## Subject: DATA MINING code:BCACsT5.11

## Chapter-1

## Introduction to Data Mining

## Basic Terminologies:

## Data: is collection of raw and unorganized facts. Data can be any character, text, word, number, etc.

## Information: is organized and processed data presented in a useful manner.

## Dataset: Collection of data or related sets of information. Most commonly a dataset corresponds to single database table.

## Database: A database is a

## collection of information that is organized so that it can be easily accessed, managed and updated.

**Database System:** is a System to organize data, to store a large number of dynamically associated data, to facilitate multi user access to computer hardware, software and data.ie, computer system with database technology.

## DBMS: It is a software or collection of programs that enables you to store, modify and extract information from a database.

## 

## Knowledge: Information read, heard or seen and understood and integrated.

## Wisdom: the ability to use knowledge and experience to make good decisions and judgments

## Data warehouse: A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources. A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision making process”

## Image result for data miningData Mining: Extraction of Knowledge from Databases.

## 

## Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and AI .

## Definitions:

## Extraction of interesting and potentially useful patterns or knowledge from large amounts of Data.

* The iterative and interactive process of discovering valid, novel, useful, and understandable knowledge ( patterns, models, rules etc.) from **Massive** databases

**History of Data Mining**

## Increasing power of technology and complexity of data sets has lead Data Mining to evolve from static data delivery to more dynamic and proactive information deliveries; from tapes and disks to advanced algorithms and massive databases (see the table below). In the late 80`s Data Mining term began to be known and used within the research community by statisticians, data analysts, and the management information systems (MIS) communities. By the early 1990`s, data mining was recognized as a sub-process or a step within a larger process called Knowledge Discovery in Databases (KDD).

## Why to Mine Data

## The Explosive Growth of Data: from terabytes to petabytes

## In the real world, Data is rich but information poor.

## Major sources of abundant data

## Business: Web, e-commerce, transactions, stocks, …

## Science: Remote sensing, bioinformatics, scientific simulation, …

## Society and everyone: news, digital cameras,

## To find new and Hidden data which database alone can’t.

## To discover patterns and new relationships.

## Clean ,Integrate and Transform Data.

## Analyze, Classify and group the Data.

## Present the Data to make proactive, knowledge-driven decisions

## Applications of Data Mining

|  |  |
| --- | --- |
| Applications | Usage |
| **Communications** | **Data mining techniques are used in communication sector to predict customer behavior to offer highly targeted and relevant campaigns.** |
| **Insurance** | **Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing customers.** |
| **Education** | **Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.** |
| **Manufaturing** | **With the help of Data Mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime.** |
| **Banking** | **Data mining helps finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc.** |
| **Service**  **Providers.** | **Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze billing details, customer service interactions, complaints made to the company to assign each customer a probability score and offers incentives.** |
| **E-commerce** | **E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, who use Data mining techniques to get more customers into their eCommerce store.** |
| **Super Markets** | **Data Mining allows supermarket's develop rules to predict if their shoppers were likely to be expecting. By evaluating their buying pattern, they could find woman customers who are most likely pregnant. They can start targeting products like baby powder, baby shop, diapers and so on.** |
| **Crime**  **Investigation** | **Data Mining helps crime investigation agencies to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border** |
| **Bioinformatics** | **Data Mining helps to mine biological data from massive datasets gathered in biology and medicine to find new drugs.** |

## Social Data Mining:

## Social media is data are largely user generated content which is vast , noisy, distributed, unstructured and dynamic.

## Vast amounts of user generated content are created on social media sites everyday i.e eg: facebook, whatsapp, twitter, instagram etc, which is a platform for blogging, chatting, media sharing, social news and other activities.

## Systematically analyzing the valuable information from the social media is called social data mining. It is important because WWW is vast, People share more data, More devices in turn produce more data and platform for Communication and Marketing.

## Use of Social Data Mining.

## Con directly study opinions and behaviors of millions of users to gain insight into

## Human behavior 2. Market Analysis

## General process in Data Mining.

## 

## Data Mining Techniques for Social Data Mining

## Text Mining:

## text mining is an emerging technology that attempts to extract meaningful information from unstructured textual data. It is also called as Text analytics, which helps in deriving high quality information. Some of the applications include web mining, medical, resume filtering etc.

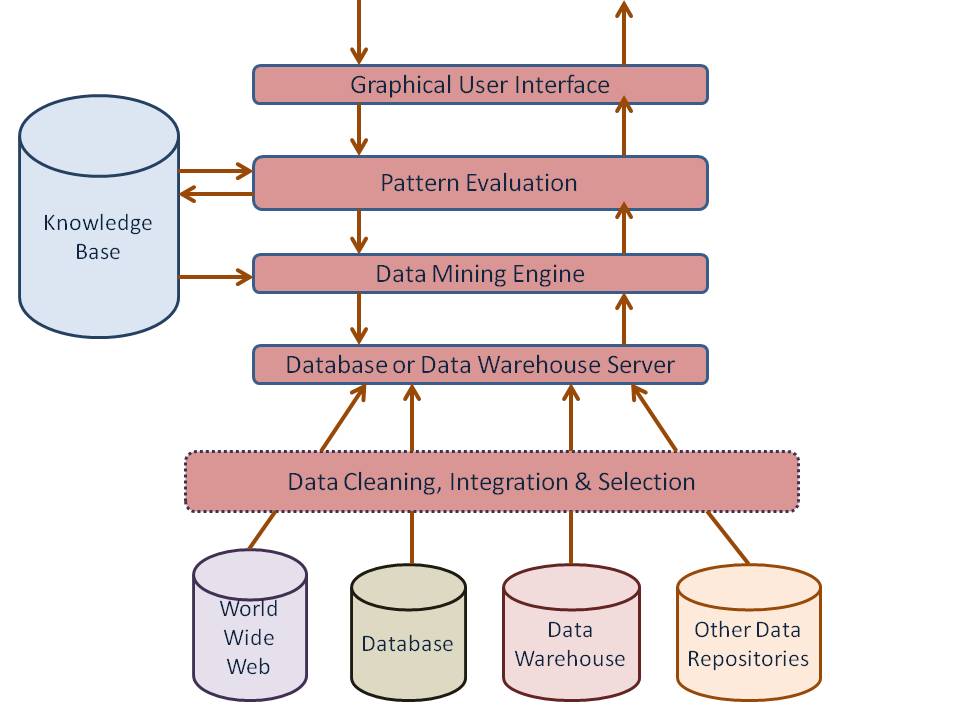
## Graph Mining:

## Graph mining is the process of extracting useful knowledge ( patterns, pathways, links, locations) from structured data that can be represented as a graph. Graph usually constitute a dominant data structure and appear essentially in all forms of information.

