Unit IV: Data Warehouse and Data Mining - Management Information System

Preview of Introduction:

Data:

Data can be described as unprocessed facts and figures. Plain collected data as raw facts cannot help in decision making. However, data is the raw material that is organized, structured, and interpreted to create useful information systems. Data is defined as 'groups of non-random symbols in the form of text, images, voice representing quantities, action and objects.'

Information:

Information is interpreted data, created from organized, structured and processed data in a particular context. The term data and information are closely related. Data are raw material resources that are processed into finished information products. In practice, the database today may contain either data or information.

According to Davis and Olson, "Information is a data that has been processed into a form that is meaningful to the recipient and is of real or perceived value in the current or the prospective action or decision of recipient."

Types of Data:

Data is the raw material from which useful information is derived. The word data is the plural of datum. Data is commonly used in both singular and plural forms. It is defined as raw facts or observations.

It takes a variety of forms, including numeric data, text and voice and images. Data is a collection of facts, which is unorganized but can be made organized into useful information. Hence types of data can be classified as:

1. Variety of forms
2. Numeric data
3. Text data
4. Voice data
5. Images data

Field:

Every field in a table has properties. These properties define the field's characteristics and behavior. The most important property for a field is its data type. A field's data type determines what kind of data it can store.

For example, a field whose data type is Text can store data that consists of either text or numerical characters, but a field whose data type is Number can store only numerical data. A field's data type determines many other important field qualities, such as:

1. The maximum size of a field value.
2. Whether the filed can be indexed.
3. Which formats can be used with the field.

When we create a new field in design view, we specify the field's data type and optionally its other properties.

Record:

In computer data processing, a record is a collection of data items arranged for processing by a program. Multiple records are contained in a file or data set.

The organization of data in the record is usually prescribed by the programming language that defines the record's organization and/or by the application that processes it. Typically, records can be of fixed length or be of variable length with the length information contained within the record.

In a database, a record is a group of a field within a table that is relevant to a specific entity. For example, in a table called customer contact information, a row would likely contain fields such as ID number, name, street address, city, telephone number and so on.

Table:

A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often de-normalized. A fact table works with dimension tables.

A fact table holds the data to be analyzed and a dimension table stores data about how the data in the fact table can be analyzed.

Thus, the fact table consists of two types of columns. The foreign keys column allows joins with dimension tables and the measure columns contain the data that is being analyzed. Suppose that a company sells products to customers. Every sale is a fact that happens and the fact table is used to record these facts.

File System:

Traditionally, file processing system was used to manage information. It stores data in various files of different application programs to extract or insert data to the appropriate file.

File processing system has several drawbacks due to which database management system is required.

File is a collection of information that is used to organize and access information but in this, the information is allocated in an unstructured and non-relation manner so that there is some limitation as compared to the database. Database management system removes problems found in the file processing system.

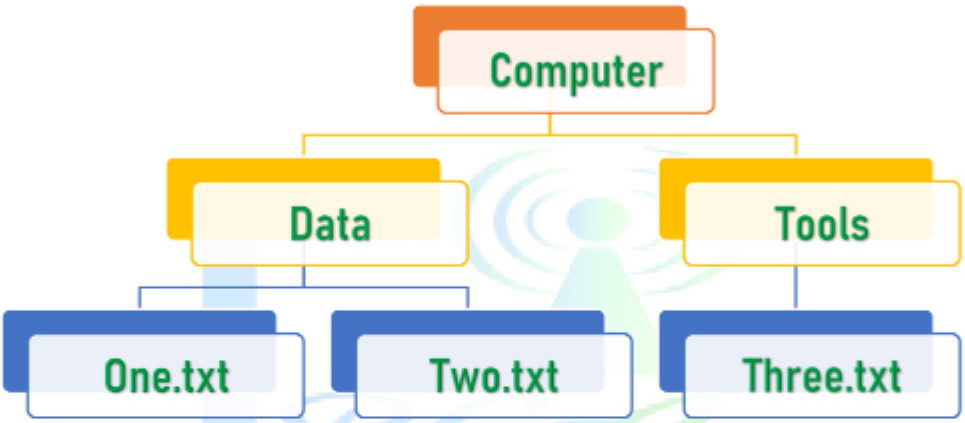


Fig: File System

Problems of File Processing Systems:

1. Data Redundancy and Inconsistency:

In file processing system, the different programmer creates files and writes application programs to access it. After a long period files may exist with different formats and application programs may be written in many different programming languages.

Moreover, the same information may be duplicated in several files. We have to pay for higher storage and access cost for such redundancy.

It may leads database in an inconsistent state because update made in one file may reflected in one file but it may not be reflected in another file where same information exists in another file.

2. Difficulty in Accessing Data:

In the file processing system, we cannot easily access required data stored in a particular file. For each new task, we have to write a new application program. The file processing system cannot allow data to be retrieved conveniently and efficiently.

3. Data Isolation:

Since data are scatter in different files and data may store in different format, so it is difficult to write program to retrieve appropriate data.

4. Integrity Problem:

In database, we required to enforce certain type consistency constraints to ensure the database correctness or to enforce certain business rules. It is in fact called integrity constraints, the integrity of database need not be violated.

In file processing system integrity constraint becomes the part of the application program. The programmer needs to write the appropriate code to enforce it. When new constraints are required to add or change the existing one, it is difficult to change the program to enforce it.

5. Atomicity Problem:

Failures may lead the database in an inconsistent state with partial updates. For example, failure occurs while transferring fund from account A to B. There would be the case that certain amount from account A is retrieved and it is updated but failure occurs just before it is deposited to Account B, such case may lead database in an inconsistent state.

6. Concurrent Access Problem:

Concurrent accessed increase the overall performance of the system providing a fast response time but uncontrolled concurrent accesses can lead inconsistencies in the system.

File processing system allows concurrent access but it is unable to coordinate different application programs so database may lead in an inconsistent state. Example: Two people reading a balance and updating it at the same time.

7. Security Problems:

Since file processing system consists of a large number of applications programs and it is added in ad hoc manner. So, it is difficult to enforce security to each application to allow accessing only part of data/database for the individual database users.

Database:

A database is a collection of information that is used to organize and access information according to the logical structure of that information. A database supports both Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP).

Fig: Database

Database that supports OLTP is known as an operational database. A database management system has three components:

1. Data Definition Language (DDL)
2. Data Manipulation Language (DML)
3. Data Dictionary

The data definition language is the formal language, programmers use to specify the structure of the content of the database. The data definition language defines each data element as it appears in the database before that data element is translated into the forms required by application programs.

Most DBMS, have a specialized language called a data manipulation language that is used in conjunction with some conventional third or fourth-generation programming languages to manipulate the data in the database.

This language contains commands that permit end users and programming specialists to extract data from the database to satisfy information requests and develop applications. The most prominent data manipulation language today is Structured Query Language (SQL).

The third element of a DBMS is a data dictionary. This is an automated or manual file that stores definitions of data elements and data characteristics, such as usage, physical representation, ownership (who in the organization is responsible for maintaining the data), authorization and security.

Many data dictionaries can produce lists and reports of data use, groupings, and program locations and so on.

Objectives of Database Approach:

Traditionally data was organized in file formats. DBMS was all new concepts then and all the research was done to make it overcome all the deficiencies in the traditional style of data management. The modern database has the following objectives:

1. Real World Entity:

Modern DBMS is more realistic and users real-world entities to design its architecture. It uses behavior and attributes too. For example, a school database may use students as an entity and their age as their attribute.

