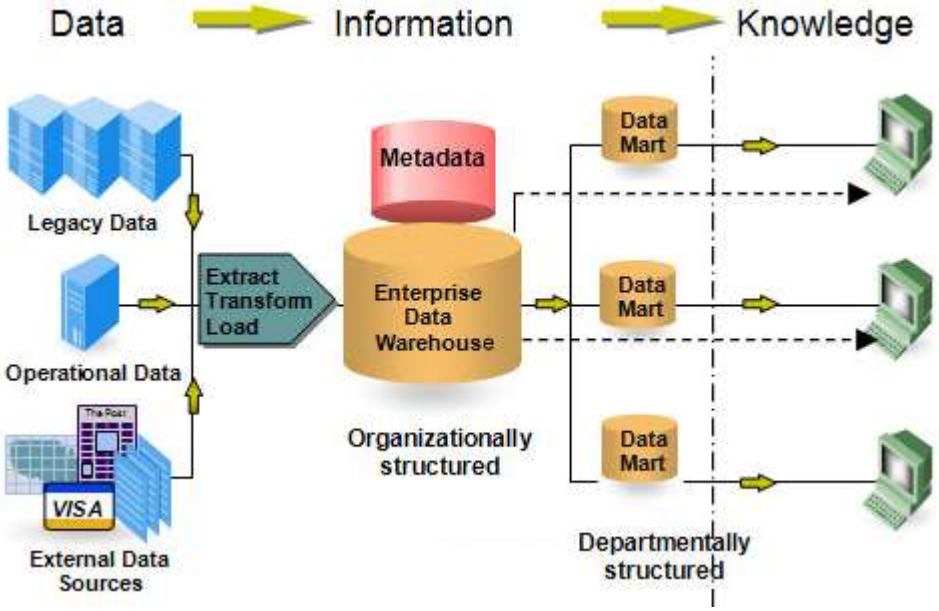
- *Jill has a job.*
- *Jill has a typical lunch time.*
- *Jill went to work today.*
- *Jill and Megan are not vegetarians.*
- *Jill probably has a day job.*
- *Jill and Megan are having a relatively normal/typical day.*
- *Jill and Megan probably don't have special dietary needs.*
- *Jill's work is likely close to Iron Pit BBQ.*

- **Database** : An organized collection of such data in which data are managed in tabular form with relationship.

- **Data Warehouse** : System that organizes all the data available in an organization, makes it accessible & usable for the all kinds of data analysis and also allows to create a lots of reports by the use of mining tools.

- **Data Mart**, is to meet the particular demands of a specific group of users within the organization, such as human resource management (HRM). Generally, an organization's data marts are subsets of the organization's data warehouse.

- **Metadata** has been identified as a key success factor in data warehouse projects. It captures all kinds of information necessary *to extract, transform and load data from source systems into the data warehouse, and afterwards to use and interpret the data warehouse contents.* Metadata summarizes basic information about data, *which can make finding and working with particular instances of data easier*. For example, *author*, *date created* and *date modified* and *file size* are examples of very basic document metadata.



* Overview

- Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective/ future analyses offered by data mining move beyond the analyses of past events provided by retrospective/ review tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They clean databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.
- Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line.

* The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective (expecting in future) and proactive/ active information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently developed:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented/ abnormal rates. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms defines techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective (looking back), static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	prospective (expecting in future) and proactive/ active information delivery

Table 1. Steps in the Evolution of Data Mining.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

* Approaches of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Predictive Data Mining:** Prediction of trends and behaviours - Data mining automates the process of finding predictive information in large databases. Traditionally, required extensive hands-on analysis but now, directly from the data — quickly.

  Example:

  - Targeted marketing: Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings.

- Forecasting economic failure and other forms of default, and identifying segments of a population likely to respond similarly to given events.

X: Vectors of independent variables.
Y: Dependent variables
Y = f(X)

- Users don't care about the model, they simply interested in accuracy of predictions.
- Using unknown examples, the model is trained and the unknown function is learned from data
- The more data with known outcomes is available the better is the predictive power of model.
- Used to predict outcomes whose inputs are known but the output values are not realized yet.
- Never 100% accurate.
- The performance of a model on past data is not predicting the known outcomes.
- Suitable for unknown data set.
- Typical questions answered by predictive models are:
  - ✓ . Who is likely to respond to next product?
  - ✓ . Which customers are likely to leave in the next six months?