## Graph mining and network analytics is critical to a variety of application domains, ranging from community detection in social networks, malicious program analysis in computer security, to searches for functional modules in biological pathways and structural analysis in chemical compounds.

## Data Mining Architecture

Data mining is a very important process where potentially useful and previously unknown information is extracted from large volumes of data. There are a number of components involved in the data mining process. These components constitute the architecture of a data mining system.



The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

## a) Data Sources

Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful. Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

### Different Processes

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable A number of techniques may be performed on the data as part of cleaning, integration and selection.

## b) Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

## c) 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.

## d) Pattern Evaluation Modules

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.

## e) Graphical User Interface

The graphical user interface module communicates between the user and the data mining system. This module helps the user to use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

## f) Knowledge Base

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

**Data Mining Tasks:**

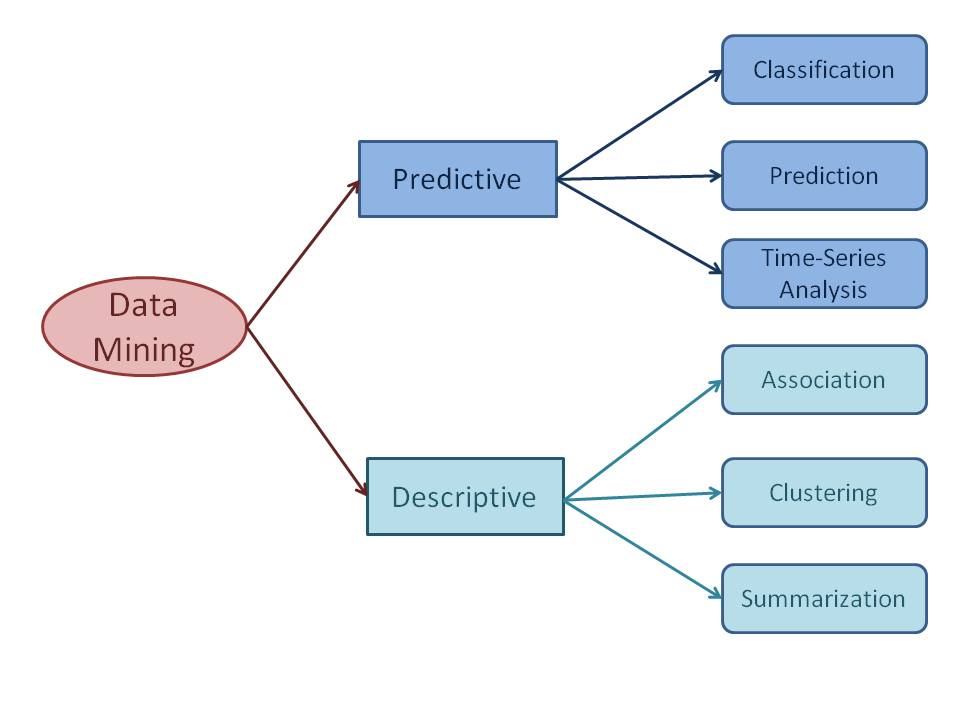
**These are the functions which are involved in collecting, analyzing searching patterns or relationships from data sources.**

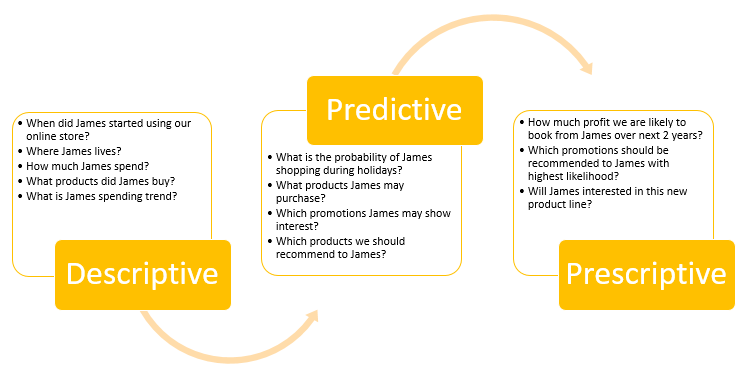
The data mining tasks can be classified generally into two types based on what a specific task tries to achieve. Those two categories are **descriptive tasks** and **predictive tasks.**

**The descriptive data mining tasks characterize the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave.**

**Different Data Mining Tasks**

There are a number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks. **A data mining system can execute one or more of the above specified tasks as part of data mining.**





**1.Predictive data mining tasks** come up with a model from the available data set that is helpful in predicting unknown or future values of another data set of interest.

**a) Classification**

**Classification derives a model to determine the class of an object based on its attributes**. A collection of records will be available, each record with a set of attributes. goal of classification task is assigning a class attribute to new set of records as accurately as possible.

**Classification can be used in direct marketing,** that is to reduce marketing costs by targeting a set of customers who are likely to buy a new product.Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past.

**b) Prediction**

**Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest.** For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

**c) Time - Series Analysis**

**Time series is a sequence of events where the next event is determined by one or more of the preceding events. Time series reflects the process being measured and there are certain components that affect the behavior of a process**. This methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time- series analysis.

**2.Descriptive data mining tasks usually** finds data describing patterns and comes up with new, significant information from the available data set. A retailer trying to identify products that are purchased together can be considered as a descriptive data mining task.

**d) Association**

**Association discovers the connection among a set of items. Association identifies the relationships between objects. Association analysis is used for commodity management, advertising, catalog design, direct marketing etc**. A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion of same kind of products. If a retailer finds that beer and nappy are bought together mostly, he can put nappies on sale to promote the sale of beer.

**e) Clustering**

**Clustering is used to identify data objects that are similar to one another. The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on.** For example, an insurance company can cluster its customers based on age, residence, income etc. This group information will be helpful to understand the customers better and hence provide better customized services.