2. Relation Based Tables:

DBMS allows entities and relations among them to form as tables. This eases the concept of data saving. A user can understand the architecture of the database just by looking at table name, etc.

3. Isolation of Data and Application:

A database system is entirely different than its data, where the database is said to the active entity, data is said to be passive one on which the database works and organizes. DBMS also stores metadata which is data about data, to ease its own process.

4. Less Redundancy:

DBMS follows rules of normalization, which splits a relation when any of its attributes is having redundancy in values. Following normalization, which itself is the mathematically rich and scientific process, make the entire database to contain as less redundancy as possible.

5. Consistency:

DBMS always enjoy the state on consistency where the previous form of data storing applications like file processing does not guarantee this. Consistency is a state where every relation in the database remains consistent. There exist methods and techniques, which can detect the attempt of leaving the database in an inconsistent state.

6. Query Language:

DBMS is equipped with a query language, which makes it more efficient to retrieve and manipulate data. A user can apply as many and different filtering options, as he or she wants. Traditionally it was not possible where a file processing system was used.

7. ACID Properties:

DBMS follows the concepts for ACID properties, which stands for Atomicity, Consistency, Isolation and Durability. These concepts are applied to transactions, which manipulate data in the database. ACID properties maintain database in a healthy state in the multi-transactional environment and in case of failure.

8. Multiuser and Concurrent Access:

DBMS support multi-user environment and allows them to access and manipulate data in parallel. Though there are restrictions on transactions when they attempt to handle the same data item, but users are always unaware of them.

9. Multiple Views:

DBMS offers multiples views for different users. A user who is in the sales department will have a different view of a database than a person working in the production department. This enables the user to have a concentrated view of the database according to their requirements.

10. Security:

Features like multiple views offer security at some extent where users are unable to access data of other users and departments. DBMS offers methods to impose constraints while entering data into database and retrieving data at a later stage.

DBMS offers many different levels of security features, which enables multiple users to have a different view with different features.

For example a user in the sales department cannot see data of purchase department is one thing, additionally how much data of sales department he can see, can also be managed. Because DBMS is not saved on disk as a traditional file system it is very hard for a thief to break the code.

Database System and Hierarchy:

Data hierarchy refers to the systematic organization of data, often in a hierarchical form. Data organization involves characters, fields, records, files and so on. This concept is a starting point when trying to see what makes up data and whether data has a structure.

For example, how does a person make sense of data such as 'employee', 'name', 'department', 'Marcy Smith', 'Sales Department' and so on, assuming that they are all related?

One way to understand them is to see these terms as smaller or larger components in a hierarchy. One might say that Marcy Smith is one of the employees in the Sales Department or an example of an employee in that Department.

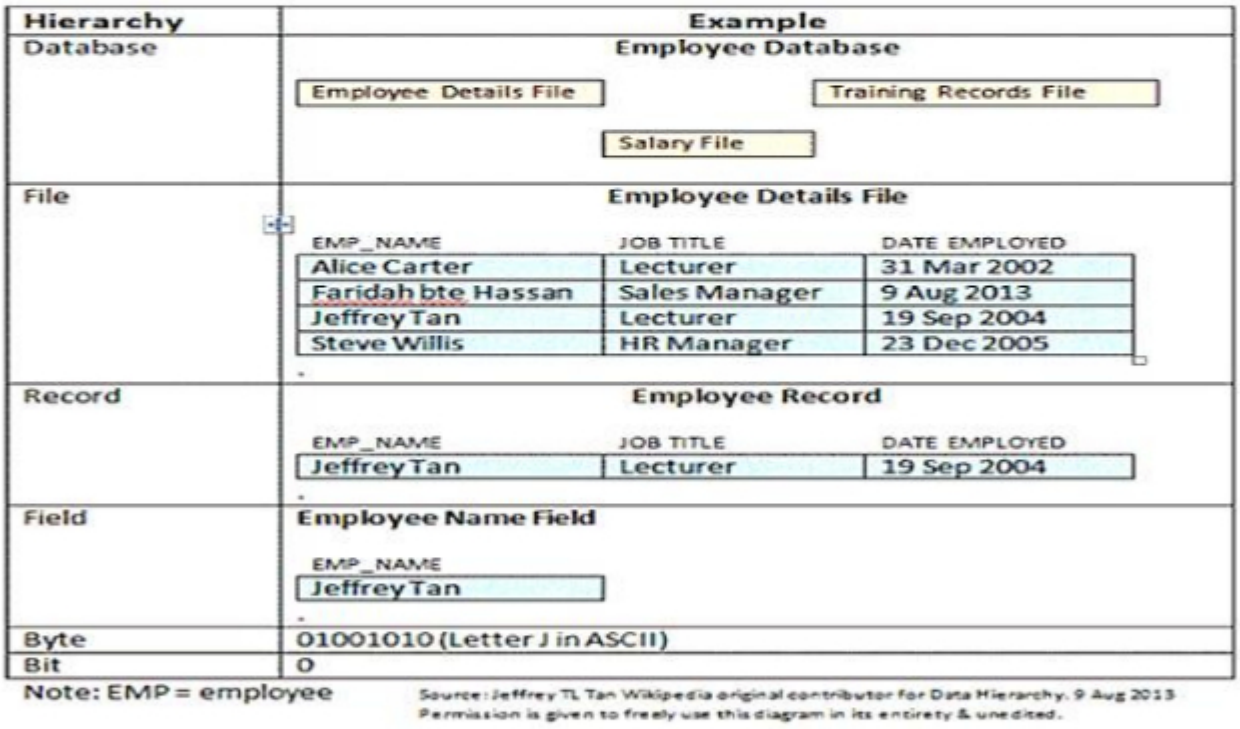
The data we want to capture all our employees, and not just Marcy, is the name, ID number, address etc.

Purpose of the Data Hierarchy:

"Data hierarchy" is a basic concept in data and database theory and helps to show the relationships between smaller and larger components in a database or data file. It is used to give a better sense of understanding about the components of data and how they are related.

Components of the Data Hierarchy:

The components of the data hierarchy are listed below.



Data Hierarchy Diagram – with Employee Database example

Fig: Components of the Data Hierarchy

A **data** field holds a single fact or attribute of an entity. Consider a date field, e.g. "19 September 2004". This can be treated as a single date field (e.g. birth date), or three fields, namely, day of the month, month and year.

A **record** is a collection of related fields. An Employee record may contain a name field(s), address fields, birthdate field and so on.

A **file** is a collection of related records. If there are 100 employees, then each employee would have a record (e.g. called Employee Personal Details record) and the collection of 100 such records would constitute a file (in this case, called Employee Personal Details file).

Files are integrated into a **database**. This is done using a Database Management System. If there are other facets of employee data that we wish to capture, then other files such as Employee Training History file and Employee Work History file could be created as well.

Types of Database Models:

A database model shows the logical structure of a database, including the relationships and constraints that determine how data can be stored and accessed. Individual database models are designed based on the rules and concepts of whichever broader data model the designers adopt. Most data models can be represented by an accompanying database diagram.

1. Hierarchical Database Model:

This is the oldest form of a database. This data model organizes the data in the tree structure i.e. each child node can have only one parent node and at the top of the structure, there is a single parenthesis node.

