UNIT -1
Data Warehouse

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

Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

Integrated: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

Data Warehouse Design Process:

A data warehouse can be built using a *top-down approach*, a *bottom-up approach*, or a *Combination of both.*

The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood.

The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments.

In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

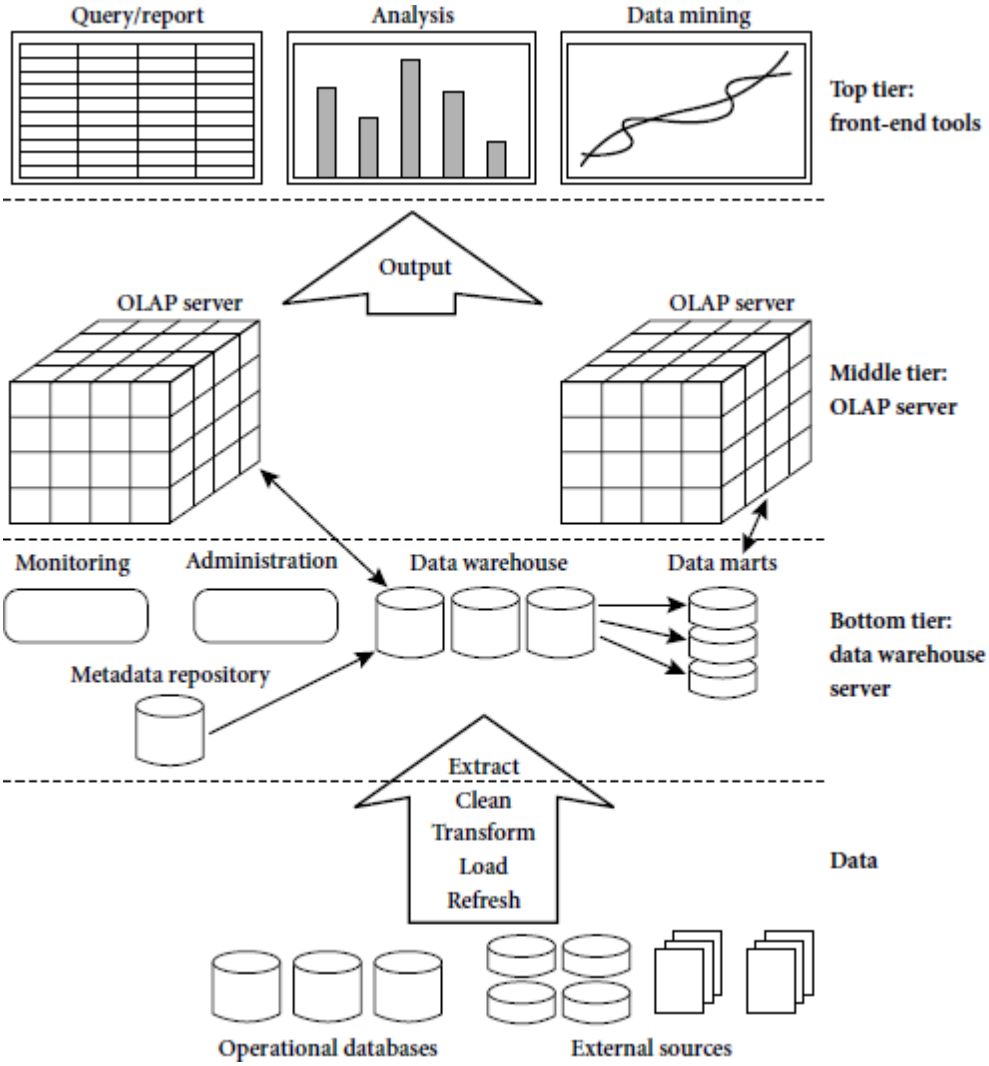
The warehouse design process consists of the following steps:

Choose a business process to model, for example, orders, invoices, shipments, inventory, account administration, sales, or the general ledger. If the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.

Choose the grain of the business process. The grain is the fundamental, atomic level of data to be represented in the fact table for this process, for example, individual transactions, individual daily snapshots, and so on.

Choose the dimensions that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transaction type, and status.

Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.



Tier-1:

The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse . The data are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.

Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Link in gand Embedding for Databases) by Microsoft and JDBC (Java Database Connection).

This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

Tier-2:

The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model or a multidimensional OLAP.

OLAP model is an extended relational DBMS that maps operations on multidimensional data to standard relational operations.

A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

Tier-3:

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

Data Warehouse Models:

There are three data warehouse models.

1. **Enterprise warehouse:**

An enterprise warehouse collects all of the information about subjects spanning the entire organization.

It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.

It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.

An enterprise data warehouse may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms. It requires extensive business modeling and may take years to design and build.