**f) Summarization**

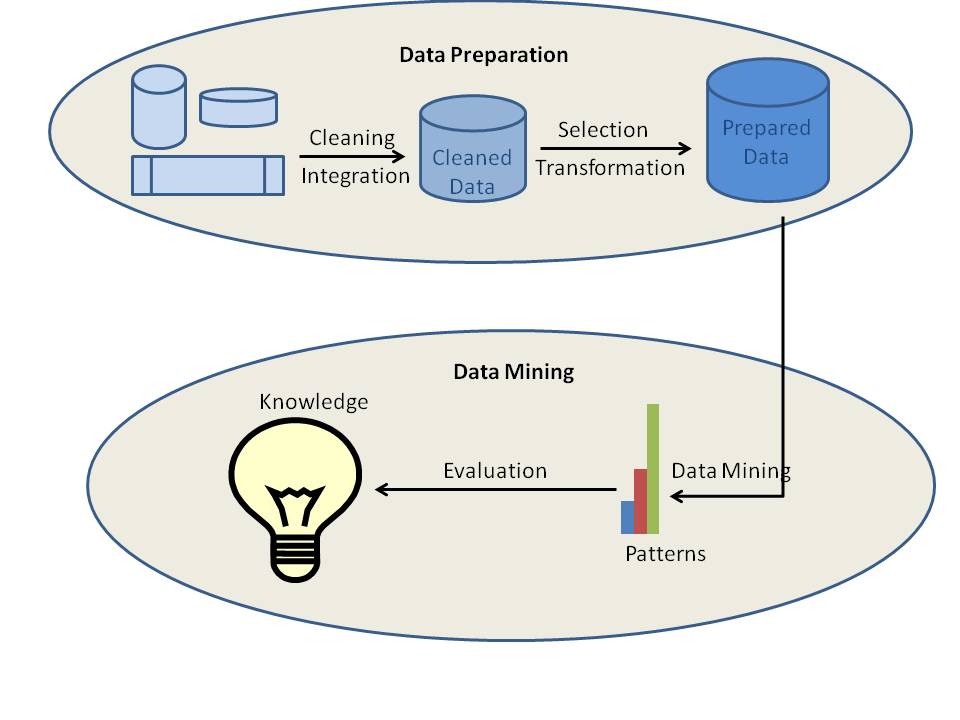
**Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data. For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc.** Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis. Data can be summarized in different abstraction levels and from different angles.

**Data Mining Processes**

**Data Mining is a very complex process than we think involving a number of processes. The whole process of data mining cannot be completed in a single step. In other words, you cannot get the required information from the large volumes of data as simple as that.**

## Types of Data Mining Processes

**Different data mining processes can be classified into two types: data preparation or data preprocessing** and **data mining.** In fact, the first four processes, that are data cleaning, data integration, data selection and data transformation, are considered as **data preparation processes**.The last three processes including data mining, pattern evaluation and knowledge representation are integrated into one process called **data mining.**



### a) Data Cleaning

**Data cleaning is the process where the data gets cleaned. Data in the real world is normally incomplete, noisy and inconsistent. The data available in data sources might be lacking attribute values, data of interest** etc. **Data cleaning involves a number of techniques including filling in the missing values manually or combined computer and human inspection, etc. The output of data cleaning process is adequately cleaned data.**

### b) Data Integration

**Data integration is the process where data from different data sources are integrated into one. Data lies in different formats in different locations.** Data could be stored in databases, text files, spreadsheets, documents, data cubes, Internet and so on. Data integration is a really complex and tricky task because data from different sources does not match normally.. Data integration tries to reduce redundancy to the maximum possible level without affecting the reliability of data.

### c) Data Selection

Data mining process requires large volumes of historical data for analysis. So, usually the data repository with integrated data contains much more data than actually required. From the available data, data of interest needs to be selected and stored. Data selection is the process where the data relevant to the analysis is retrieved from the database.

### d) Data Transformation

**Data transformation is the process of transforming and consolidating the data into different forms that are suitable for mining. Data transformation normally involves normalization, aggregation, generalization etc.** For example, a data set available as "-5, 37, 100, 89, 78" can be transformed as "-0.05, 0.37, 1.00, 0.89, 0.78". Here data becomes more suitable for data mining. After data integration, the available data is ready for data mining.

### e) Data Mining

Data mining is the core process where a number of complex and intelligent methods are applied to extract patterns from data. Data mining process includes a number of tasks such as association, classification, prediction, clustering, time series analysis and so on.

### f) Pattern Evaluation

The pattern evaluation identifies the truly interesting patterns representing knowledge based on different types of interestingness measures. A pattern is considered to be interesting if it is potentially useful, easily understandable by humans.

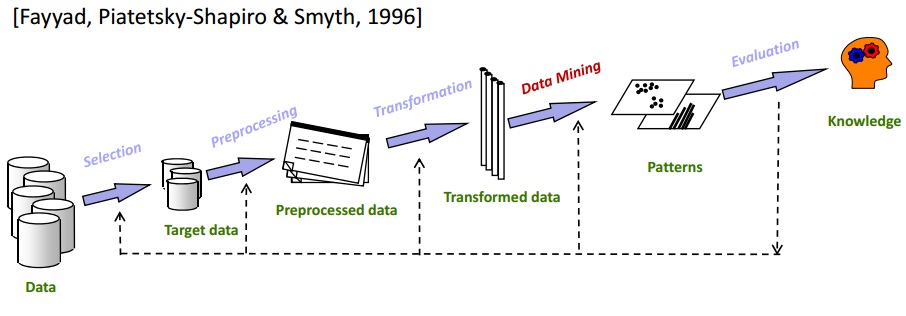
### g) Knowledge Representation

The information mined from the data needs to be presented to the user in a required way. Different knowledge representation and visualization techniques are applied to provide the output of data mining to the users.

**KDD (Knowledge Discovery in Databases)**

**KDD process**

Knowledge Discovery in Databases is the overall process of searching for hidden knowledge in the massive amounts of data that we are technically capable of generating and storing. The term KDD was coined and co-chaired by Usama Fayyad and Ramasamy Uthurusamy in the year 1996.



1. **Data cleaning: it is a preprocessing task inorder**  to remove noise and inconsistent data.
2. **Data integration**, where multiple data sources may be combined to get the relevant data.
3. **Data selection**,the data relevant to the analysis task are retrieved from the database.
4. **Data transformation**, Here the data are transformed and consolidated into forms appropriate for mining by pre forming summary or aggregation operations.
5. **Data mining**, which is an essential process where intelligent methods are applied to extract data patterns.
6. **Pattern evaluation** to identify the truly interesting patterns representing knowledge based on interesting measures.
7. **Knowledge presentation**, where visualization and knowledge representation techniques are used to present mined knowledge to users.

**Difference between KDD and Data mining process.**

KDD is a multi-step process that encourages the conversion of data into useful information. Data mining is one among the steps of Knowledge Discovery in Databases(KDD) Data mining is a very crucial step of the KDD process. Data Mining is an implementation of KDD

**Data mining Issues:**

Data mining systems face a lot of challenges and issues in today’s world, some of them are

**1. Mining Methodology & User Interaction**

* **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
* **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
* **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
* **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
* **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
* **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
* **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

**2. Performance Issues**

There can be performance-related issues such as follows −

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

**3. Diverse Data Types Issues**

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

**Data mining Metrics.**

**Data mining metrics may be defined as a set of measurements which can help in determining the efficacy of a Data mining Method / Technique or Algorithm.** They are important to help take the right decision as like as choosing the right data mining technique or algorithm. Data mining metrics generally fall into the categories of accuracy, reliability, and usefulness.

**Accuracy** is a measure of how well the model correlates an outcome with the attributes in the data that has been provided.