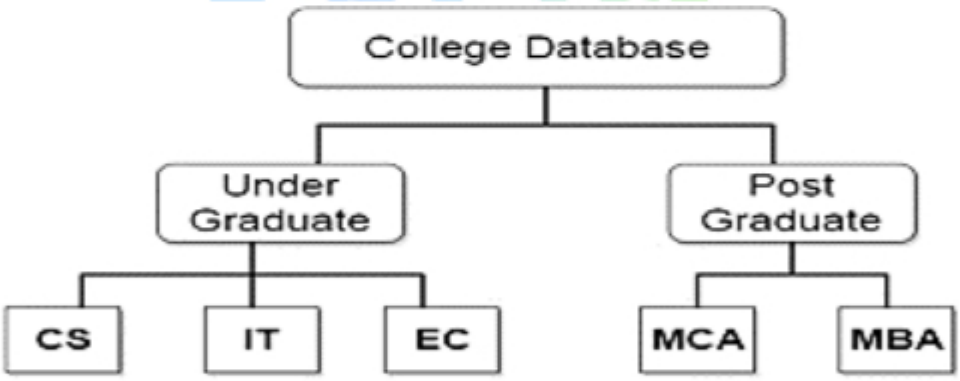


Fig: Hierarchical Database Model

In this model a database record is a tree that consists of one or more groupings of fields called segments, which make up the individual nodes of the tree. This model uses a **one-to-many relationship**.

Advantage: Data access is quite predictable in structure and retrieval and updates can be highly optimized by a DBMS.

Disadvantage: The link is permanently established and cannot be modified which makes this model rigid.

2. Network Database Model:

The Network database model was developed as an alternative to the hierarchical database. This model expands on the hierarchical model by providing multiple paths among segments i.e. more than one parent-child relationship. Hence this model allows **one-to-one, one-to-many and many-to-many relationships.**

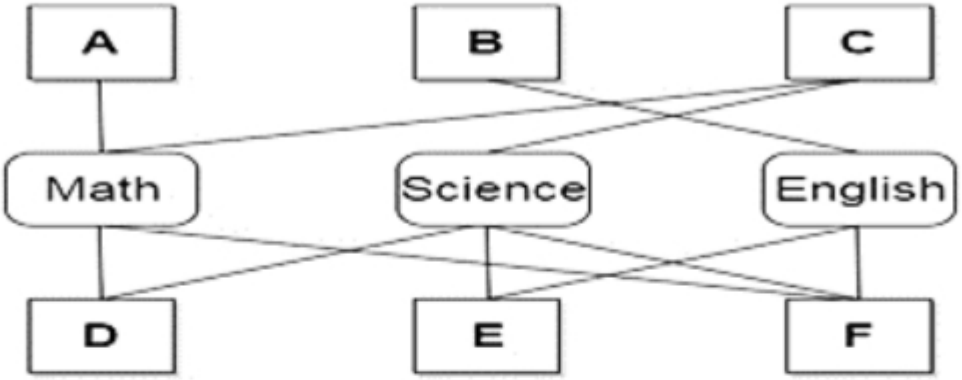


Fig: Network Database Model

Supporting multiple paths in the data structure eliminate some of the drawbacks of the hierarchical model, the network model is not very practical.

Disadvantage: It can be quite complicated to maintain all the links.

3. Relational Database Model:

The key differences between previous database models and relational database model is in terms of flexibility. A relational database represents all data in the database as simple two-dimensional tables called relations.

Each row of a relational table, called tuple, represents a data entity with columns of the table representing attributes (fields). The allowable values for these attributes are called the **domain**.

Student Table

Student_ID	Student_Name
101	Shubham
102	Rajat

Course Table

Course_ID	Course_Name
14	Java
16	Android

College Table

Batch_Year	Student_ID	Course_ID	Teacher_Name	Teacher_Number
2012-16	101	14	Jack	9876543
2013-17	102	16	Tom	9823451

Fig: Relational Database Model

Each row in a relational table must have a unique primary key and also has some secondary keys which correspond with primary keys in other tables

Advantage: Provides flexibility that allows changes to the database structure to be easily accommodated. It facilitates multiple views of the same database for different users.

For example: COLLEGE table has Batch_Year as primary key and has secondary keys Student_ID and Course_ID, these keys serve as primary keys for STUDENT and COURSE tables.

4. Object-Oriented Database Model:

The relational database model has a wide variety of applications. However, it does not easily support the distribution of one database across several servers. Due to this, the object-oriented database management system was developed.

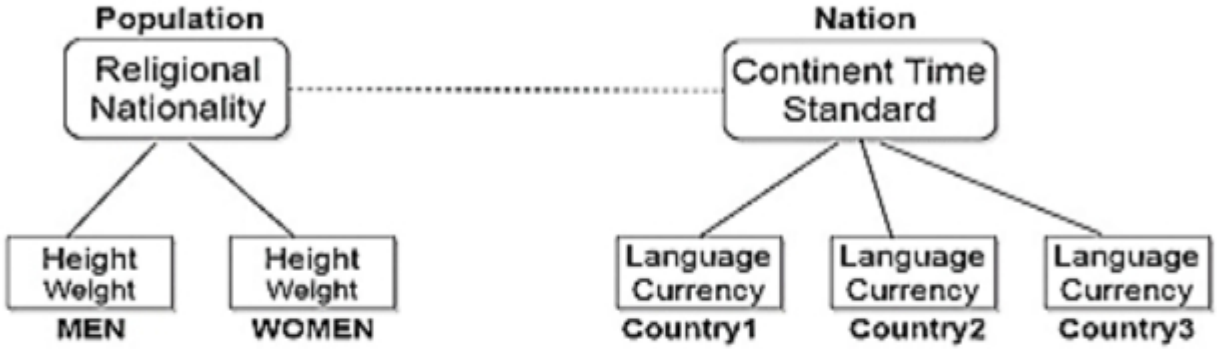


Fig: Object-Oriented Database Model

In these databases, the users can define own data access methods, the representation of data and the method of manipulating it. An object-oriented database stores and maintains objects.

Example: The class population is the root of class hierarchy, which includes the Nation class. The Population class is also the root of two sub-class, men and women. The Nation class is the root of other sub-classes country1, country2, country3. Note that each class has its own set of attributes apart from the root class's attributes.

5. Multidimensional Database (MDB):

A multidimensional database (MDB) is a type of database that is optimized for data warehouse and online analytical processing (OLAP) applications. Multidimensional databases are frequently created using input from existing relational databases.

Whereas a relational database is typically accessed using a Structured Query Language (SQL) query, a multidimensional database allows a user to ask questions like "How many Aptiva's have been sold in Nebraska so far this year?" and similar questions related to summarizing business operations and trends.

An OLAP application that accesses data from the multidimensional database is known as a MOLAP (multidimensional OLAP) application.

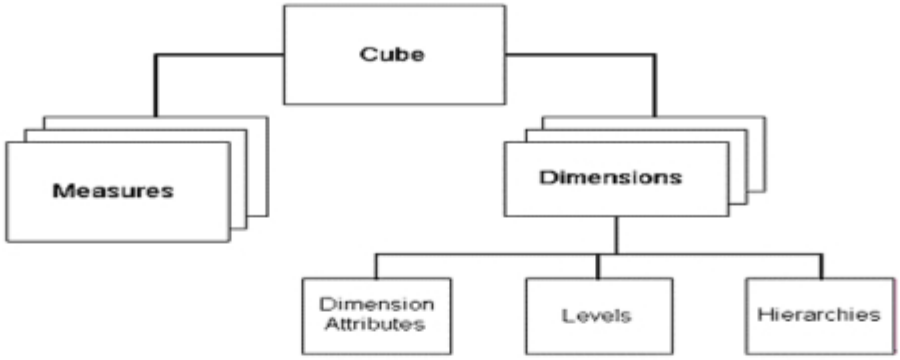


Fig: Multidimensional Database Model

A multidimensional database - or a multidimensional database management system (MDDBMS) - implies the ability to rapidly process the data in the database so that answers can be generated quickly.