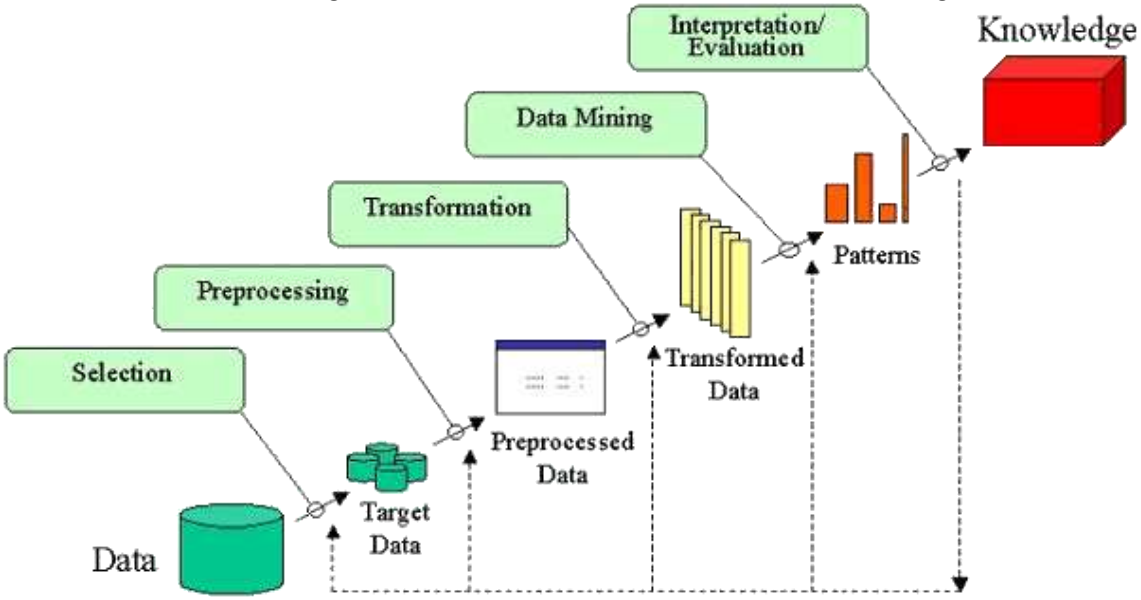
- **Descriptive Data Mining:** Discovery of previously unknown patterns **-** Data mining tools sweep through databases and Detect or identify previously hidden patterns in one step.

Example:

- Analysis of retail sales data, to identify likely unrelated products that are often purchased together.
- Detecting fraudulent/ fake credit card transactions and identifying anomalous/ abnormal data that could represent data entry keying errors.
- It characterizes the general properties of data in the database.
- It finds patterns in data the user determinants which ones are important.
- Mostly used during data exploration.
- Typical questions answered by descriptive data mining are: .
  - ✓ . What is in the data?
  - ✓ . What doesn't look like?
  - ✓ . Are there any unusual patterns?
  - ✓ . What does the data suggest for customer segmentation?
- User may have no idea on which kind of patterns are interesting?
- Functionalities of descriptive data mining are: Clustering, Summarization, Visualization, and Association.

KDD (Knowledge Discovery in Databases)

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.
Major KDD application areas include marketing, fraud detection, telecommunication and manufacturing.



1. Developing an understanding of
   - the application domain
   - the relevant prior knowledge
   - the goals of the end-user
2. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.
3. Data cleaning and preprocessing.
   - Removal of noise or outliers.
   - Collecting necessary information to model or account for noise.
   - Strategies for handling missing data fields.
   - Accounting for time sequence information and known changes.
4. Data reduction and projection.
   - Finding useful features to represent the data depending on the goal of the task.
   - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
5. Choosing the data mining task.
   - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
6. Choosing the data mining algorithm(s).
   - Selecting method(s) to be used for searching for patterns in the data.
   - Deciding which models and parameters may be appropriate.
   - Matching a particular data mining method with the overall criteria of the KDD process.
7. Data mining.
   - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
8. Interpreting mined patterns.
9. Consolidating discovered knowledge.

The terms knowledge discovery and data mining are distinct.

- KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, pre-processing, sampling, and projections of the data prior to the data mining step.
- Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

*** Data Mining Process**

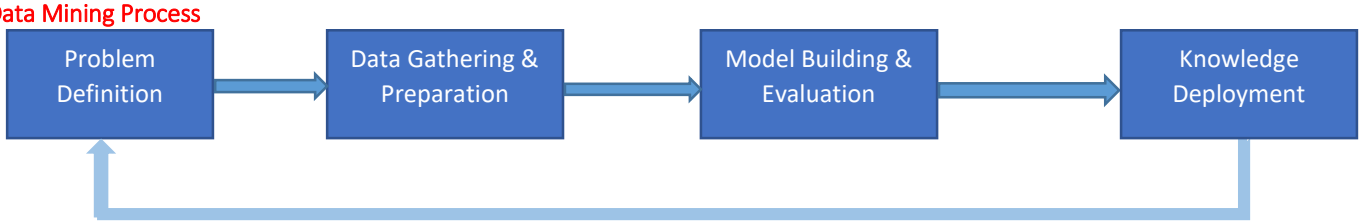


Fig: Data mining process flow

- **Problem Definition:**
  - Focuses on Understanding the project objectives and requirements in terms of business perspective.

  E.g.: How can I sell more of my product to customer? Which customers are most likely to purchase the product?

  - **Business understanding**: In the business understanding phase:
    - First, it is required to understand business objectives clearly and find out what are the business's needs.
    - Next, we have to assess the current situation by finding the resources, assumptions, constraints and other important factors which should be considered.
    - Then, from the business objectives and current situations, we need to create data mining goals to achieve the business objectives within the current situation.
    - Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible.
  - **Data understanding**
    - First, the data understanding phase starts with initial data collection, which we collect from available data sources, to help us get familiar with the data. Some important activities must be performed including data load and data integration in order to make the data collection successfully.
    - Next, the "gross" or "surface" properties of acquired data need to be examined carefully and reported.

- Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting, and visualization.
- Finally, the data quality must be examined by answering some important questions such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"

- **Data Gathering and Preparation:**
  The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.
  - Data Collection & Exploration.
  - Identify data quality, patterns in data.
  - Data preparation phase covers all the tasks involved to build the model.
  - Data preparation tasks are likely to be performed multiple and not in any prescribed order.

- **Model Building and Evaluation:**
  Modelling:
  - First, modeling techniques have to be selected to be used for the prepared dataset.
  - Next, the test scenario must be generated to validate the quality and validity of the model.
  - Then, one or more models are created by running the modeling tool on the prepared dataset.
  - Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives.

  Evaluation: In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to the new patterns that have been discovered in the model results or from other factors.
  - Various modelling techniques are applied and calibrated the parameters to optimal values.
  - Evaluate how well the model satisfies the originally stated business goal.
  - Does the model achieve the business objective?
  - Have all business issues been considered?

- **Knowledge Deployment:**
  The knowledge or information, which we gain through data mining process, needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, in deployment phase, the plans for deployment, maintenance, and monitoring have to be created for implementation and also future supports. From the project point of view, the final report of the project needs to summary the project experiences and review the project to see what need to improved created learned lessons.
  - Use data mining within a target environment e.g. business goal, objectives.
  - Insight and actionable information can be derived from data for decision making.

Data Mining Vs. Query Tools

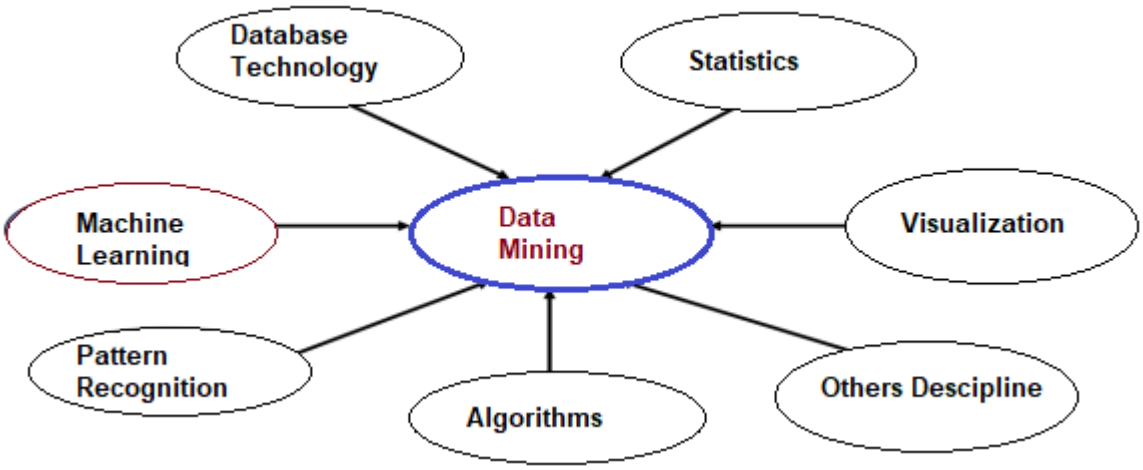
i. SQL can find normal queries from the database such as what is an average turnover? Whereas data mining tools find interesting patterns and facts such as what are the important trends in sells?
ii. Data mining is much more faster than SQL in trend and pattern analysis since it uses algorithm like machine learning, genetic algorithm.
iii. If we know exactly what we are looking for, we use SQL nut if we know only unclearly what we are looking for we use data mining.
iv. Hybrid information can't be easily being traced using SQL.

Why Data Mining?

Data mining is a combination of multidisciplinary field. It can be applied in many fields and can be done using many algorithm and techniques.

* Functions of Data Mining

- **Anomaly detection** (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- **Association rule learning** (dependency modelling) – Searches for relationships between variables. **For example:** *market basket analysis: a supermarket might gather data on customer purchasing habits. e.g. if you buy shoes, then you can buy brand new socks.*
- **Clustering** – is the task of discovering groups and structures in the data that are in some way or another "similar".
- **Classification** – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "real" or as "spam".
- **Regression** – attempts to find a function, which models the data with the least error i.e. for estimating the relationships among data or datasets.
- **Summarization** – providing a more compact representation of the data set, including visualization and report generation.



*** Techniques used in data mining**

Data mining techniques can yield 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. When data mining tools are implemented on high performance parallel processing systems, they can analyse massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyse huge quantities of data. Larger databases, in turn, yield improved predictions. Databases can be larger in both depth or more columns and breadth or more rows:

Techniques 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. E.g. Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .
- Genetic Algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- Nearest Neighbour 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. Sometimes called the k-nearest neighbour technique.
- Rule Induction: The extraction of useful if-then rules from data based on statistical significance.