**Reliability** assesses the way that a data mining model performs on different data sets. A data mining model is reliable if it generates the same type of predictions or finds the same general kinds of patterns regardless of the test data that is supplied.

**Usefulness** includes various metrics that tell us whether the model provides useful information. For example, a data mining model that correlates store location with sales might be both accurate and reliable, but might not be useful, because you cannot generalize that result by adding more stores at the same location.

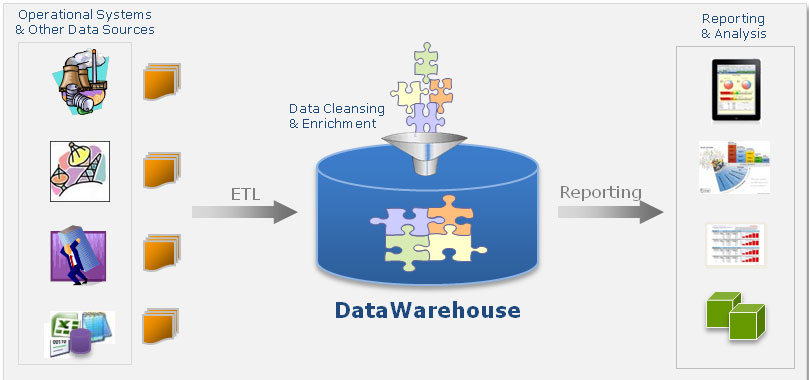
**Data Ware House**

A data warehouse refers to a data repository that is maintained separately from an organization’s operational databases. Data warehouse systems allow for integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historic data for analysis.

The two major functions of a data warehouse are:

• Maintaining past and present records

• Helping organizations to take effective business decisions with precise data analysis.

****

**William H. Inmon** first coined the concept of Data Warehouse (DW) in 1990. According to him

“**A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision making process”**

**data warehousing** :it is the process of constructing and using data warehouses. The construction of a data warehouse requires data cleaning, data integration, and data consolidation.

**Key features of DW**

**Subject-oriented:** A data warehouse is organized around major subjects such as customer, supplier, product, and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view of particular subject issues by excluding data that are not useful in the decision support process.

**Integrated:** A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

**Time-variant:** Data are stored to provide information from an historic perspective (e.g., the past 5–10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, a time element.

**Nonvolatile:** A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data. initial loading of data and access of data.

**DW operates with data extracted from multiple sources- internally built systems, third-party business groups, purchased applications and others. Many operations like production, transactions, sales and marketing, human resourcing are included from these source locations. With the buzz of big data and E-commerce, DW involves dealing with terabytes and petabytes of consumers and products data generated from each website click.**

Data mining uses various analytic tools to create summary reports, which are helpful in taking business decisions. The use of data mining in business applications leads to [Business Analytics](https://intellipaat.com/ibm-cognos-training/) and [Business Intelligence.](https://intellipaat.com/tutorial/msbi-tutorial/introduction-of-msbi/)

**Operational Databases and Warehouses(OLTP and OLAP)**

**Operational Databases**

The major task of online operational database systems is to perform online transaction and query processing. These systems are called online transaction processing(OLTP) systems. They cover most of the day-to-day operations. of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting

**Data warehouses**

Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users. These systems are known as **online analytical processing (OLAP) systems.**

The major distinguishing features of Operational Databases (OLTP) and Data warehouses (OLAP) are summarized as

* **Users and system orientation**: An OLTP system is *customer-oriented* and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is *market-oriented* and is used for data analysis by knowledge workers, including managers, executives, and analysts.
* **Data contents**: An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historic data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for proper decision making.
* **Database design**: An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a *star* or a *snowflake* model and a subject-oriented database design.
* **View**: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historic data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.
* **Access patterns**: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations (because most data warehouses store historic rather than up-to-date information), although many could be complex queries.

Difference B/W Online Transaction Processing**(OLTP)** & Online Analytical Processing**(OLAP)**



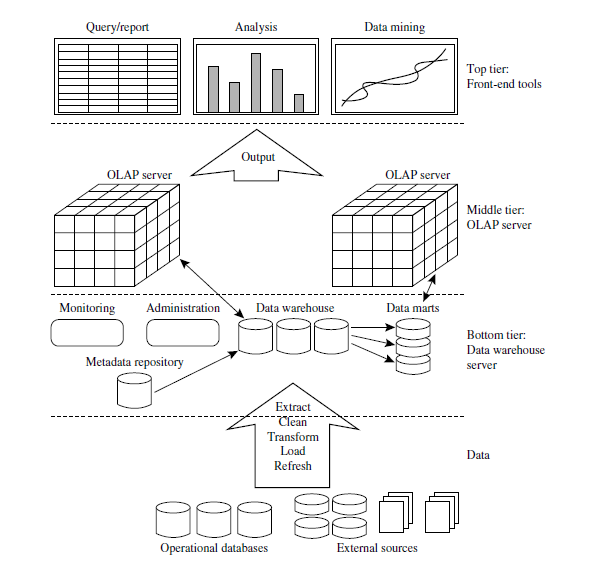
**Data Ware House Architecture(3 tier )**

The Data ware house architecture mainly classified into three tiers

Bottom Tier: Data ware house server

Middle Tier: OLAP server

Top Tier: Front end tools

****

* **Bottom Tier** − The bottom tier of the architecture is the data warehouse server along with the metadata. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, clean, Transform, Load, and refresh functions.
* **Middle Tier** − In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.
  + By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
  + By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.
* **Top-Tier** − This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

**Data Warehouse Terminologies**

**Metadata**

In simple terms, metadata provides the answers to all your data-related questions in the data warehouse. It refers to data about data giving users detailed explanation of the syntax and semantics and describing all relevant attributes of the data in DWH. For instance, the number of tables in a DB can be referred as metadata. It is commonly known as

* Table of Contents for the data
* Data Catalog
* Data directory
* DWH Roadmap

**Data Mart**

It is the subset of data warehouse contain only small slices of data warehouse. Eg: Data pertaining to particular department. Two types of data mart. Dependent and Independent.

Dependent/Direct: sourced directly from data warehouse

Independent/Indirect: sourced from one or more sources

**Data ware house server:**

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

**Olap Server:**

Online Analytical Processing server is based on multidimensional model. It allows managers and analysts to get an insight of information through fast, consistent and interactive access to information.

**Front end tools:** This involves data mining tools and softwares to analyse and identify useful patterns and generate the reports of the same.

OLAP

**OLAP is a specialized tool that creates a multidimensional view of data for the user to do the analysis.**

Short for Online Analytical Processing, this has a category of software tools that provides analysis of data stored in a database. OLAP tools enable users to analyze different dimensions of multidimensional data. For example, it provides time series and trend analysis views. OLAP often is used in data mining.

