Data Mining

Databases today can range in size into the terabytes — more than 1,000,000,000,000 bytes of data. Within these masses of data lies hidden information of strategic importance. But when there is such enormous amount of data how could we draw any meaningful conclusion? We are data rich, but information poor. The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation.

Data mining is being used both to increase revenues and to reduce costs. The potential returns are enormous. Innovative organizations worldwide are using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud.

What is data mining?

Mining is a brilliant term characterizing the process that finds a small set of precious bits from a great deal of raw material.

Definition:

- Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.
- Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in unique ways that are both understandable and useful to the data owner.
- Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large database.

Many other terms carry a similar or slightly different meaning to data mining, such as *knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology,* and *data dredging.*

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. While others view data mining as simply an essential step in Knowledge discovery.

Applications:

Market Basket Analysis Fraud Detection Intrusion Detection

Customer Segmentation Bio Informatics

Knowledge Discovery in Database (KDD)

1. **Data Cleaning** - noise and inconsistent data is removed.
2. **Data Integration** - multiple data sources are combined.
3. **Data Selection** – data appropriate to the analysis task are retrieved from the database.
4. **Data Transformation** - data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. **Data Mining** - intelligent methods are applied in order to extract data patterns.
6. **Pattern Evaluation** - data are evaluated to identify the truly interesting patterns.

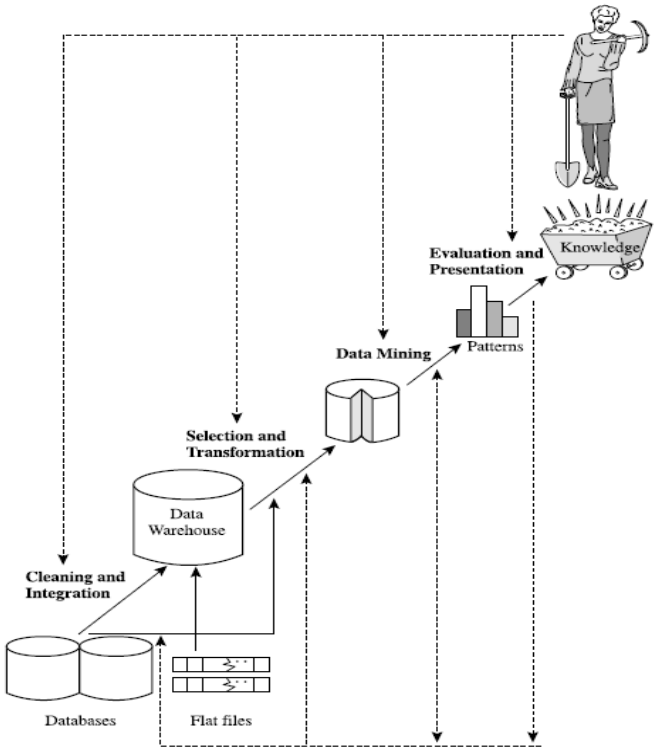


Figure 1. Data mining as a step in Knowledge Discovery

7. **Knowledge Presentation** - knowledge is represented using visualization techniques.

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. So, according to this view, data mining is only one step in entire knowledge discovery process. However, in industry, in media, and in the database research area, the term data mining is becoming more popular than the longer term of knowledge discovery from data.

Based on this view, the architecture of a typical data mining system has following major components:

Data sources: Data Sources consists of one or a set of databases, data warehouses, spreadsheets, Word Wide Web or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base (KB): a knowledge base is a centralized repository for information: a public library or a database of related information about a particular subject. A KB is not a static collection of information, but a dynamic resource that may itself have the capacity to learn, as part of an artificial intelligence (AI) expert system.

Data mining engine: The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

Pattern evaluation module: The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