*** Major issues in Data Mining**

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multidimensional space
  - Data Mining-an interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling uncertainty, noise, or incompleteness of data
  - Patter evaluation and pattern or constraint-guided mining
- User Interaction
  - Interactive mining: dynamically change the focus of a search
  - Incorporation of background knowledge
  - Ad hoc data mining and data mining query languages
  - Presentation and visualization of data mining results
- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining algorithms
- Diversity of Database types
  - Handling complex types of data
  - Mining dynamic, networked and global data repositories
- Data Mining and Society
  - social impacts of data mining: should address individual privacy and data protection rights
  - privacy-preserving data mining: to observe data sensitivity and preserve people's privacy while performing successful data mining
  - invisible data mining: hidden data mining algorithms e.g. web search engines

* Applications of Data Mining

Major in two fields DM is applied:-

1. **BI (Business Intelligence):** provides historical, current and predictive views of business operations. E.g. reporting, OLAP, business performance management, competitive intelligence, predictive analytics.
2. **Web Search Engines**: searches information from web by crawling, indexing and searching techniques.

A wide range of companies have deployed successful applications of data mining, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships.

Two critical factors for success with data mining are:

i. a large, well-integrated data warehouse and
ii. a well-defined understanding of the business process within which, data mining is to be applied (such as customer prospecting, retention, campaign management, and so on).

Some successful application areas include: -

- **A medicine company:** can analyse its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems.
- **Credit card company:** A credit card company can influence its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified.
- **Transportation company:** to identify the best prospects for its services. Using data mining to analyse its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects.
- **Supermarket or CRM:** to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.
- **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: misuse detection, anomaly detection, used must be efficient and scalable, and capable of handling network data of high volume, dimensionality, and heterogeneity.
  - Association and correlation analysis, aggregation to help select and build discriminating attributes: Association and correlation mining can be applied to find relationships between system attributes describing the network data. Such information can provide insight regarding the selection of useful attributes for intrusion detection. New attributes derived from aggregated data may also be helpful, such as summary counts of traffic matching a particular pattern.
  - Analysis of Stream data: frequently encountered together, find sequential patterns, and identify outliers, real-time intrusion detection.
  - Distributed data mining: analyze network data from several network locations in order to detect these distributed attacks.
  - Visualization and query tools: Visualization tools should be available for viewing any anomalous patterns detected. Such tools may include features for viewing associations, clusters, and outliers.
- **Telecommunication Industry**

  Today the telecommunication industry provides 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 telecom industry is rapidly expanding. Therefore, 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. Examples for which data mining improves telecommunication services are: –
  - Multidimensional Analysis of Telecommunication data: dimensions such as calling-time, duration, location of caller, location of callee, and type of call.
  - Fraudulent pattern analysis and the identification of unusual patterns.

    (1) identify potentially fraudulent users and their atypical usage patterns;
    (2) detect attempts to gain fraudulent entry to customer accounts; and
    (3) discover unusual patterns that may need special attention, such as *busy-hour frustrated call attempts, switch and route congestion patterns, and periodic calls from automatic dial-out equipment (like fax machines) that have been improperly programmed*
  - Multidimensional association and sequential patterns analysis: promote the sales of specific long-distance and cellular phone combinations and improve the availability of particular services in the region.
  - Mobile Telecommunication services: unusually busy mobile phone traffic at certain locations may indicate something abnormal happening in these locations. Moreover, ease of use is crucial for attracting customers to adopt new mobile services

- Use of visualization tools in telecommunication data analysis: Tools for OLAP visualization, linkage visualization, association visualization, clustering, and outlier visualization

- **Medical: Biological Data Analysis**

In the field of biology such as genomics, proteomics, functional Genomics and biomedical research, 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: Genomic and proteomic data sets are often generated at different labs and by different methods. They are distributed, heterogenous, and of a wide variety. The semantic integration of such data is essential to the cross-site analysis of biological data
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences: identify highly conserved residues among genomes, and such conserved regions can be used to build phylogenetic trees to infer evolutionary relationships among species
- Discovery of structural patterns and analysis of genetic networks and protein pathways.: discover approximate and frequent structural patterns and to study the regularities and irregularities among such interconnected biological networks.
- Association and path analysis: identifying co-occurring gene sequences and linking genes to different stages of disease development
- Visualization tools in genetic data analysis: Alignments among genomic or proteomic sequences and the interactions among complex biological structures are most effectively presented in graphic forms, transformed into various kinds of easy-to-understand visual displays

- **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: Multidimensional data analysis methods should be used to analyze the general properties of banking and financial data. For example, one may like to view the debt and revenue changes by month, by region, by sector, and by other factors, along with maximum, minimum, total, average, trend, and other statistical information. Data warehouses, data cubes, multi feature and discovery-driven data cubes, characterization and class comparisons, and outlier analysis all play important roles in financial data analysis and mining.
- Loan payment prediction and customer credit policy analysis: factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus the total monthly income), payment to-income ratio, customer income level, education level, residence region, and credit history. Analysis of the customer payment history may find that, say, payment-to income ratio is a dominant factor, while education level and debt ratio are not.
- Classification and clustering of customers for targeted marketing: use classification to identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.
- Detection of money laundering and other financial crimes: To detect money laundering and other financial crimes, it is important to integrate information from multiple databases (like bank transaction databases, and federal or state crime history databases).