**Why OLAP ?**

## Traditional SQL’s data-comparison ability is limited. For example, SQL can manage queries, such as a list of sales agents, versus sales volume histories. However, with larger data volumes, it can be overwhelming just to use SQL and tough to translate data into information that easily facilitates decision making. It is difficult to answer certain questions in SQL, such as why product sales are higher mid-month, or why female sales agents consistently outsell their male counterparts during the summer.

**OLAP Server**

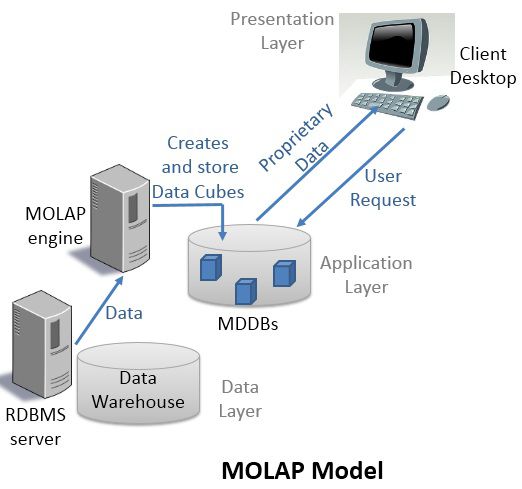
**Online analytical processing (OLAP) is a high-level concept that describes a category of tools that helps in the analysis multi-dimentional queries. It is based on multidimensional model. It allows managers and analysts to get an insight of information through fast, consistent and interactive access to information**.

OLAP came about because of the tremendous complexity and sheer growth associated with business data during the 1970s as the volume and type of information became too heavy for adequate analysis through simple structured query language (SQL) queries.

## There are mainly two types of OLAP server

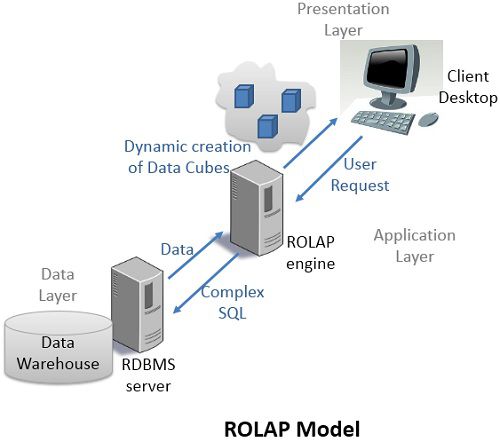
## MOLAP(Multidimensional Online Analytical Processing)

* Multidimensional Data Cube: Most OLAP products are developed based on a structure where the cube is patterned as a multidimensional array. These multidimensional OLAP (MOLAP) products usually offers improved performance when compared to other approaches mainly because they can be indexed directly into the structure of the data cube to gather subsets of data. When the number of dimensions is greater, the cube becomes sparser. That means that several cells that represent particular attribute combinations will not contain any aggregated data. This in turn boosts the storage requirements, which may reach undesirable levels at times, making the MOLAP solution untenable for huge data sets with many dimensions. Compression techniques might help; however, their use can damage the natural indexing of MOLAP.



**ROLAP** (Relational Online Analytical Processing)

* Relational OLAP: Relational OLAP make use of the relational database model. The ROLAP data cube is employed as a bunch of relational tables (approximately twice as many as the quantity of dimensions) compared to a multidimensional array. Each one of these tables, known as a cuboid, signifies a specific view.

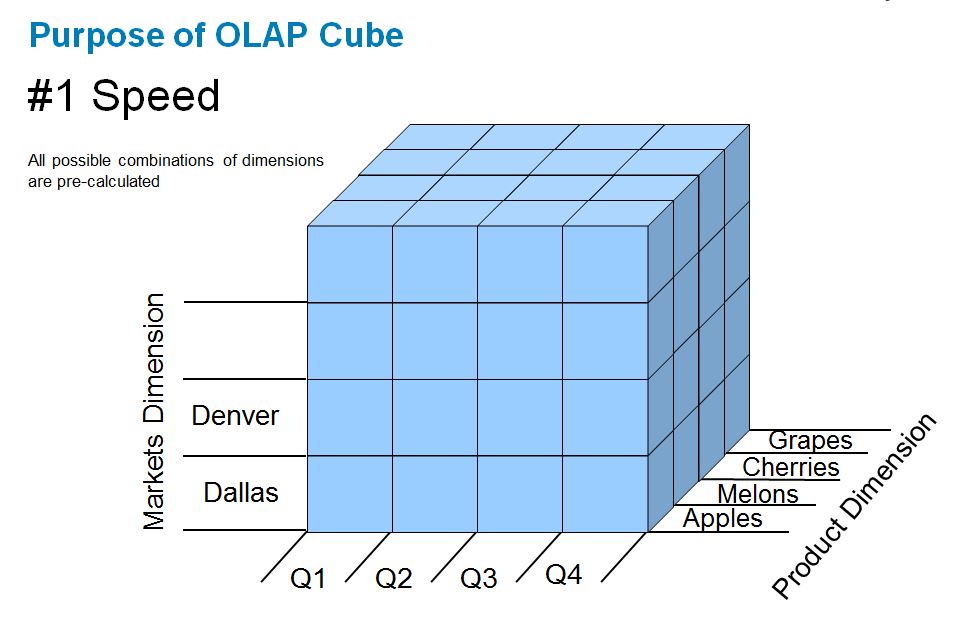


## Key Differences Between ROLAP and MOLAP

1. ROLAP stands for Relational Online Analytical Processing whereas; MOLAP stands for Multidimensional Online Analytical Processing.
2. In both the cases, ROLAP and MOLAP data is stored in the main warehouse. In ROLAP data is directly fetched from the main warehouse whereas, in MOLAP data is fetched from the proprietary databases MDDBs.
3. In ROLAP, data is stored in the form of relational tables but, in MOLAP data is stored in the form of a multidimensional array made of data cubes.
4. ROLAP deals with large volumes of data whereas, MOLAP deals with limited data summaries kept in MDDBs.
5. ROLAP engines use complex SQL to fetch data from the data warehouse. However, MOLAP engine creates prefabricated and precalculated data cubes to present multidimensional view of data to a user and to manage data sparsity in data cubes, MOLAP uses Sparse matrix technology.
6. ROLAP engine creates a multidimensional view of data dynamically whereas, MOLAP statically stores multidimensional view of data in proprietary databases MDDBs for a user to view it from there.