2. **Data mart:**

A data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in data marts tend to be summarized.

Data marts are usually implemented on low-cost departmental servers that are UNIX/LINUX- or Windows-based. The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years. However, it may involve complex integration in the long run if its design and planning were not enterprise-wide.

Depending on the source of data, data marts can be categorized as independent or dependent. Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area. Dependent data marts are sourced directly from enterprise data warehouses.

3. Virtual warehouse:

A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized.

A virtual warehouse is easy to build but requires excess capacity on operational database servers.

4. Meta Data Repository:

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports.

The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control).

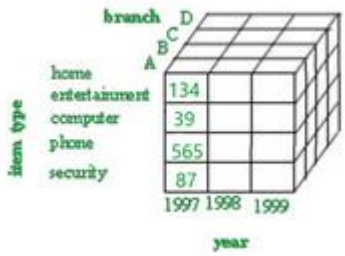
Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

Business metadata, which include business terms and definitions, data ownership information, and charging policies.

What is OLAP?

OLAP stands for Online Analytical Processing, which is a technology that enables multi-dimensional analysis of business data. It provides interactive access to large amounts of data and supports complex calculations and data aggregation. OLAP is used to support business intelligence and decision-making processes.

Grouping of data in a multidimensional matrix is called data cubes. In Dataware housing, we generally deal with various multidimensional data models as the data will be represented by multiple dimensions and multiple attributes. This multidimensional data is represented in the data cube as the cube represents a high-dimensional space. The Data cube pictorially shows how different attributes of data are arranged in the data model. Below is the diagram of a general data cube.



The example above is a 3D cube having attributes like branch(A,B,C,D), item type(home, entertainment, computer, phone, security), year(1997,1998,1999) .

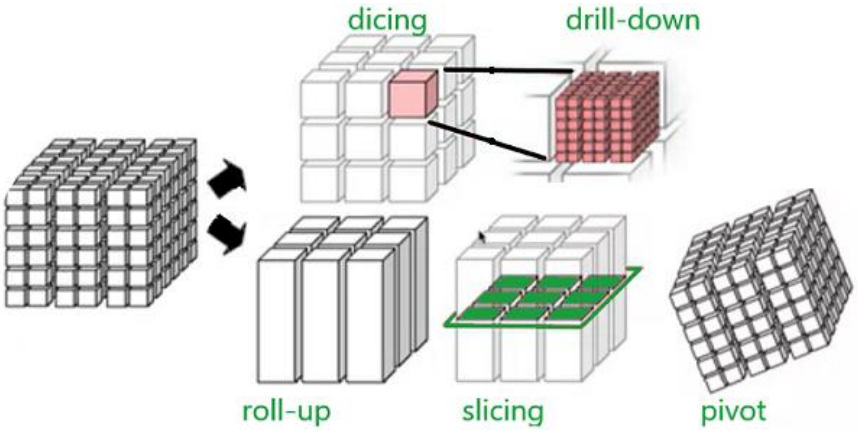
Data cube classification:

The data cube can be classified into two categories:

- **Multidimensional data cube:** It basically helps in storing large amounts of data by making use of a multi-dimensional array. It increases its efficiency by keeping an index of each dimension. Thus, dimensional is able to retrieve data fast.

- **Relational data cube:** It basically helps in storing large amounts of data by making use of relational tables. Each relational table displays the dimensions of the data cube. It is slower compared to a Multidimensional Data Cube.

Data cube operations:



Data cube operations are used to manipulate data to meet the needs of users. These operations help to select particular data for the analysis purpose. There are mainly 5 operations listed below-

- **Roll-up**: operation and aggregate certain similar data attributes having the same dimension together. For example, if the data cube displays the daily income of a customer, we can use a roll-up operation to find the monthly income of his salary.

- **Drill-down**: this operation is the reverse of the roll-up operation. It allows us to take particular information and then subdivide it further for coarser granularity analysis. It zooms into more detail. For example- if India is an attribute of a country column and we wish to see villages in India, then the drill-down operation splits India into states, districts, towns, cities, villages and then displays the required information.

- **Slicing**: this operation filters the unnecessary portions. Suppose in a particular dimension, the user doesn't need everything for analysis, rather a

particular attribute. For example, country="Jamaica", this will display only about Jamaica and only display other countries present on the country list.

- **Dicing**: this operation does a multidimensional cutting, that not only cuts only one dimension but also can go to another dimension and cut a certain range of it. As a result, it looks more like a subcube out of the whole cube(as depicted in the figure). For example- the user wants to see the annual salary of Jharkhand state employees.

- **Pivot**: this operation is very important from a viewing point of view. It basically transforms the data cube in terms of view. It doesn't change the data present in the data cube. For example, if the user is comparing year versus branch, using the pivot operation, the user can change the viewpoint and now compare branch versus item type.