- **Retail Industry**

Retail Industry collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. 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: The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers can help improve company profits.
- Customer Retention- analysis of customer loyalty: With customer loyalty card information, one can register sequences of purchases of particular customers. Customer loyalty and purchase trends can be analyzed systematically.
- Product recommendation and cross-referencing of items: By mining associations from sales records, one may discover that a customer who buys a digital camera is likely to buy another set of items. Such information can be used to form product recommendations.

* Data Warehouse

- A data warehouse is a united repository for all the data that an enterprise's various business systems collect. The repository may be physical or logical.
- Data warehousing emphasizes the capture of data from diverse sources for useful analysis and access, but does not generally start from the point-of-view of the end user who may need access to specialized, sometimes local databases. The latter idea is known as the data mart.
- In most of the organization, there occur large databases in operation for normal daily transactions called operational database.
- A data warehouse is a large database built from the operational database

- **Data Warehouse**

- System that organizes all the data available in an organization, makes it accessible & usable for the all kinds of data analysis and also allows to create a lot of reports by the use of mining tools.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."
- Typically, a data warehouse is housed on an enterprise mainframe server or increasingly, in the cloud. Data from various online transaction processing (OLTP) applications and other sources is selectively extracted for use by analytical applications and user queries.

- **Data warehousing:**
  - The process of constructing and using data warehouses.
  - Is the process of extracting & transferring operational data into informational data & loading it into a central data store (warehouse)

Characteristics of Data Warehouse: *A data warehouse should be…*

i. **Time – dependent**
- There must be a connection between the information in the warehouse and the time when it was entered.
- One of the most important aspect of the warehouse as it relates to data mining, because information can then be sourced according to period.
- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element".

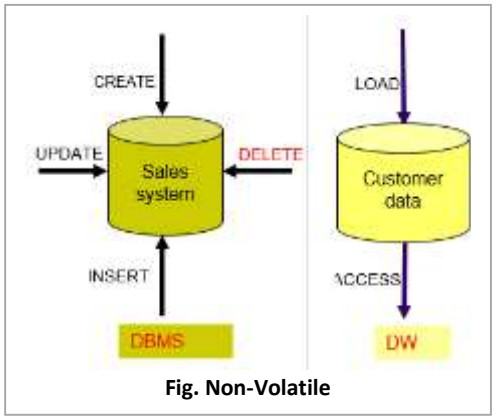


Fig. Non-Volatile

ii. **Non-Volatile**
- Data in a warehouse is never updated, but used only for queries.
- End-users who want to update data must use operational database.
- A data warehouse will always be filled with historical data.
- A physically separate store of data transformed from the operational databases.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing: *initial loading of data* and *access of data*.

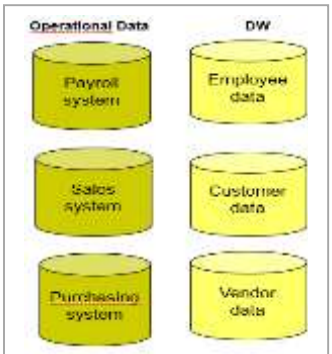


Fig. Subject-Oriented

iii. **Subject Oriented**
- Not all the information in the operational database is useful for a data warehouse. A data warehouse should be designed especially for decision support and expert system with specific related data.
- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

iv. **Integrated**
- Constructed by integrating multiple, heterogeneous data sources: relational databases, flat files, on-line transaction records.
- Data cleaning and data integration techniques are applied: Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources. E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

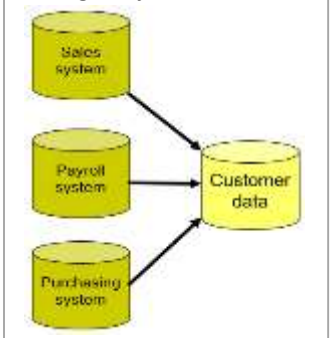
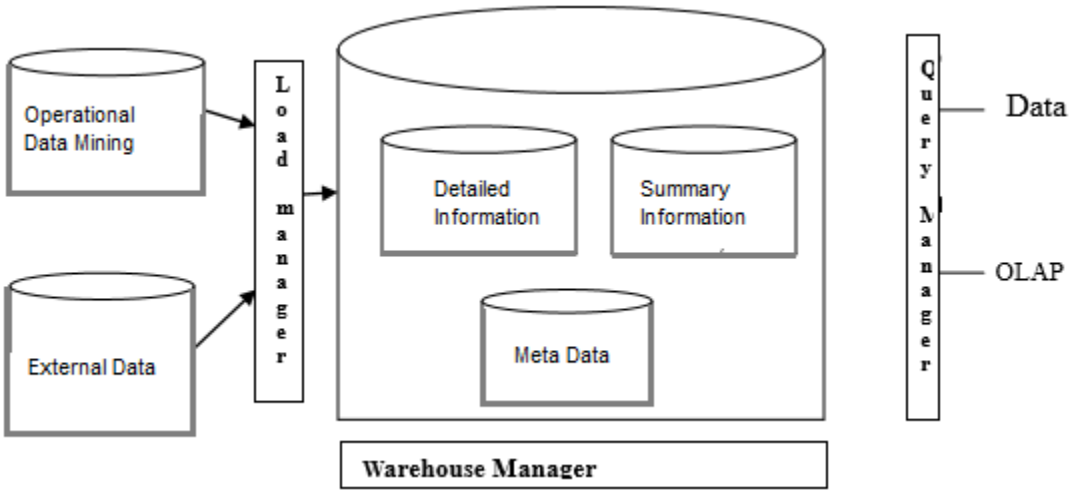