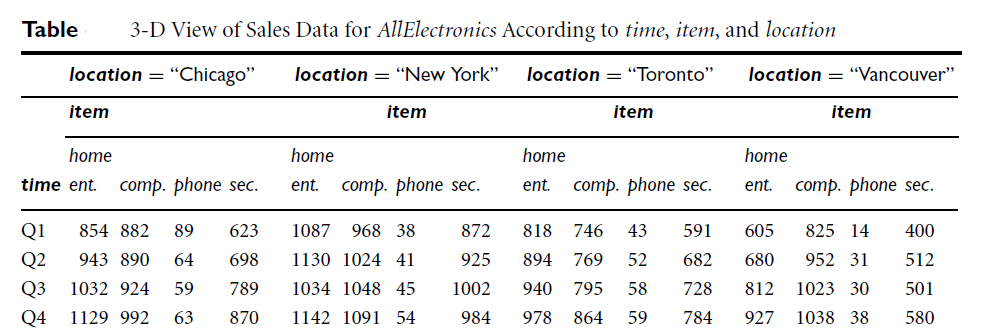
**OLAP Cube or Data Cube is a**[**multidimensional database**](https://searchoracle.techtarget.com/definition/multidimensional-database)**that is optimized for data warehouse and**[**online analytical processing (OLAP)**](https://searchdatamanagement.techtarget.com/definition/OLAP)**applications.**

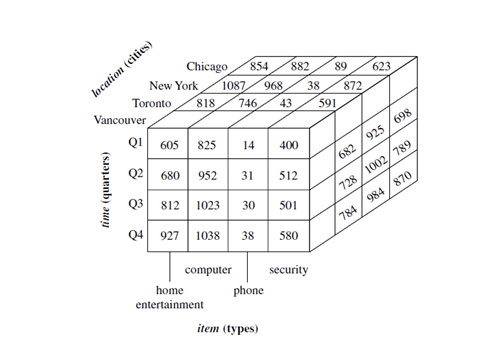
An OLAP cube is a method of storing data in a multidimensional form, generally for reporting purposes. In OLAP cubes, data (measures) are categorized by dimensions. OLAP cubes are often pre-summarized across dimensions to drastically improve query time over relational databases.



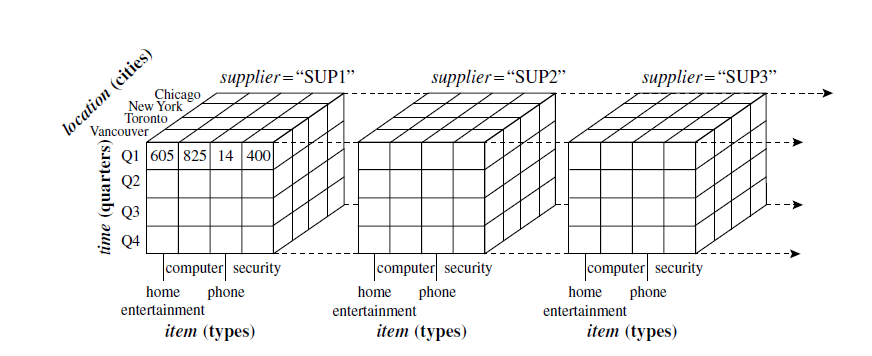
**Note:** The query language used to interact and perform tasks with OLAP cubes is [multidimensional expressions (MDX)](https://searchsqlserver.techtarget.com/definition/multidimensional-expressions-MDX). The MDX language was originally developed by Microsoft in the late 1990s, and has been adopted by many other vendors of multidimensional databases.

**Data Cube Construction.** Depending on the dimension whether it is 2, 3 or 4 table, cube, cuboid, tesseract(4D) can be constructed respectively.

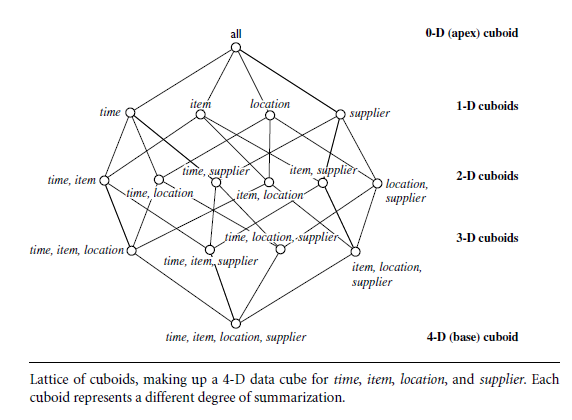




4D representation by projecting cubes. Supplier is mentioned for different regions of sales.

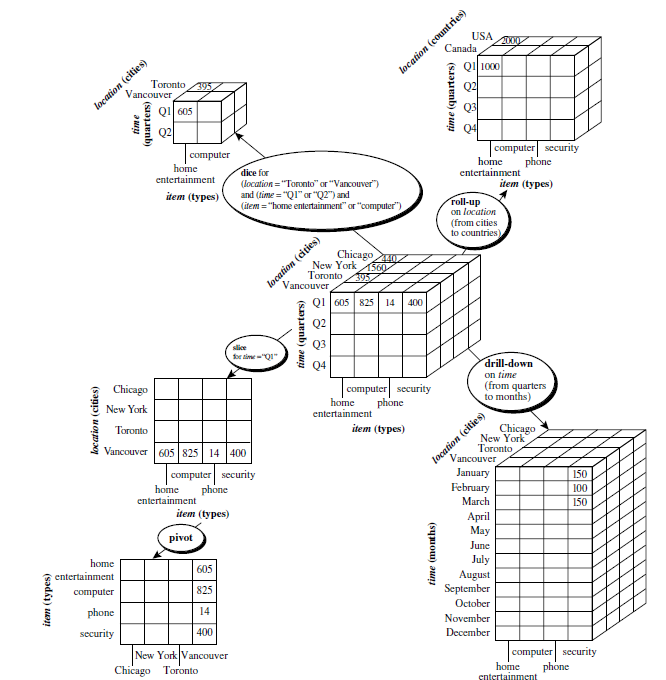


**Lattice of Cuboid**



**OLAP Operations**

In multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. It provides flexibility to view data from different perspectives.



**Roll-up:** The roll-up operation (also called the *drill-up* operation by some vendors) performs aggregation on a data cube, either by *climbing up a concept hierarchy* for a dimension. *country*.” The roll-up operation shown aggregates the data by ascending the *location* hierarchy from the level of *city* to the level of *country*.

**Drill Down:** Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either *stepping down a concept* *hierarchy* for a dimension or *introducing additional dimensions*. Drill-down occurs by descending the *time* hierarchy fromthe level of *quarter* to the more detailed level of *month*.

**Slice:** The *slice* operation performs a selection on one dimension of the given cube, resulting in subcube.

**Dice:** The *dice* operation defines a subcube by performing a selection on two or more dimensions.

**Pivot:**(also called *rotate*) is a visualization operation that rotates the data axes in view to provide an alternative data presentation.