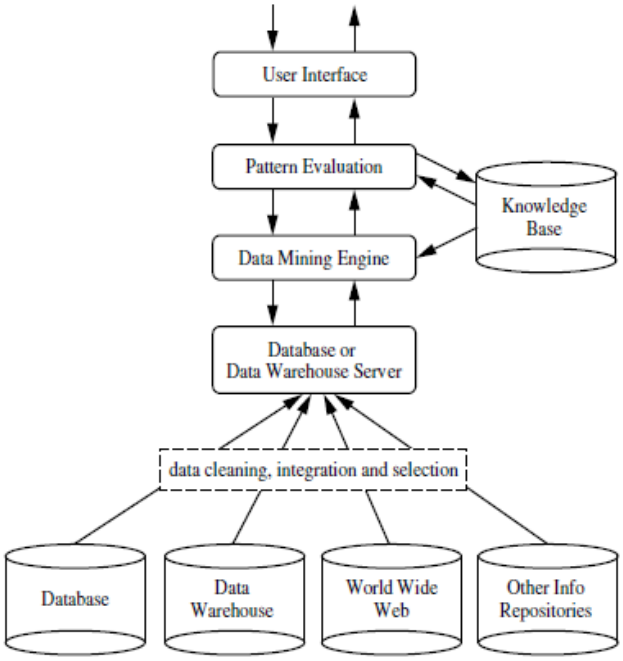


Figure 2. A Typical Data Mining Architecture

User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

https://genuinenotes.com

Data Mining Techniques

Association

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction.

Classification

Classification is a classic data mining technique based on machine learning. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics to classify each item in a set of data into one of a predefined set of classes or groups.

Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

Regression

Regression uses existing values to forecast what other values will be. A regression task begins with a data set in which the target values are known. For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time. The data might track age, height, weight, developmental milestones, family history, and so on. Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.

Data Warehouse

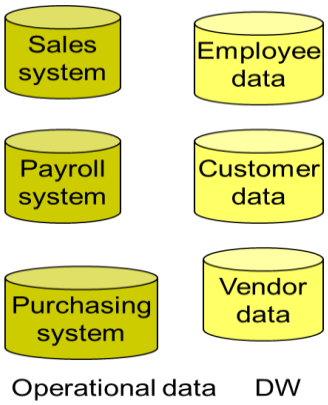
In most of the organization, there occur large database in operation for normal daily transactions called operational database. A data warehouse is a large database built from the operational database that organizes all the data available in an organization, makes it accessible & usable for the all kinds of data analysis and also allows to create a lots of reports by the use of mining tools.

"A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process."

A data warehouse should be:

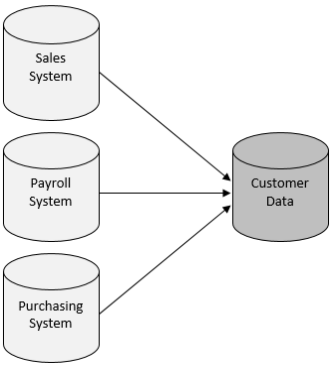
Subject oriented:

- Focus is on Subject Areas rather than Applications
- Organized around major subjects, such as customer, product, sales.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.



Integrated

- Constructed by integrating multiple, heterogeneous data sources. The result is that data-once it resides in the data warehouse- has a single physical corporate appearance.
- Integration tasks handles naming conventions as well as physical attributes of data.

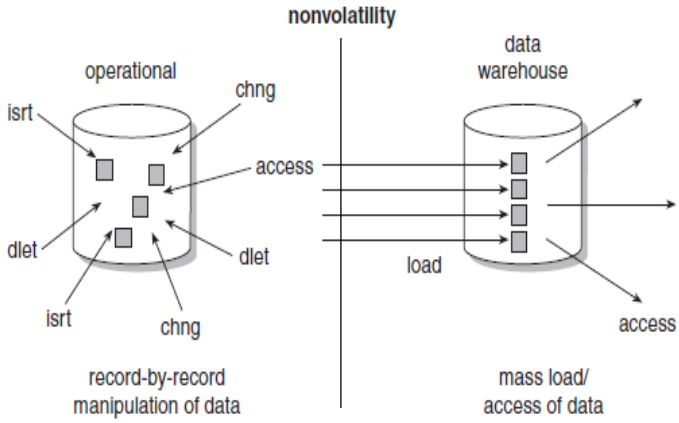


Time variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data. (60 to 90 days)
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)

Non volatile

- Data Warehouse is relatively static in nature.
- Data warehouse data is loaded (usually en masse) and accessed, but it is not updated (in the general sense).
- Data warehouse data is a sophisticated series of snapshots, each taken at one moment in time. The effect created by the series of snapshots is that the data warehouse has a historical sequence of activities and events.



Architecture of a Data Warehouse System

A typical data warehouse system has three main phases:

- **Data acquisition**
  - Relevant data collection
  - Recovering: transformation into the data warehouse model from existing models
  - Loading: cleaning and loading in the DW
- **Storage**
- **Data extraction**

  Tool examples: Query report, SQL, multidimensional analysis (OLAP tools), data mining

These three tasks are performed by following personnel:

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. Also called ETL (Extract Transform and Load).