Fig. Integrated

*Architecture of a Data Warehouse



- **Load Manager**: The system components that perform all the operations necessary to support the extract and load process. It fast loads the extracted data into a temporary data store and performs simple transformations into a structure similar to the one in the data warehouse. It performs the following operations:
    - Extract the data from the source systems: transferred from source systems, and made available to the DW.
    - Fastest load the extracted data into a temporary data store and speed data processing in DW
    - Perform simple transformation into a structure similar to one in DW
- **Warehouse Manager**: Performs all the necessary operations to support the warehouse management process. It analyses the data to perform consistency and referential checks. It also transforms and merges the source data in the temporary data store into the published data warehouse with creating indexes and business views. Update all existing aggregations and back up data in the data warehouse. It performs the following operations:
    - Analyze the data to perform consistency and referential integrity check.
    - Transform and merge the source data in the temporary data store into the DW.
    - Generate renormalization if appropriate.
    - Backup totally the data within the DW.
    - Create Index & Views (combine a number of partitions into a single fact table)
    - Generate the summaries
- **Query Manager**: Performs all the operations necessary to support the query management process by directing queries to the appropriate tables. In some cases, it also stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate. It performs the following operations:
    - Direct queries e appropriate tables
    - Schedule the execution of user queries.

- **Detailed Information**: Stores all the detailed information to determine the business requirements to analyse the level at which to recollect detailed information in the data warehouse. Detailed information can be managed by the topics:
    - Data warehouse schemas
    - Fact data
    - Dimension data
    - Partitioning data
- **Summary Information**: Stores all the predefined aggregations generated by the warehouse manager. It is a transient area which will change on an ongoing basis in order to respond to changing query profiles. It is essentially a replication to detailed information. The implication of summary data is that the data:
    - Exists to speed up the performance of common queries
    - Increases operational cost
    - May have to be updated every time new data is loaded into the DW
    - May not have to be backed up, because it can be generated fresh from the detailed info

    - Guideline1: avoid creating a summary that require more than 200 centralized summary tables on an ongoing basis.
    - Guideline2: inform users that summary table accessed infrequently will be dropped on an ongoing basis
- **Meta Data**: Meta data is data about data which describes how information is structured within a data warehouse. It maps data stores to common view of information with the data warehouse.

***** Approaches to Data Warehousing

1. **Top Down**: The top down approach spins off data marts for specific groups of users after the complete data warehouse has been created.
2. **Bottom Up:** The bottom up approach builds the data marts first and then combines them into a single, all-encompassing data warehouse.

Top-Down Approach: *First DW is built then data is loaded into Data Mart*

- Data is extracted from the various source systems. The extracts are loaded and validated in the stage area. Validation is required to make sure the extracted data is accurate and correct. You can use the ETL (Extract, Transfer, Load) tools or approach to extract and push to the data warehouse.
- Then, data is extracted from the data warehouse in regular basis by applying various aggregation, summarization techniques on extracted data and loaded back to the data warehouse.
- Once the aggregation and summarization are completed, various data marts extract that data and apply the some more transformation to make the data structure as defined by the data marts.

Advantages of top-down design are:

- Provides reliable dimensional views of data across data marts, as all data marts are loaded from the data warehouse.
- This approach is good against business changes. Creating a new data mart from the data warehouse is very easy.

Disadvantages of top-down design are:

- This methodology is inflexible to changing departmental needs during implementation phase.
- It represents a very large project and the cost of implementing the project is significant.

Bottom-Up Approach: *First Data Marts are created, then data is loaded into DW*

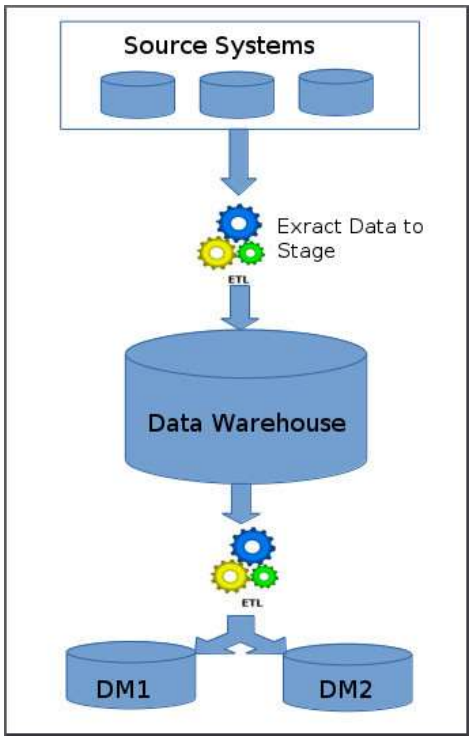
- The data flow in the bottom up approach starts from extraction of data from various source system into the stage area (ETL) where it is processed and loaded into the data marts that are handling specific business process.
- After data marts are refreshed the current data is once again extracted in stage area and transformations are applied to create data into the data mart structure. The data is the extracted from Data Mart to the staging area is aggregated, summarized and so on loaded into Enterprise DW (EDW) and then made available for the end user for analysis and enables critical business decisions.