**What is Multidimensional schemas?**

Multidimensional schema is especially designed to model data warehouse systems. The schemas are designed to address the unique needs of very large databases designed for the analytical purpose (OLAP).

Types of Data Warehouse Schema:

Following are 3 chief types of multidimensional schemas each having its unique advantages.

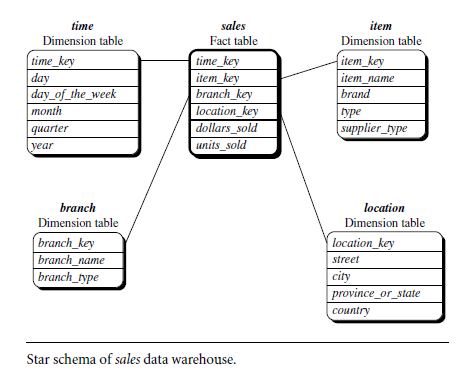
* Star Schema
* Snowflake Schema
* Galaxy Schema

**Star Schema**

It is the simplest style of data warehouse schema. It is called a star schema because the entity relationship diagram of this resembles a star, with points radiating from central table. A star query is join between a fact table and number of dimension tables.

Each dimension table is joined to fact table using primary key to foreign key join but dimension table are not joined together. Fact table contain key and measure.

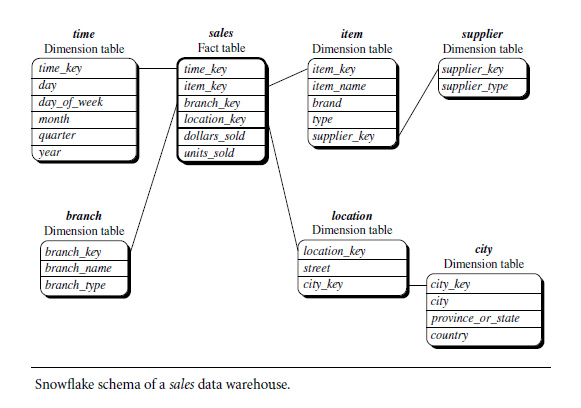
**Ex: Star schema for Electronics sales database.**

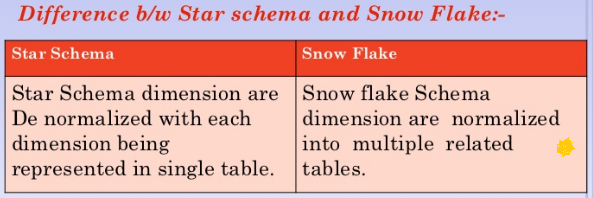


**Snow-flake**

**Snowflake is a variant of star schema where dimension tables are organized into hierarchy of tables by normalizing them which looks like snowflake.**

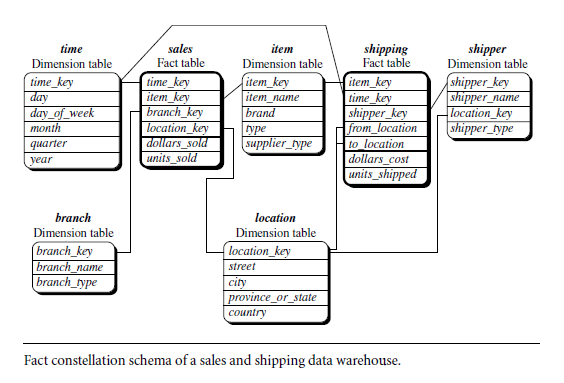
**It is represented by centralized fact table which are connected to multiple dimensions, hierarchy only for dimension tables not for fact table.**





**Fact Constellation**

Sophisticated applications may require multiple fact tables to *share* dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.



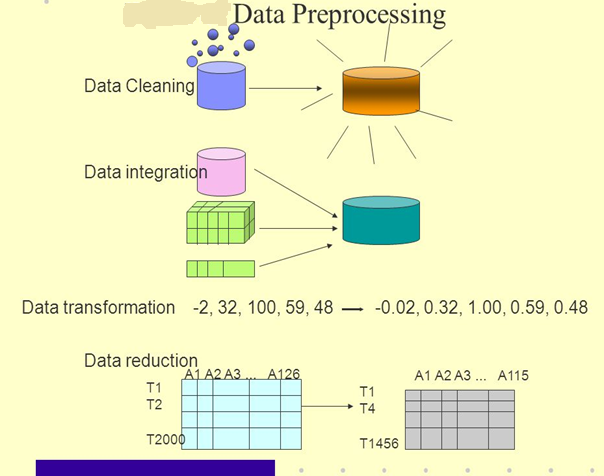
**Note:** Fact constellation is a better alternative of Star schema

**Data Pre-processing / Data preparation**

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Because Data in the real world is dirty.

**Noisy Data:** Unstructured, incomplete, irrelevant, redundant data which result in programming errors, improper analysis sometimes leads to hardware failure are called noisy data.

**Steps in Data Pre-processing**

****

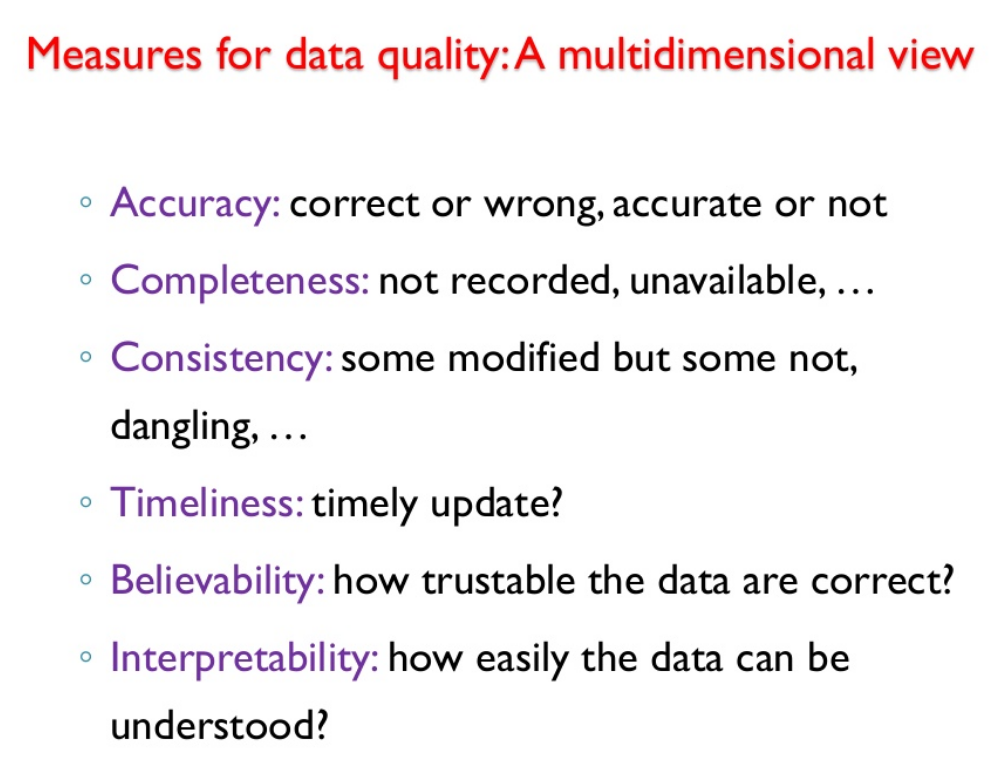
**The Major tasks in Data preprocessing/Data preparation are**