Advantages of data cubes:

- **Multi-dimensional analysis**: Data cubes enable multi-dimensional analysis of business data, allowing users to view data from different perspectives and levels of detail.
- **Interactivity:** Data cubes provide interactive access to large amounts of data, allowing users to easily navigate and manipulate the data to support their analysis.
- **Speed and efficiency**: Data cubes are optimized for OLAP analysis, enabling fast and efficient querying and aggregation of data.
- **Data aggregation:** Data cubes support complex calculations and data aggregation, enabling users to quickly and easily summarize large amounts of data.
- **Improved decision-making:** Data cubes provide a clear and comprehensive view of business data, enabling improved decision-making and business intelligence.
- **Accessibility**: Data cubes can be accessed from a variety of devices and platforms, making it easy for users to access and analyze business data from anywhere.
- Helps in giving a summarized view of data.
- Data cubes store large data in a simple way.
- Data cube operation provides quick and better analysis,
- Improve performance of data.

Disadvantages of data cube:

- **Complexity**: OLAP systems can be complex to set up and maintain, requiring specialized technical expertise.

- **Data size limitations**: OLAP systems can struggle with very large data sets and may require extensive data aggregation or summarization.

- **Performance issues**: OLAP systems can be slow when dealing with large amounts of data, especially when running complex queries or calculations.

- **Data integrity:** Inconsistent data definitions and data quality issues can affect the accuracy of OLAP analysis.

- **Cost:** OLAP technology can be expensive, especially for enterprise-level solutions, due to the need for specialized hardware and software.

- **Inflexibility:** OLAP systems may not easily accommodate changing business needs and may require significant effort to modify or extend.

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information. This chapter cover the types of OLAP, operations on OLAP, difference between OLAP, and statistical databases and OLTP.

Types of OLAP Servers

We have four types of OLAP servers −
- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.
ROLAP includes the following −
- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

Specialized SQL Servers

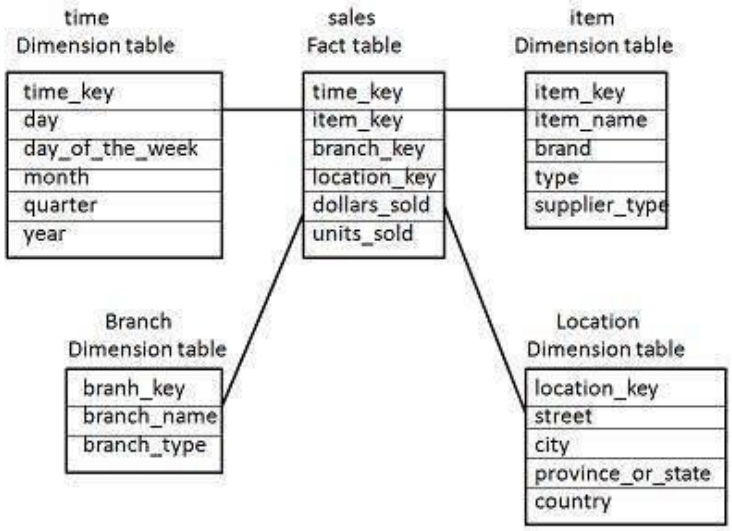
Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

Data Warehousing – Schemas

Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

Star Schema

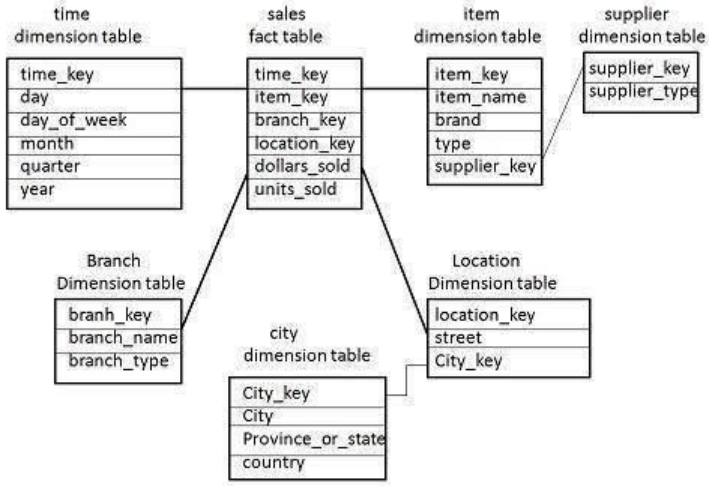
- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.



- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Snowflake Schema

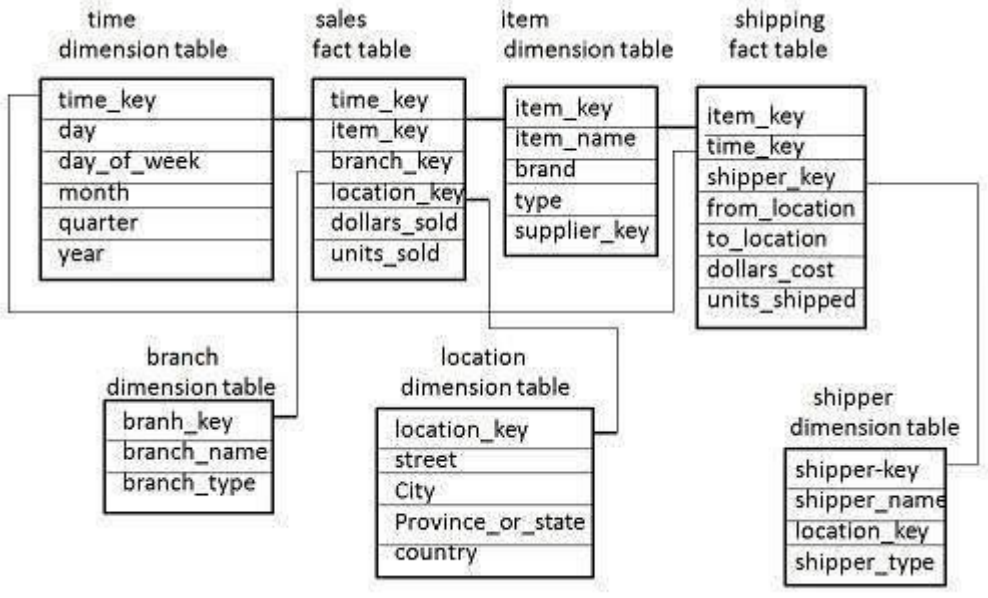
- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.



- Now the item dimension table contains the attributes item key, item name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier key and supplier type.

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

UNIT-2

INTRODUCTION TO DATAMINING

Datamining has the great impact on information industry, and also in society. As, we have huge amount of data, turning that data into useful information and knowledge is nothing but mining.

The information and knowledge gained can be used for applications ranging from market analysis, fraud detection and customer rentention to production control and science exploration.

Datamining can be ~~used~~ viewed as a result of natural evolution of information technology.

The term datamining is evolved from the motto,

" NECESSITY IS THE MOTHER OF INVENTION "

Evolution of database technology is as follows

1960's & Earlier – Data Collection & Database creation.

1970's – early 1980's – Database management Systems.

Mid-1980's to Present – Advanced Database Systems.

- Data can now be stored in many different kinds of databases and information repositories. One data repository architecture emerged is the DATAWAREHOUSE. (A repository of multiple heterogeneous datasources organized Under a Unified schema at a single site inorder to provide decision making).

- Dw technology includes data cleaning, data integration and OLAP. i.e., analysis techniques with functionalities such as summarization, consolidation and aggregation. and also ability to view information from different perceptions.

- The abundant data, coupled with powerful dataanalysis tools describedas datarich, but information Poor.

- Without powerful analytic tools, it is not possible to handle such a huge amount of data.

- Info Data Repositories will become datatombs, if the required information is not ~~date~~ going to be available.

Datamining :— It refers to extracting or mining knowledge from large amounts of data.

If we mine gold from rocks or sand, it is referred as GOLD MINING rather than rock or sand mining. Thus, datamining should have been more appropriately named "knowledge mining from Data", so but they call it as "KNOWLEDGE MINING"

Datamining is also known as knowledge mining from data/ knowledge extraction / data or pattern analysis / Data archeology / data dredging.

Many of them treat datamining as a synonym for another term "KNOWLEDGE DISCOVERY FROM DATA / KDD."

*→ Datamining simply an essential step in the process of KDD.

KDD is an iterative sequence of following steps

1. Data cleaning (To remove noisy and inconsistent data)
2. Data Integration (Where multiple data sources are combined).
3. Data selection (Data i.e., relevant to the analysis task are retrieved from the database).

4. Data Transformation (where data are transformed into forms appropriate for mining by performing operations like summary or aggregation)

5. DataMining (an essential process where intelligent methods are applied inorder to extract data patterns).

6. Pattern Evaluation (Identifying the interesting patterns representing knowledge based on some measures).

7. Knowledge Presentation (Where Visualization and Knowledge representation techniques are used to present the mined knowledge to the user).

→ Steps 1 to 4 – Different forms of Data Preprocessing, where data has been prepared for mining. The DM step may interact with user or knowledge base.

→ The interesting patterns are presented to user and may be stored as new knowledge in the knowledge base.

Datamining is the process of discovering interesting knowledge from large amounts of data stored in db's, datawarehouses or other information repositories.