Advantages of bottom-up design are:

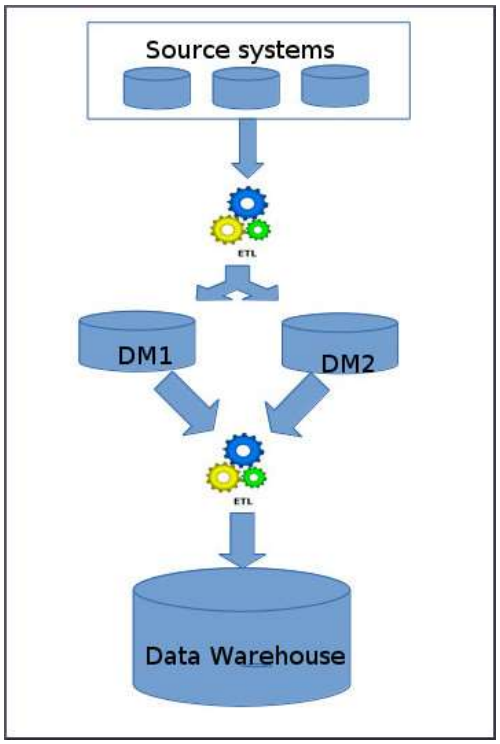
- This model contains reliable data marts and these data marts can be delivered quickly.
- As the data marts are created first, reports can be generated quickly.
- The data warehouse can be extended easily to accommodate new business units. It is just creating new data marts and then integrating with other data marts.

Disadvantages of bottom-up design are:

- The positions of the data warehouse and the data marts are reversed in the bottom-up approach design.



Fig(a): Top-Down Approach



Fig(b): Bottom-Up Approach

***** Benefits of Data Warehouse

A data warehouse maintains a copy of information from the source transaction systems. This architectural complexity provides the opportunity to:

- Integrate data from multiple sources into a single database and data model. Large cluster of data to single database, so a single query engine can be used to present data is an operational data store (ODS) and enabling a central view across the enterprise.
- Mitigate the problem of database isolation level lock contention in transaction processing systems caused by attempts to run large, long running, analysis queries in transaction processing databases.
- Maintain data history, even if the source transaction systems do not.
- Improve data quality, by providing consistent codes and descriptions, flagging or even fixing bad data.
- Present the organization's information consistently.
- Provide a single common data model for all data of interest regardless of the data's source.
- Restructure the data so that it makes sense to the business users and also it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems.
- Make decision–support queries easier to write.
- Optimized data warehouse architectures allow data scientists to organize and disambiguate repetitive data

Difference between data warehouse and data mart

Attributes	Data warehouse	Data mart
Scope	Enterprise-wide data	Department-wide data
Focus	Multiple subject areas	Single subject area
Design	Difficult to design	Easy to design
Time	Takes more time to build	Less time to build
Memory	Larger memory	Limited memory

Types of data marts

i. **Dependent Data Mart –** This data mart depends on the enterprise data warehouse and works in top-down manner.
ii. **Independent Data Mart –** This data mart does not depend on the enterprise data warehouse and works in bottom-up manner.

Database vs. Data Warehouse

Attributes	Database	Data Warehouse
Definition	Any collection of data organized for storage, accessibility, and retrieval.	A type of database that integrates copies of transaction data from disparate source systems and provisions them for analytical use.
Types	usually applies to an OLTP application database, which we'll focus on throughout this table. Other types of databases include OLAP (used for data warehouses), XML, CSV files, flat text, and even Excel spreadsheets.).	A data warehouse is an OLAP database. An OLAP database layers on top of OLTPs or other databases to perform analytics. They differ according to how the data is modeled. Most data warehouses employ either an enterprise or dimensional data model.
Similarities	Both OLTP and OLAP systems store and manage data in the form of tables, columns, indexes, keys, views, and data types. Both use SQL to query the data.	
How used	Typically constrained to a single application: one application equals one database. OLTP allows for quick real-time transactional processing. It is built for speed and to quickly record one targeted process (ex: patient admission date and time).	Accommodates data storage for any number of applications: one data warehouse equals infinite applications and infinite databases. OLAP allows for one source of truth for an organization's data. This source of truth is used to guide analysis and decision-making within an organization *(ex: total patients over age 18 who have been readmitted, by department and by month*).
Service Level Agreement (SLA)	OLTP databases must typically meet 99.99% uptime. System failure can result in disorder and complaints. The database is directly linked to the front-end application. Data is available in real time to serve the here-and-now needs of the organization.	With OLAP databases, SLAs are more flexible because occasional downtime for data loads is expected. The OLAP database is separated from frontend applications, which allows it to be scalable. Data is refreshed from source systems as needed (typically this refresh occurs every 24 hours). It serves historical trend analysis and business decisions.
Optimization	Optimized for performing read-write operations of single point transactions. An OLTP database should deliver sub-second response times. Performing large analytical queries on such a database is a bad practice, because it impacts the performance of the	Optimized for efficiently reading/retrieving large data sets and for aggregating data. Because it works with such large data sets, an OLAP database is heavy on CPU and disk bandwidth. A data warehouse is designed to handle large analytical queries.