Warehouse Manager (Data Manager): It is the system component that performs analysis of data to ensure consistency. The data from various sources and temporary storage are merged into data

warehouse by the warehouse manager. The job of backing-up and archiving data as well as creation of index is performed by this manager.

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.

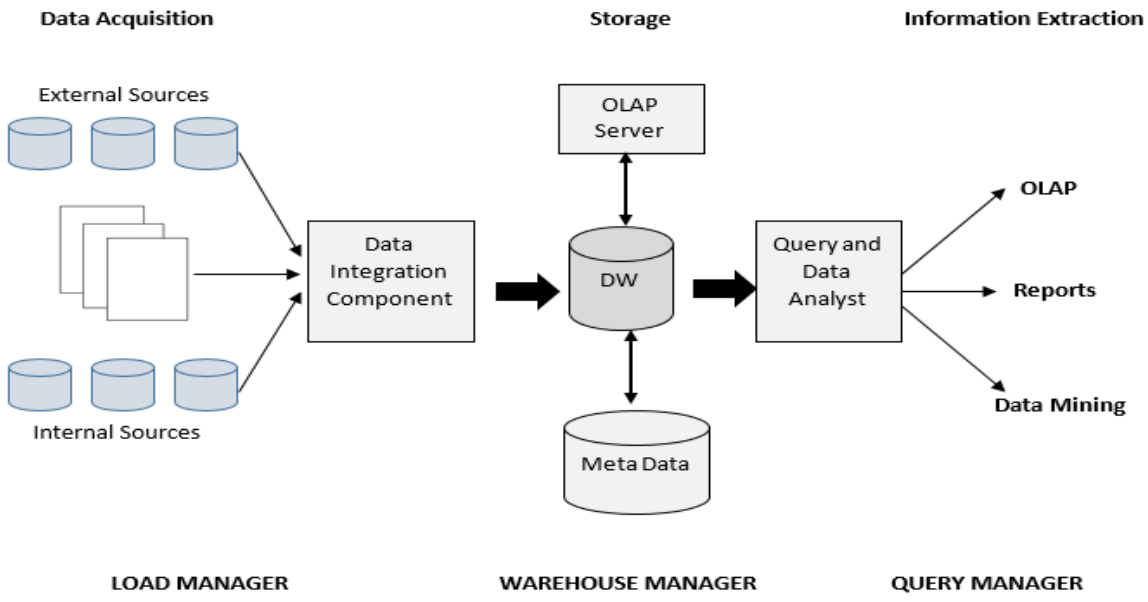


Figure 3. Data Warehouse Architecture

Origin of Data Mining

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history. The traditional data analysis techniques have encountered following practical difficulties in meeting the challenges that motivated the development of data mining:

Scalability: because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes or even petabytes are becoming common. If the data mining algorithms are to handle these massive data sets, then they must be scalable. For example scalability can be improved by using sampling or developing parallel and distributed algorithms.

High dimensionality: it is now common to encounter data sets with hundreds of attributes instead of the handful common a few decades ago. Data sets with temporal or spatial components often

have high dimensionality. The traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high dimensional data.

Heterogeneous and complex data: as the role of data mining in business, science, medicine and other field has grown, so has the need for technique that can handle heterogeneous attributes. The traditional data analysis techniques only deals with data sets containing attributes of the same type, either continuous or categorical.

Data ownership and distribution: sometimes the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques.

Non-traditional analysis: the traditional statistical approach is based on a hypothesize-and-test paradigm. In this approach, a hypothesis is purposed, an experiment is designed to gather data and then the data is analyzed with respect to hypothesis. Unfortunately, this process is extremely labor intensive. Current data mining techniques require this task to be done in huge scale. So, the data mining techniques have been motivated by the desire to automate the process of hypothesis generation and evaluation.

So, to meet all those challenges, researchers from different disciplined began to focus on developing more efficient and scalable tools that could handle diverse types of data. This work, which culminated in the field of data mining, built upon the methodology and algorithms that researchers had previously used.

Data mining draws upon the ideas such as

- Sampling, estimation and hypothesis testing from statistics
- Search algorithm, modeling techniques and learning theories from artificial intelligence
- Various other ideas from pattern recognition and machine learning
- And database systems are of course needed for providing support for efficient storage, indexing and query processing.