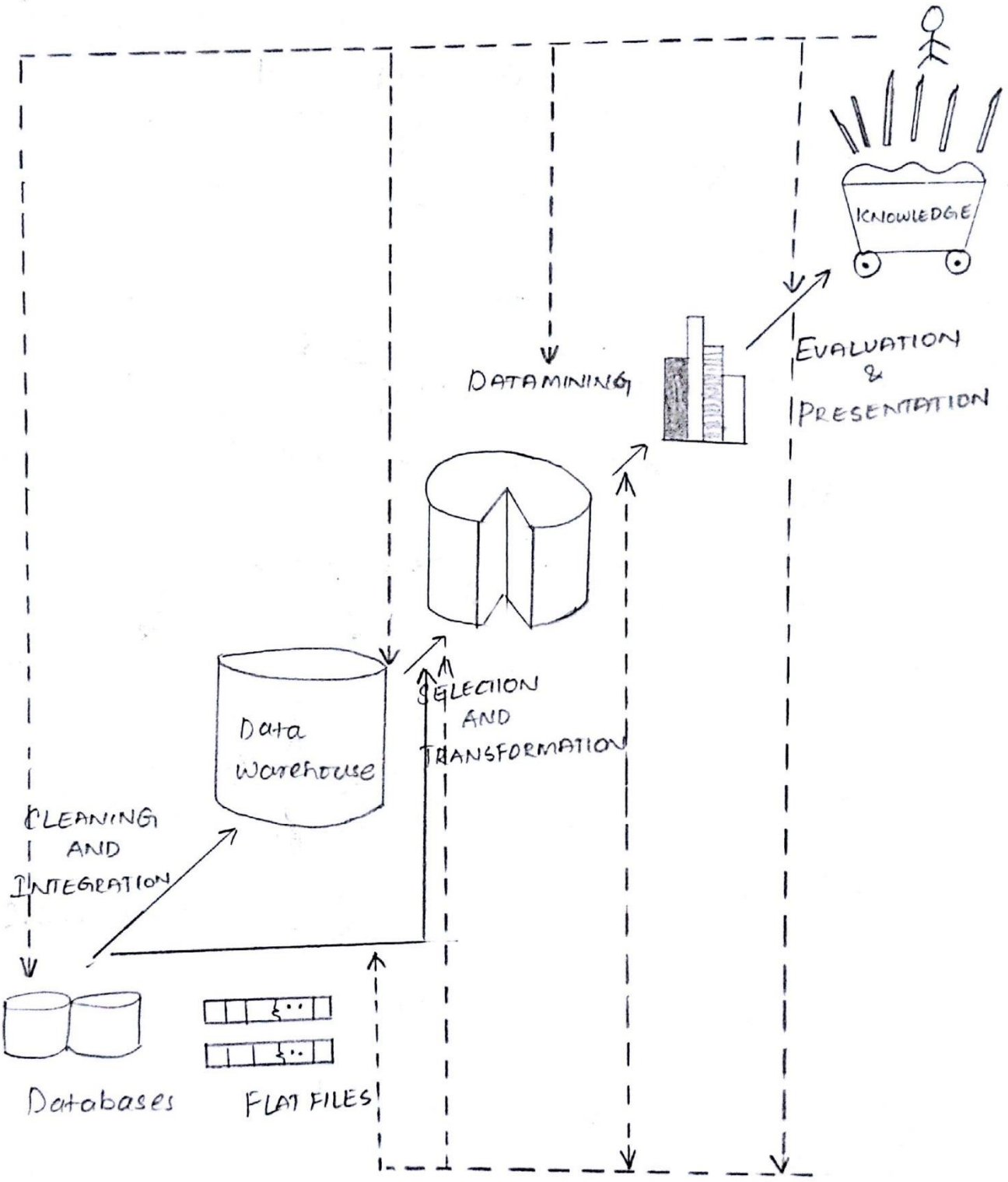


FIG: DATAMINING AS A STEP IN THE PROCESS OF KNOWLEDGE DISCOVERY

The architecture of a typical datamining ~~is~~ may have the following components:

1, Database, Datawarehouse, WWW or Other Information repository :- on the following repositories, data cleaning and data integration techniques are performed.

2, Database or Datawarehouse Server :- It is responsible for fetching relevant data, based on user's datamining request.

3, Knowledge Base :- This is the domain knowledge i.e., used to search or evaluate the interestingness of resultant patterns. Knowledge may contain concept hierarchies, ~~or~~ used to organize attributes or their values.

4, Datamining Engine :- This is the essential part of datamining System, and it have a set of functional modules for tasks such as Charcterization, association and correlational analysis, Classification, Prediction, Cluster-analysis, Outlier analysis, and evolution analysis.