Several vendors provide products that use multidimensional databases. Approaches to how data is stored and the user interface vary.

Conceptually, a multidimensional database uses the idea of a data cube to represent the dimensions of data available to a user. For example, "sales" could be viewed in the dimensions of the product model, geography, time, or some additional dimension.

In this case, "sales" is known as the measure attribute of the data cube and the other dimensions are seen as feature attributes. Additionally, a database creator can define hierarchies and levels within a dimension (for example, state and city levels within a regional hierarchy).

Database Management System:

A database management system (DBMS) is software that permits an organization to centralize data, manage them efficiently and provide access to the stored data by application programs.

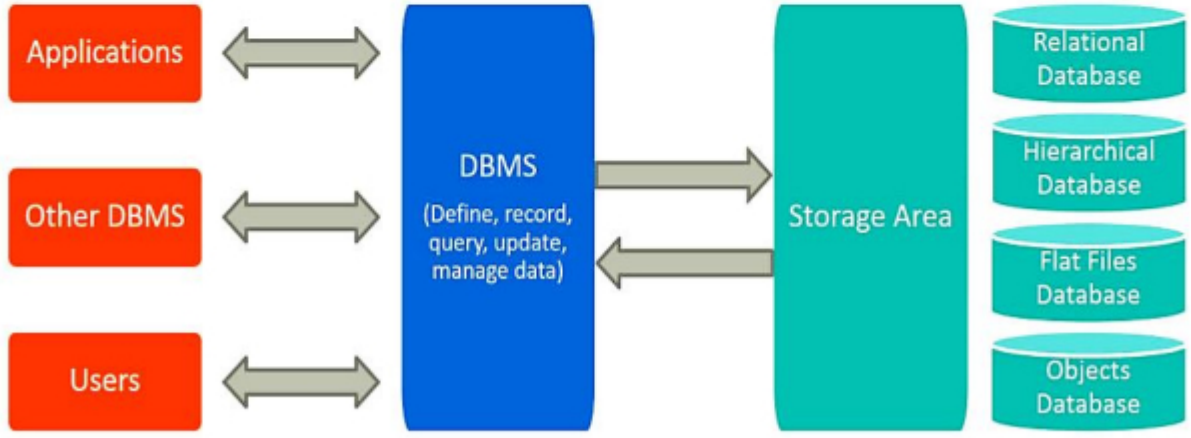


Fig: Database Management System

The DBMS acts as an interface between the application programs and the physical data files. When the application program calls for a data item, the DBMS finds this item in the database and presents it to the application program.

The DBMS relieves the programmer or end-user from the task of understanding where and how the data are actually stored by separating the logical and physical views of the data.

The logical view presents data as they would be perceived by end-users of business specialists, whereas the physical view shows how data are actually organized and structured on physical storage media.

Database Management System Tools:

A DBMS contains five important tools or components:

Fig: Database Management System Tools

1. DBMS Engine:

Accept logical requests from the various other DBMS subsystems convert them into their physical equivalent and actually accesses the database and dictionary as they exist on a storage device.

2. Data Definition Subsystem:

Helps to create and maintain the data dictionary and define the structure of the files in a database.

3. Data Manipulation Subsystem:

Helps to add, change and delete information in a database and mine it for valuable information.

4. Application Generation System:

Contains facilities to help to develop a transaction intensive applications.

5. Data Administration Subsystem:

Help to manage the overall database environment by providing facilities for backup and recovery, security management, query optimization, concurrency control and change management.

Data Repository:

The term "data repository" is often used interchangeably with a data warehouse or a data mart. It is a more generalized term, favored when the specific type of data storage entity is not known or is irrelevant to the context.

The purpose of a data repository is to keep a certain population of data isolated so that it can be mined for greater insight or business intelligence or to be used for a specific reporting need.

Data repository is an integral part of a data warehouse system. It contains the following metadata:

1. **Business Metadata:** It contains the data ownership information, business definition, and changing policies.
2. **Operational Metadata:** It includes currency of data and data lineage. The currency of data refers to the data being active, archived or purged. Lineage of data means a history of data migrated and transformation applied to it.
3. **Data for Mapping from Operational Environment to Data Warehouse:** It includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.
4. **The Algorithms for Summarization:** It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

Examples of Data Repositories:

The term data repository can be used to describe several ways to collect and store data:

1. A **data warehouse** is a large data repository that aggregates data usually from multiple sources or segments of a business, without the data being necessarily related.

2. A **data lake** is a large data repository that stores unstructured data that is classified and tagged with metadata.

3. **Data marts** are subsets of the data repository. These data marts are more targeted to what the data user needs and easier to use. Data marts also are more secure because they limit authorized users to isolated data sets. Those users cannot access all the data in the data repository.

4. **Metadata repositories** store data about data and databases. The metadata explains where the data source, how it was captured, and what it represents.

5. **Data cubes** are lists of data with three or more dimensions stored as a table - as we may find in a spreadsheet.

Benefits of Data Repositories:

There is value to storing and analyzing data. Businesses can make decisions based upon more than anecdote and instinct. However, using data repositories as part of data management is another level of investment that can improve business decisions, such as:

1. Isolation allows for easier and faster data reporting or analysis because the data is clustered together.
2. Database administrators have easier time tracking problems because data repositories are compartmentalized.
3. Data is preserved and archived.

Data Warehouse:

A data warehouse is a repository of information gathered from multiple sources and stored under a unified schema at a single site. So, a data warehouse is the database that stores current, historical and external data of potential interest to decision-makers through the company.

The data originate in many operational transactional systems such as a system for sales, customer accounts, marketing information, manufacturing information, etc. and may include data from website transactions also.

The data warehouse consolidates and standardizes information from different operational databases so that the information can be used across the enterprise for management analysis and decision making.

In relational database model information is represented in a series of two-dimensional tables or files, but in a data warehouse, most of the data warehouses are multi-dimensional. The data warehouse is not transaction-oriented. They exist to support the decision making task in an organization.

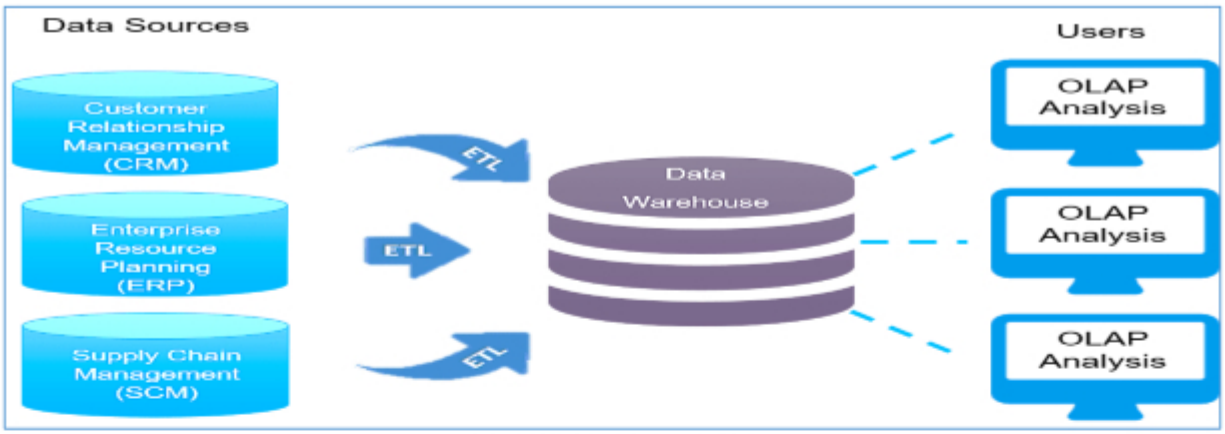


Fig: Data Warehouse

Key Characteristics of a Data Warehouse:

1. Multidimensional:

In the relational database model, information is represented in a series of two-dimensional tables but in a data warehouse they are multi-dimensional means that they contain layers of columns and rows. The layers in a data warehouse represent information according to different dimensions. This multidimensional representation of information is referred to as hypercube.