### a) Data Cleaning

**Data cleaning is the process where the data gets cleaned. Data in the real world is normally incomplete, noisy and inconsistent. The data available in data sources might be lacking attribute values, data of interest** etc. **Data cleaning involves a number of techniques including filling in the missing values manually or combined computer and human inspection, etc. The output of data cleaning process is adequately cleaned data.**

**Cleansing:** It is the process of detecting and correcting(removing redundant , irrelevant and making data complete and consistent) dirty data.

**Measures of Data quality:** The data quality can be measured based on some metrics.

****

**Data cleaning as a process:** Missing values, inconsistencies lead to inaccurate data. The first step in data cleaning as a process is Discrepancy detection. Discrepancies can be caused by several factors

* Poorly designed data entry forms
* Human error in data entry

Data should be examined regarding

* **Unique rule**: each value must be different from all other attribute value
* **Consecutive rule**: No missing value between lowest and highest values of the attribute
* **Null rule**: Specifies the use of blank, question marks special characters etc

**Cleaned data:** Accurate , complete, consistent data which is free from noise.

### b) Data Integration

**Data integration is the process where data from different data sources are merged/grouped into one. Data lies in different formats in different locations. Data could be stored in databases, text files, spreadsheets, documents, data cubes, Internet and so on**. Data integration is a really complex and tricky task because data from different sources does not match normally.. Data integration tries to reduce redundancy to the maximum possible level without affecting the reliability of data.

**Benefits:**

It can help to reduce , avoid redundancies and inconsistencies

It improves the accuracy and speed of the subsequent process.

**C) Data Reduction:** It is the process of obtaining a reduced representation of the data much smaller in volume. to select a target data for mining. Strategies for data reduction include.

**i. Data Cube aggregation:** aggregation operations are applied to construction of a cube.

**ii. Attribute Subset selection:** weakly, irrelevant, redundant data identified or removed.

**iii. Dimensionality reduction:** encoding mechanisms or olap operations are used to reduce the data set size.

**iv. Numerosity reduction:** the data are replaced or estimated by alternative , smaller data representations such as:

* **parametric Model:** Graphs, static models
* **Non parametric Model:** clustering, sampling, histogram models.

**D) Data Transformation:** It is the process of transforming and consolidating the data into different forms that are suitable for mining. Data transformation normally involves normalization, aggregation, generalization etc. For example, a data set available as "-5, 37, 100, 89, 78" can be transformed as "-0.05, 0.37, 1.00, 0.89, 0.78". Here data becomes more suitable for data mining. The transformation involves:

* smoothing: removing noise data
* aggregation: summarization , data cube construction
* generalization: Hierarchy of representation of Data
* Normalization: Scaled data to represent the specified range

**Machine Learning**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

**Types of machine Learning**

* **Sepervized Learning**
* **Unsupervized Learning**
* **Semi supervised Learning**

**1.Supervised Machine Learning**

**The majority of practical machine learning uses supervised learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. Y = f(X)The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.**

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process.

Supervised learning problems can be further grouped into **classification** and **outlier analysis**.

Supervised learning is more commonly used in applications where historical data predict future events, such as fraudulent credit card transactions, diagnosing diseases etc.

**2.Unsupervised Machine Learning**

**Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.**

These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped **into clustering** and **association** problems.

Unsupervised learning is more commonly used in applications Basically to associate products, online recommendations, identification of data outliers, and segments, topics etc.

**3.Semi supervised learning**

Semi-supervised learning is a bit of both supervised and unsupervised learning and uses both labeled and unlabeled data for training. In a typical scenario, the algorithm would use a small amount of labeled data with a large amount of unlabeled data.

* **Supervised**: All data is labeled and the algorithms learn to predict the output from the input data.
* **Unsupervised**: All data is unlabeled and the algorithms learn to inherent structure from the input data.
* **Semi-supervised**: Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.

**Pattern Matching:**

**Pattern:** A regular or intelligent form or sequence.

**Dataset:** A Dataset is made up of Objects.

**A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity.**

**Data Object:** A data object represents an entity—in a sales database, the objects may be customers, store items, in sales; in a medical database, the objects may be patients; in a university database, the objects may be students’ professors and courses.

**Attribute:**An attribute is a data field, representing a characteristic or feature of a data object. Attributes describing a customer object can include, for example, customer ID, name, and address

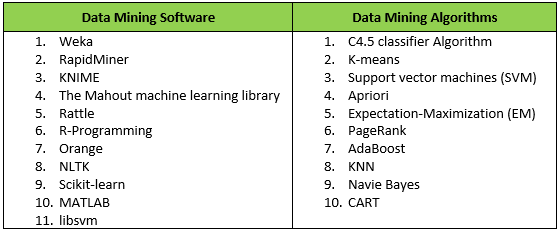
**Interestingness** is an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity.

**Pattern matching** is one of the methods for classification of data, it is used to classify data into predefined groups or classes. It finds the occurrences of a predefined pattern in the data. Applications include speech recognition, Information retrieval, time series analysis etc.

**Frequent pattern:** patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent item sets, frequent subsequences (also known as sequential patterns), and frequent substructures. Frequent patterns are the foundation for many data mining tasks like **Association, Correlation, Clustering, pattern analysis, classification and data warehousing.**

**DataMining Tools / Softwares**

Huge amount of data generated every second and it is necessary to have knowledge of different tools that can be utilized to handle this huge data and apply interesting data mining algorithms and visualizations in quick time. Some of them are listed below.



*1.Weka*



Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

*2.Orange*



Python users playing around with data sciences might be familiar with Orange. It is a Python library that powers Python scripts with its rich compilation of mining and machine learning algorithms for data pre-processing, classification, modelling, regression, clustering and other miscellaneous functions. Orange also comes with a visual programming environment and its workbench consists of tools for importing data, and dragging and dropping widgets and links to connect different widgets for completing the workflow.

*3.R Program*



R is a free software environment for statistical computing and graphics written in C++. R Studio is IDE specially designed for R language.It is one of the leading tools used to do data mining tasks and comes with huge community support as well as packaged with hundreds of libraries built specifically for data mining.