5, Pattern Evaluation :- This component is responsible for interestingness of patterns and interact with datamining modules, so that it can focus search towards interesting patterns.

6, User Interface :- This module communicates between Users and the datamining System, allows the user to interact with System by specifying datamining query or task. ↳ It also allows the User to browse database and datawarehouses, evaluated patterns and visualize the patterns in different forms.

Datamining is Performed on :-

1, Relational Databases (DBMS)

2, Datawarehouses

3, Transactional Databases - Each record represents a transaction.

4, Advanced Data and Information Systems and Advanced applications

5, Object - relational databases.

6, Temporal Databases, Sequencedatabases & Time-series DB

↓, relational data include time-related attributes

↓, Sequence of ordered Events

↓, Stores Sequence of Values obtained over repeated measurement of time

7, Spatial Databases and SpatiotemporalDatabases

↓ Spatial-related Information like geographical maps, ULSI, medical and Satellite imagedb's

↓, If the information present in Spatialdb's are changing wrt time Ex:- Moving vehicles etc.

8, Text Databases and Multimedia Databases

↓ ↓

contains word description of objects. | Stores Image, audio & Video data.

9, Heterogeneous Databases and Legacy Databases

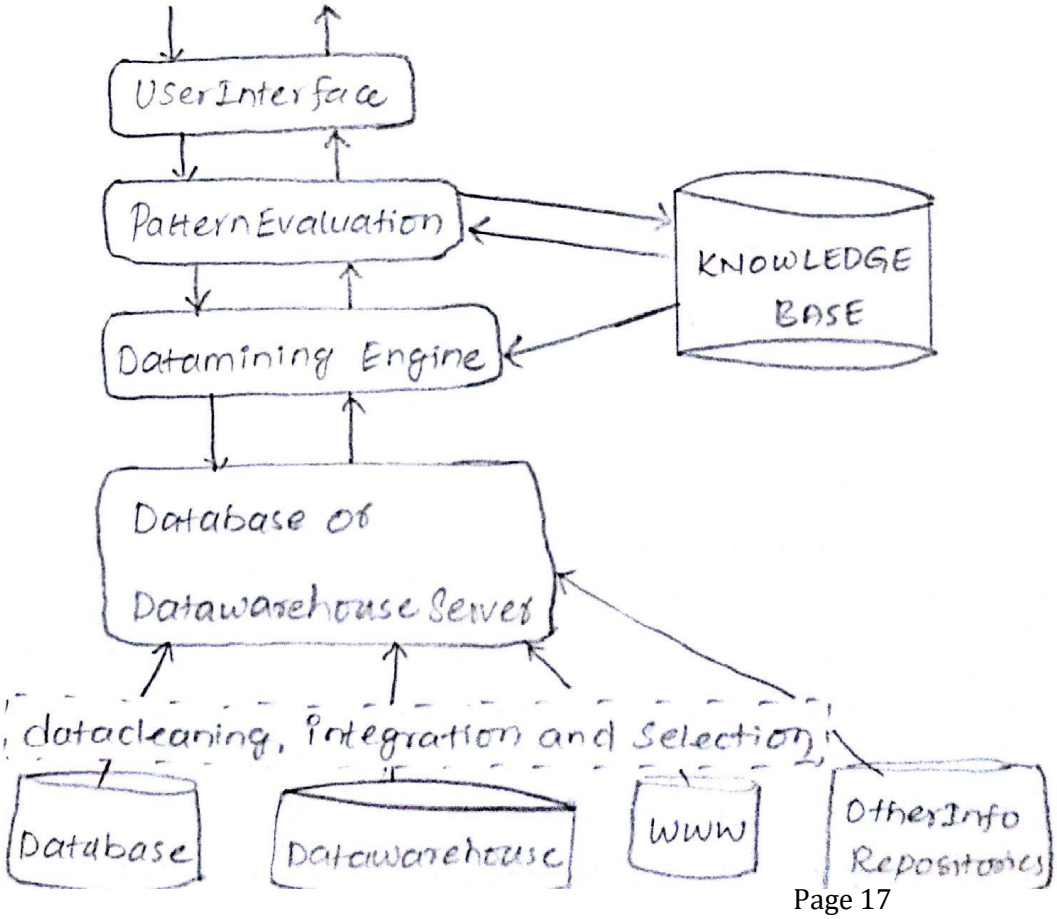
↓ ↓

set of Interconnected & autonomous component db's | It is a group of heterogeneous databases that combines different kinds of data systems.

10, Datastreams:

A Some applications may involve generation and analysis of a new kind of data, called STREAMDATA, where data flow in and out of an window dynamically.

ARCHITECTURE OF A TYPICAL DATA - MINING SYSTEM



Datamining Functionalities:-

What kind of patterns can be mined.