2. Support Decision Making:

Data warehouses support decision making because they contain summarized information, support business activities and decision-making tasks, not a transaction processing.

3. Subject Oriented:

A data warehouse is subject-oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc.

4. Integrated:

A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.

5. Time-Variant:

The data collected in a data warehouse is identified with a particular period. The data in a data warehouse provides information from the historical point of view.

6. Non-volatile:

Non-volatile means the previous data is not erased when new data is added to it. A data ware house is kept separate from the operational database and therefore frequent changes in the operational database are not reflected in the data warehouse.

Types of Data Warehouse:

Information processing, analytical processing and data mining are the three types of data warehouse applications that are discussed below:

1. Information Processing:

A data warehouse allows processing the data stored in it. The data can be processed employing querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.

2. Analytical Processing:

The data warehouse supports the analytical processing of the information stored in it. The data can be analyzed through basic OLAP operations, including slice and dice, drill down, drill up and pivoting.

3. Data Mining:

Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and perfection. These mining results can be presented using visualization tools.

Pros of Data Warehousing:

1. Speedy Data Retrieving:

How many times have we had a piece of information that we wanted to retrieve but forgot where it was placed? Once we input our information into our DW, we will never lose track of this data again. By conducting a quick search, we'll be able to find the statistic and further analyze it without having to waste time in our search.

2. Error Identification & Correction:

Many of us have a natural tendency to skip over dotting the I's and crossing the T's when inputting raw information, but data warehouses help eliminate user oversight. Before loading data, our data warehouse makes it a point to show us inconsistencies and correct them. This is extremely helpful for those who may be careless or hasty when it comes to gathering data.

3. Easy Integration:

Our DW also adds value to operational business applications like CRM systems when the warehouse is successfully integrated. The complexity of its makeup allows a data warehouse to translate the information in a simpler, digestible format to the point where our team members can easily understand what's been placed in front of them.

Cons of Data Warehousing:

1. Time Consuming Preparation:

While a major part of a data warehouse's responsibility is to simplify our business data, most of the work that will have to be done on our part is inputting the raw data. Now, while the job the DW does for us is helpful and extremely convenient, this is the most work we'll have to manually perform, as the DW performs many other functions for us.

2. Difficulty in Compatibility:

Depending on the system we currently have in place, the use of data warehouse technology could likely require a helping hand from an independent BI team. With the intricacies of operating systems, software and programs, it can be difficult for a business owner to figure out how to properly make use of their data warehouse.

Especially since the costs of these tools are investments in our business, we'll want to ensure that our system is working exactly the way we intend it to.

3. Maintenance Costs:

One of the pros and cons of our DW is its ability to consistently update. This is great for the business owner who wants the best and latest features, however, these upgrades don't usually come cheap.

Including regular maintenance for our system, we can expect to shell out more than our initial investment should we want to have the latest technology at our fingertips.

4. Limited Use Due to Confidential Information:

If we have sensitive data that should only be viewable from a certain staff member, our DW's use will be limited. In order to maintain the security of our current system, less usage could eventually decrease the overall value of our data warehouse.

No matter our needs or concerns, our specialists at Business Impact look forward to helping we make the right the decision when it comes to selecting the right BI solution for our company.

Knowledge Discovery in Database (KDD):

Knowledge Discovery in Database (KDD) refers to the broad process of finding knowledge in data and focuse4s the high-level application of particular data mining methods.

So, it is the process of discovering useful knowledge form the huge collection of data. It is used to extract patterns or trends or relationships in data by applying different data mining tools.

It is very much useful in the business sector and the decision-making process to assist the decision-makers by providing decision-oriented information.

The main goal of the KDD process is to extract knowledge from the data in the context of a large database. It does this by using data mining methods (algorithms) to extract what is deemed knowledge and applying that knowledge in decision making.

Knowledge Discovery in Database (KDD) Process:

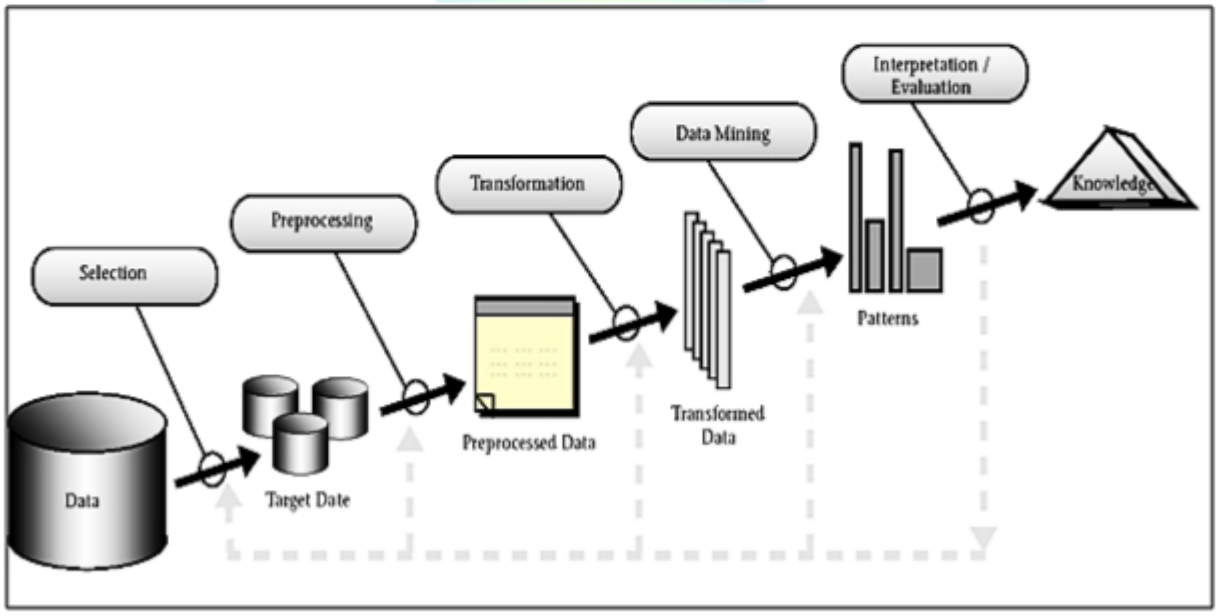


Fig: KDD Process

The overall process of finding and interpreting patterns from data involves the following steps:

1. Developing and Understanding:

This is the initial preliminary step where we develop and understand the following:

a. The application domain
b. The relevant prior knowledge
c. The goals of the end-users.

2. Creating a Target Data:

Set by selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed.

3. Preprocessing:

After collecting the target dataset from various sources, where databases may be in heterogeneous formats, we perform preprocessing. Preprocessing involves:

a. Removing redundant and unnecessary data, known as data cleansing.
b. Strategies for handling missing data fields.
c. Accounting for time sequence information and known changes.