	system for clinicians trying to use it for their day-to-day work.	
Data Organization	An OLTP database structure features very complex tables and joins because the data is normalized (it is structured in such a way that no data is duplicated).	In an OLAP database structure, data is organized specifically to facilitate reporting and analysis, not for quick-hitting transactional needs. The data is de-normalized to enhance analytical query response times and provide ease of use for business users.
Reporting/Analysis	Because of the number of table joins, performing analytical queries is very complex. They usually require the expertise of a developer or database administrator familiar with the application. Reporting is typically limited to more static, siloed needs.	With fewer table joins, analytical queries are much easier to perform. This means that semi-technical users (anyone who can write a basic SQL query) can fill their own needs. The possibilities for reporting and analysis are limitless. There's an intrinsic need for aggregating, summarizing, and drilling down into the data. A data warehouse enables you to perform many types of analysis: ▪ Descriptive (what has happened) ▪ Diagnostic (why it happened) ▪ Predictive (what will happen) ▪ Prescriptive (what to do about it)

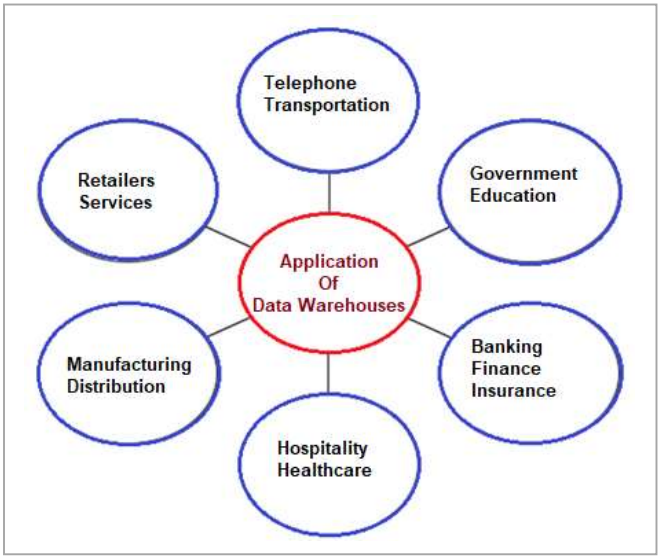
Q. Why Data Warehouse Separated from Operational/ Transactional Databases?

Ans.: Data warehousing is an efficient system which is used for making reports and doing analysis. These systems are used to store the past or previous as well as current data used for creating trending reports which is further made use in senior management reporting annually and quarterly. It works by bringing all the data in a central location which is known as data warehouse. All the data that is stored or uploaded in this data warehouse is done from the operational systems. But there is a great difference that is to be considered between operational systems and data ware housing.

The data ware housing systems and databases do not require real time data validation but the operational database do require data validation tables on a regular basis. Data ware house has only few users that is up to hundred while the operational database has many concurrent users. Single line transactions are related to the operational database while the bulk load with data ware housing database. The data ware house is flexible but the operational databases are known to provide high performance. Data warehousing is located in separate system basically to increase the performance of the system and also reduces the cost involved per analysis.

Applications of Data Warehouse

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 analyzing 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 Sector 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.
  - **Agriculture :** The agricultural census compiles a large number of agricultural parameters at the national level .District wise agricultural production area and yield of crops is compiled, analysis, mining and forecasting. Statistics on consumption of fertilizers can be turned into a data merge. Data on agricultural inputs such as seeds and fertilizers can also be effectively analysed in a data warehouse.
  - **Rural development:** Data on individuals below the poverty line can be built into a data warehouse. Drinking water census data (from drinking water mission) can be effectively utilized by OLAP and data mining technologies. Monitoring and analysis of progress made on implementation of rural development programs can also be made using OLAP and data mining technologies.
  - **Health:** Community needs assessment data, immunization data, data from national programs on controlling blindness, leprosy, malaria can all used for data warehousing implementation, OLAP and data mining applications.
  - **Planning:** At the planning commission, data warehouses can be built for the state plan data on all sectors, labour, energy, education, trade and industry, five-year plan etc.
  - **Education**: The educational survey data has been converted into a data warehouse, various types of analytical queries and reports can be answered.
  - **Commerce and trade:** Data link on trade can be analysed and converted into a data warehouse. World price monitoring system can be made to perform better by using data warehousing and data mining technologies
  - **Tourism:** Tourist arrival behaviour and performances, tourism products data, foreign exchange earnings data and hotels, travel and transportation data. Predictably the government departments have largely been satisfied with developing single management information system or limited cases a few databases which were used online for limited purposes.

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

Reference…. Er. Pratap Sapkota