- These are used to find what kind of patterns to be found in data-mining tasks.
- Datamining tasks can be classified into two categories.
  1. Descriptive
  2. Predictive (Classification & Prediction)
     ↓
     Performs inference on current data in order to make prediction.
- Descriptive mining tasks characterize the general properties of data in the database. List of descriptive functions.
  1. Class/Concept Description
  2. Mining of frequent patterns
  3. Mining of Associations
  4. Mining of Correlations
  5. Mining of Clusters.

1. Class/concept Description:- It refers the data to be associated with classes or concepts. For example, in All Electronics store, classes of items for sale include computers, Home Ent, and concepts of customers include Big spenders and Budget Spenders. Such description of class or a concept are called class/concept Descriptions. These descriptions can be derived from following 2 ways.

1, Data Charcterization – This refers to Summarizing data of D.C
class understudy. This class understudy is called as Target class.
O/p of D.C. can be represented as Pie-charts, Barcharts, Curves MD cubes etc.
ii, Data Discrimination – It refers to mapping or association Classification of a
class with some predefined group or class.
↳ Comparison of target class with one/more contrasting classes.

2, Mining of Frequent Patterns :- Frequent patterns are those patterns
that occur frequently in transactional data. Here is the list of
kind of frequent patterns

i, Frequent Item Set : It refers to set of items that frequently
appear together, in a transactional dataset. for example milk and bread.

ii, Frequent Subsequence : A sequence of patterns that occur frequently
such as purchasing a camera is followed by memorycard.
— Sequential Pattern

iii, Frequent Substructure : Substructure refers to different Structural
forms, such as graphs, trees or lattices, which may be combined with
itemsets or subsequences.
If a substructure occurs frequently, it is called a (frequent)
Structured pattern.

3, Mining of Association (Association Analysis) →

→ Mining of frequent patterns leads to discovery of interesting
associations and correlations with in data.

Association Analysis: Suppose, as a marketing manager of All Electronics you would like to determine which Items are frequently purchased together with in the Same transactions.

An example of such a rule, mined from the All Electronics Transaction-al database is,

buys(X, "computer") ⇒ buys(X, "software") [Support = 1%, Confidence = 50%]

X = a variable representing a customer

- Confidence or certainity of 50% means that if a customer buys a computer, there is a 50% chance that he/she will buy software as well.
- A 1% Support means that 1% of all transactions Under analysis Showed that computer and Software were purchased together.

This association rule involves a single attribute or predicate (i.e., buys). Association rules that contain a single attribute are called as Single-dimensional association rules.

Dropping Predicate above rule becomes,

"Computer ⟹ Software [1%, 50%]"

⇒ All Electronics relational ~~Customers~~ database relating to purchases. A datamining System may find association rules like,

$$age(X, "20...29") \wedge income(X, "20K...29K") \Rightarrow buys(X, "CD\ Player")$$

[Support = 2%, Confidence = 60%]

The above mentioned association rule, is having more than One predicate/attribute. Hence each predicate referred to as dimension, The above rule can be referred as Multidimensional Association Rule.

* An association rule is discarded, as Uninteresting, If they donot Satisfy both minimum Support threshold & minimum Confidence threshold.

CLASSIFICATION and PREDICTION :-

* classification is the process of finding a model that describes and distinguishes data classes or concepts. This model is used to predict the class of objects, whose class label is Unknown. The derived model is based on the analysis of set of training data. (i.e., data objects, whose class label is known).

Ex:- Sales mgr — AllElectronics — classify large set of Items based on responses: Good Response, Mild, No

We derive model, based on descriptive features like Price, brand, type etc

— The derived model may be represented as classification rules, decision trees, mathematical formulae or neural networks.

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch denotes an outcome of the test, and tree leaves represent classes. These trees can be easily converted to association rules.

There are many other methods for constructing classification models like Naive-Bayesian classification, Support vector machines and K-nearest neighbor classification.

Classification predicts categorical (discrete, unordered) labels, where as prediction models continous-valued functions. i.e., it is used to predict missing or unavailable numerical data values rather than class labels

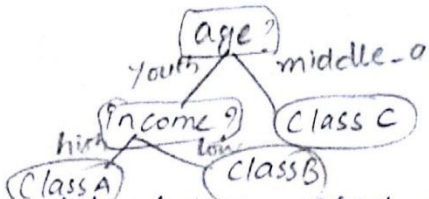
Prediction is useful for both numeric prediction and class label prediction. Regression Analysis is the method i.e., often used for numeric prediction. (IF-THEN) rules, — Classification rules

Ex- age(X, "Youth") AND income(X, "high") → class(X, "A")

age(X, "youth") AND income(X, "low") → class(X, "B")

age(X, "middle-aged") → class(X, "C") age(X, "Senior") → class(X, "C")

CLUSTER ANALYSIS :-



Decision Tree

- Unlike classification and prediction, which analyze class-labeled data objects, cluster analyzes data objects without consulting a known class label.
- Class labels are not present in the training data
- Clustering can be used to generate such labels.
- The objects are clustered based on the principle of "maximizing the intra-class similarity and minimizing the interclass similarity."