4. Transformation:

Transformation includes:

a. Converting heterogeneous databases into a homogeneous unified form.
b. Finding useful features to represent the data depending on the goal of the task.
c. Using different transformation methods or dimension reduction methods to reduce the effective number of variables under consideration or to find invariant representation for data.

5. Data Mining:

In this step we perform the following tasks:

a. Choosing the data mining task which may be either the process of classification, refraction, clustering, etc.
b. Selecting the appropriate data mining algorithm to be used for searching patterns in data.
c. Deciding which models and parameters may be appropriate.

d. Matching a particular data mining method with the overall criteria of the KDD process.
e. Finally performing data mining in search of patterns of interest in a particular representational form.

6. Interpretation/Evaluation:

In this step the data patterns provided by data mining tools are interpreted and evaluated by experts and decision-makers.

7. Knowledge:

In this step, experts consolidate knowledge discovered by data mining tools to solve the problems and to enhance the effectiveness of decision making.

Need for a Data Warehouse:

A data warehouse is kept separate from the operational database due to the following reasons:

1. An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contrast, data warehouse queries are often complex and they present a general form of data.
2. Operational databases support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database.

3. An operational database query allows reading and modifying operations, while an OLAP query needs only read-only access to stored data.

4. An operational database maintains current data. On the other hand, a data warehouse maintains historical data.

Building a Data Warehouse:

The typical architecture of data warehouse consists of:

1. Different types of databases from various internal and external sources such as operational data, customer data, manufacturing data, sales and marketing data, competitors information, etc.

2. A component that is responsible for data extraction, transformation into a common schema and data cleansing. Data must be extracted from multiple heterogeneous sources. Data must be formatted for consistency within the warehouse. This is essential because the source database may be in different form and format such as heterogeneous databases must be transformed into a common unified homogeneous format. Data must be cleaned to ensure validity. Data cleansing is a complex process but necessary to avoid redundant and unnecessary data.

3. The data warehouse itself. Data from the various sources must be installed in a data model of the data warehouse. Data may have to be converted from relational, object-oriented or legacy databases to a multi-dimensional data model.

4. An information directory is maintained that contains Metadata contents that mean information about the data contained in the warehouse.

5. Different type of data access and analytical tools are used to retrieve the necessary data from the data warehouse, to support the business decision process such tools may include OLAP (Online Analytical Processing), Data Mining, Querying and Reporting, etc.

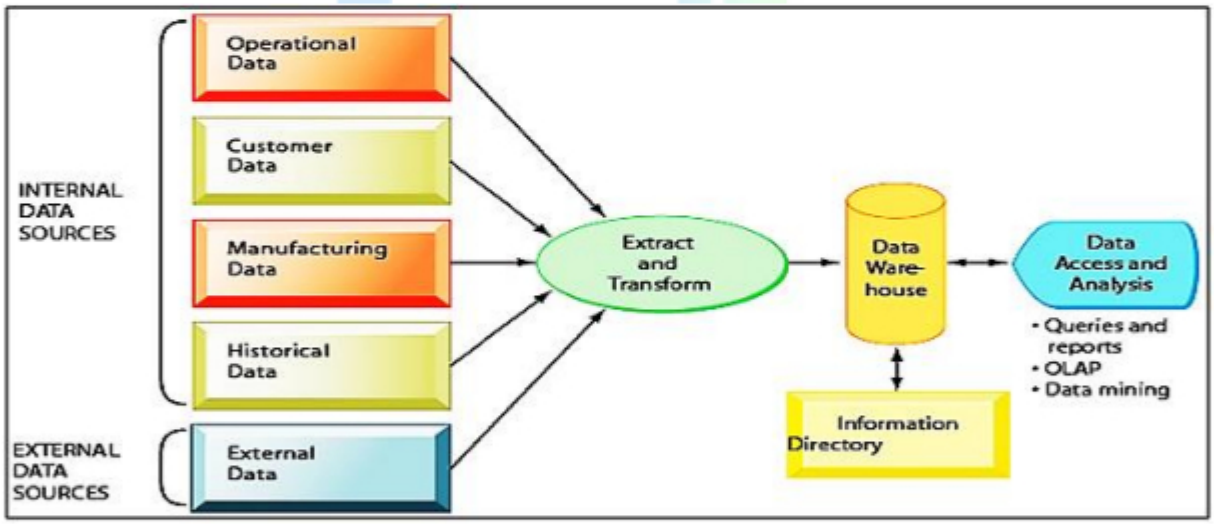


Fig: Building a Data Warehouse

To Integrate Heterogeneous Databases, We Have Two Approaches:

1. Query Driven Approach:

This is the traditional approach to integrating heterogeneous databases. This approach was used to build wrappers and integrators on top of multiple heterogeneous databases. These integrators are also known as mediators.

Now, these queries are mapped and sent to the local any processor. The results from heterogeneous sites are integrated into a global answer set.

2. Update Driven Approach:

This is an alternative to the traditional approach. Today's data warehouse systems follow update driven approach rather than the traditional approach discussed earlier.

In an update driven approach, the information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse. This information is available for direct querying and analysis.

Data Warehousing Technologies:

1. Online Analytical Processing (OLAP):

OLAP supports multi-dimensional data analysis, enabling users to view the same data indifferent ways using multiple dimensions. Each aspect of the information represents a different dimension.

OLAP is an approach to answer multi-dimensional analytical queries interactively from multiple perspectives. Generally, 3-dimensional data model, known as data cube is created and we can rotate the cube to show different aspects of information. Such tools are very useful in analyzing data from a different perspective.

At the core of the OLAP system, there exists an OLAP cube also called multi-dimensional cube or hyper-cube. It consists of numeric facts called measures that are categorized by dimensions. The measures are placed at the intersection of the hyper-cube.

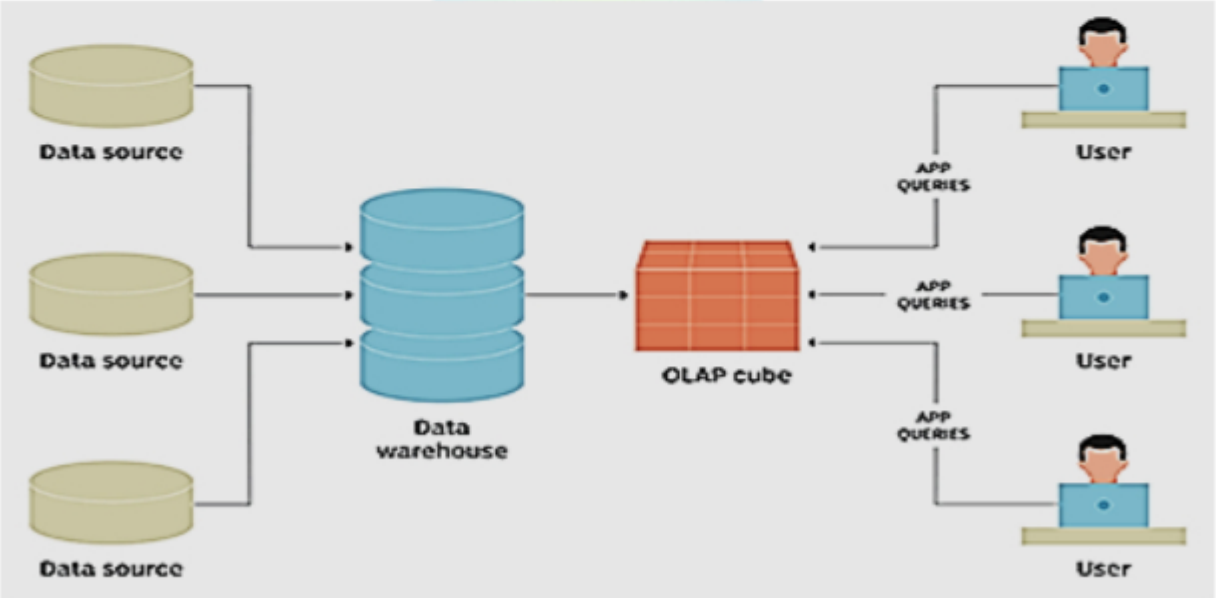


Fig: OLAP Process

Types of OLAP:

1. Relational OLAP:

ROLAP servers are placed between the relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

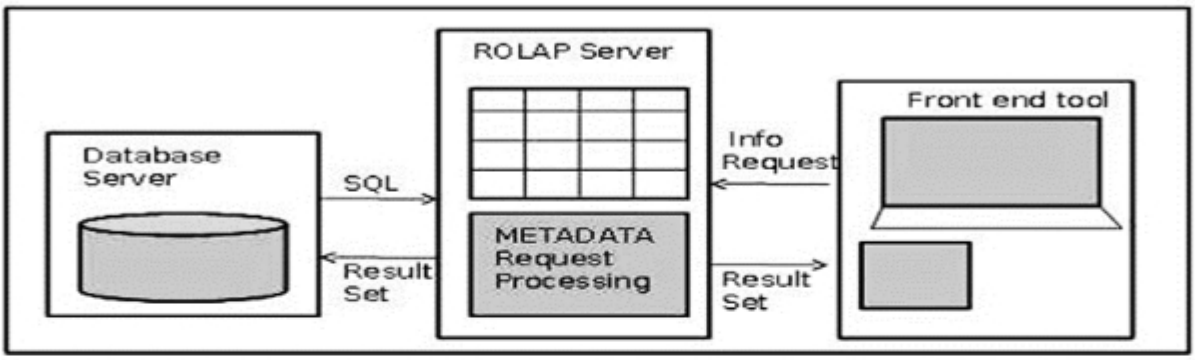


Fig: ROLAP Server

ROLAP Includes The Following:

a. Implementation of aggregation navigation logic.
b. Optimization for each DBMS back end.
c. Additional tools and services.

Points to Remember:

a. ROLAP servers are highly scalable.
b. ROLAP tools analyze large volumes of data across multiple dimensions.
c. ROLAP tools store and analyze highly volatile and changeable data.

Relational OLAP Architectures:

ROLAP includes the following components:

a. Database Server
b. ROLAP Server
c. Front end tool

Advantages:

a. ROLAP servers can be easily used with existing RDBMS.
b. Data can be stored efficiently since no zero facts can be stored.
c. ROLAP tools do not use pre-calculated data cubes.
d. DSS server of micro-strategy adopts the ROLAP approach.

Disadvantages:

a. Poor query performance.
b. Some limitations of scalability depending on the technology architecture that is utilized.

2. Multidimensional OLAP:

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

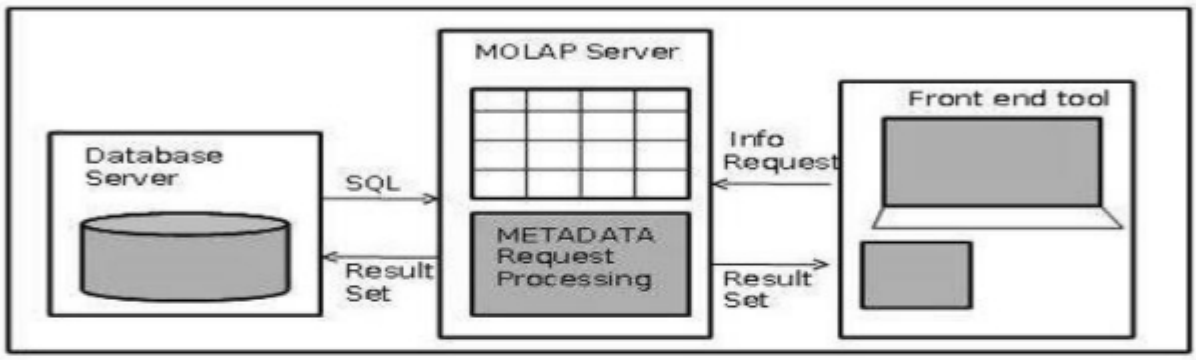


Fig: MOLAP Server

Points to Remember:

a. OLAP tools process information with consistent response time regardless of the level of summarizing or calculations selected.
b. MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis.
c. MOLAP tools need the fastest possible performance.
d. MOLAP server adopts two levels of storage representation to handle dense and sparse dataset.
e. Denser sub-cubes are identified and stored as an array structure.
f. Sparse sub-cubes employ compression technology.

MOLAP Architecture:

MOLAP includes the following components:

a. Database server
b. MOLAP server
c. Front end tools

Advantages:

a. MOLAP allows the fastest indexing to the per-computed summarized data.
b. Helps the users connected to a network who need to analyze larger, less defined data.
c. Easier to use, therefore MOLAP is suitable for inexperienced users.

Disadvantages:

a. MOLAP is not capable of containing detailed data.
b. The storage utilization may be low if the data set is sparse.

Difference between MOLAP and ROLAP:

MOLAP	ROLAP
Information retrieval is fast	Information retrieval is comparatively slow.
Uses sparse array to store datasets.	Uses relational table.
MOLAP is best suited for inexperienced users, since it is very easy to use.	ROLAP is best suited for experienced users.
Maintains a separate database for data cubes.	It may not require space other than available in the data warehouse.
DBMS facility is weak.	DBMS facility is strong.

3. 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 allow storing the large data volumes of detailed information. The aggregations are stored separately in the MOLAP store.

Hybrid OLAPs enable the use of combinations of the two OLAPs, they basically store data in both a relational database and a multidimensional database. As a result, the decision to access one of the two databases depends solely on which is best suited for the desired processing type or application.

This provides much more flexibility when handling data. For heavy data processing, the data is stored in a relational database, while for theoretical processing, the data is stored in a multidimensional database.

This approach is beneficial in the following situations:

1. If there are volumes of data that cannot be handled by one multidimensional database.
2. If there are instances of performance bottlenecks whenever data is being accessed from a server.

3. If there is a need to use existing summarized and arranged data sources.

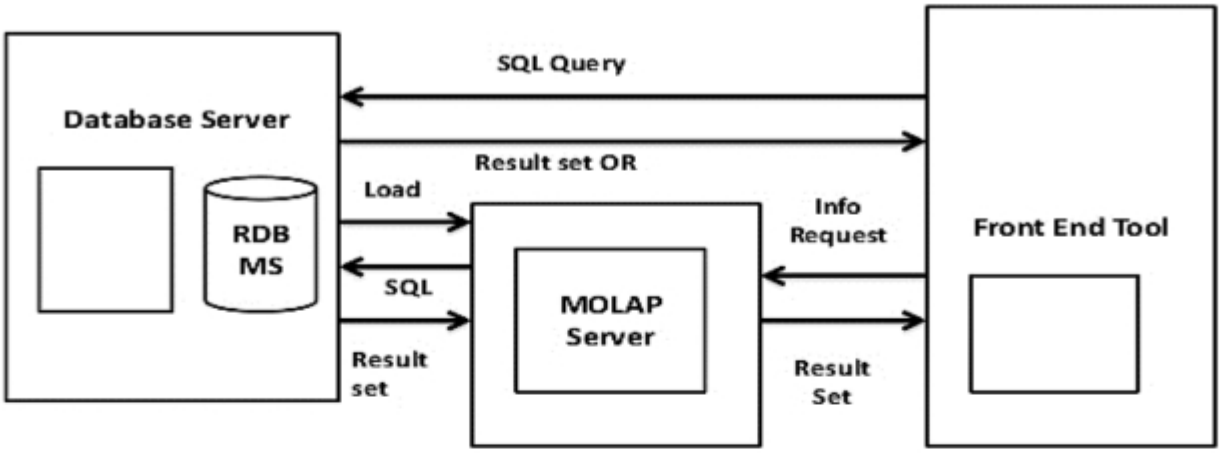


Fig: HOLAP Server

OLAP Operations (Operation of Data Warehouse):