OUTLIER ANALYSIS :-

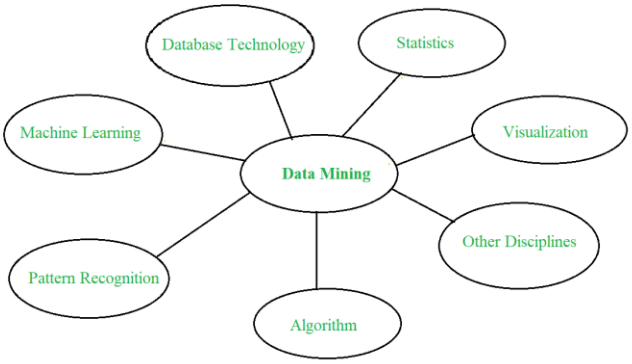
- Database may contain objects that do not comply with the general behavior or model of data. Those data objects are Outliers.
- Most of the datamining methods discard outliers as noisy data/ exceptions
- Useful in Fraud Detection (Things occur rarely).
- The analysis of outlier data, is referred as OUTLIER MINING.

EVOLUTION ANALYSIS :-

- Describes and models regularities for the objects whose behavior change over time. → This include Characterization, discrimination, association, etc.

DATA MINING TECHNOLOGIES

Data mining has incorporated many techniques from other domain fields like machine learning, statistics, information retrieval, data warehouse, pattern recognition, algorithms, and high-performance computing. Since it is a highly application-driven domain, the interdisciplinary nature is typically very significant. Research and development in data mining and its applications prove quite useful in implementing it. We will see major technologies utilized in data mining.



Machine Learning:

It has a main research area that focuses on computer programs that will automatically learn based on the given input data and make intelligent decisions. There are similarities and interrelations between machine learning and data mining. For classification and clustering approaches, machine learning is often applied to predict accuracy. Typical machine learning problems that are utilized in mining are:

1. Supervised learning that makes use of class labels to predict information
2. Unsupervised learning doesn't use class labels similar to clustering but it will discover new classes within data.
3. Semi-supervised learning will redefine the boundaries between two classes and makes use of both labeled and unlabeled examples.
4. Active learning will ask the user to label the classes that may be from unlabeled examples. It will optimize learning by acquiring data from the user.

Information Retrieval:

The technique searches for the information in the document, which may be in text, multimedia, or residing on the Web. It has two main characteristics:

1. Searched data is unstructured
2. Queries are formed by keywords that don't have complex structures.

The most widely used information retrieval approach is the probabilistic model. Information retrieval combined with data mining techniques is used for finding out any relevant topic in the document or web.

Uses: A large amount of data are available and streamed in the web, both text and multimedia due to the fast growth of digitalization including the government sector, health care, and many others.

The search and analysis have raised many challenges and hence Information Retrieval becomes increasingly important.

Statistics:

Data mining has an inherent connection with statistics. It studies the collection, and interpretation performs the analysis and helps visualize data presentation. A statistical model is used for data classes and data modeling. It describes the behavior of an object in a class and its probability. Statistical models are the outcomes of data mining tasks like classification and data characterization. Or we can use the mining task on top of the statistical models.

Advantage:

- Statistics can be used to model noise and missing data values. The tools for forecasting, predicting, or summarizing data can be availed by statistics. Statistics are useful for pattern mining. After mining a classification model, the statistical hypothesis is used for verification. A hypothetical test makes the decisions using the test data. The result is statistically significant if it is not likely to have been incurred by chance.

Disadvantage:

- When the statistical model is used on large data set, it increases the complexity cost. When data mining is used to handle large real-time and streamed data, computation costs increase dramatically.

Database System & Data warehouse:

Database systems are used in query languages, query processing, optimization, and data models. Recent database system data analytics capabilities that use data mining and warehousing techniques. Data warehousing combines data from multiple sources (heterogeneous) and gathers historical data in various timeframes. It facilitates data cubes in a multidimensional database. The OLAP facilitates a multi-dimensional database. The data mining task is used to extend the existing requirement of the database system that would enhance the capabilities and enhance users' sophisticated requirements

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. In this tutorial, we will discuss the applications and the trend of data mining.

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows –

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry –

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services –

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very

important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis –

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications –

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection –

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Data Mining System Products

There are many data mining system products and domain specific data mining applications. The new data mining systems and applications are being added to the previous systems. Also, efforts are being made to standardize data mining languages.