Since OLAP servers are based on the multi-dimensional view of data we can perform variously operations on OLAP and few common operations are:

1. Roll-up:

Roll-up performs aggregation on a data cube in any of the following ways:

a. By climbing up a concept hierarchy for a dimension
b. By dimension reduction

The diagram below illustrates how roll-up works.

a. Roll-up is performed by climbing up a concept hierarchy for the dimension location.
b. Initially the concept hierarchy was "street < city < province < country".
c. On rolling up, the data is aggregated by ascending the location hierarchy from the level of the city to the level of the country.
d. The data is grouped into cities rather than countries.
e. When roll-up is performed, one or more dimensions from the data cube are removed.

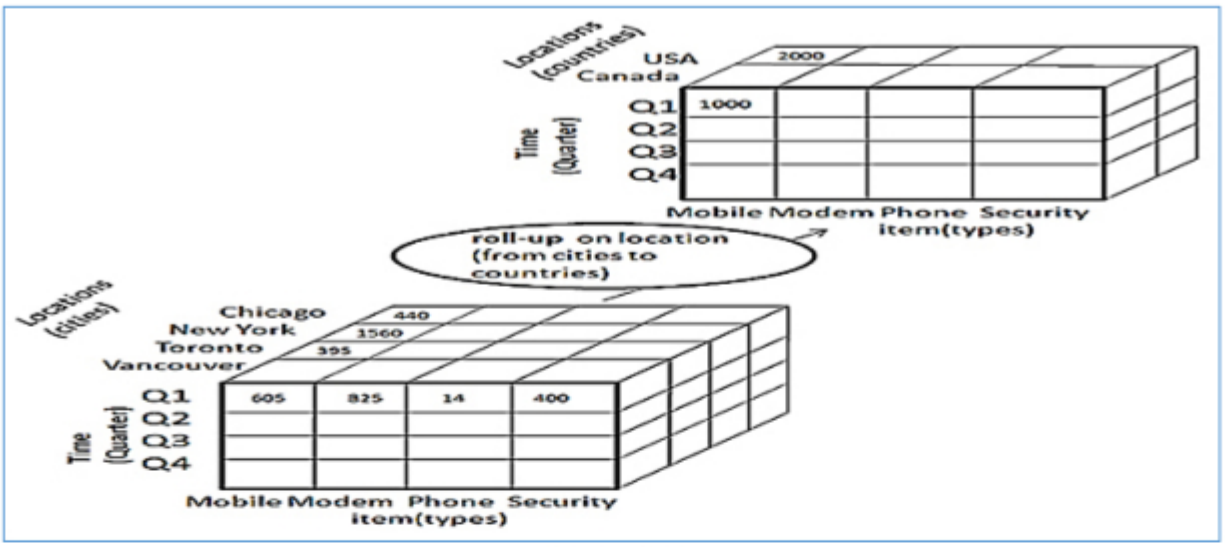


Fig: Roll-up

2. Drill-down:

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:

a. By stepping down a concept hierarchy for a dimension
b. By introducing a new dimension.

The diagram below illustrates how drill-down works:

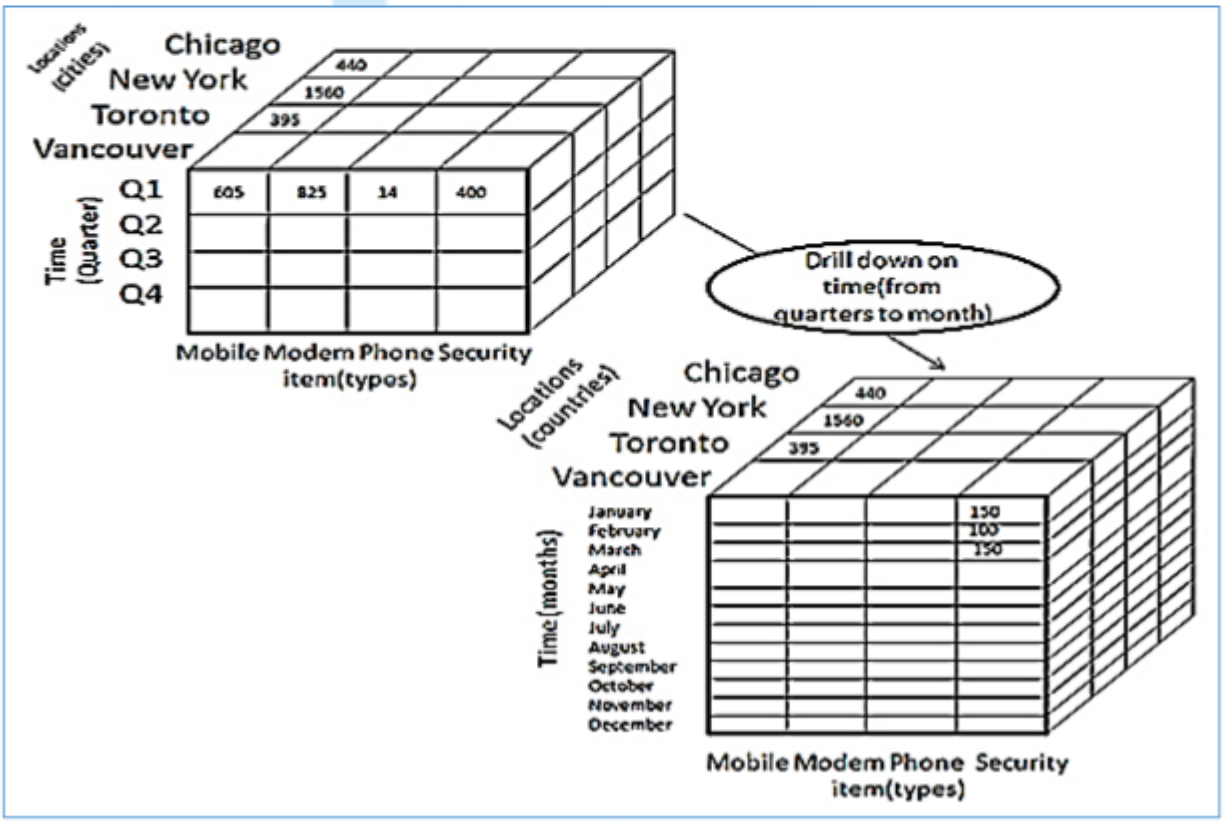


Fig: Drill-Down

a. Drill-down is performed by stepping down a concept hierarchy for the dimension time.
b. Initially, the concept hierarchy was "day < month < quarter < year."
c. On drilling down, the time dimension is descended from the level of the quarter to the level of the month.
d. When drill-down is performed, one or more dimensions from the data cube are added.
e. It navigates the data from less detailed data to highly detailed data.

3. Slice:

The slice operation selects one particular dimension from a given cube and provides a new sub-cube.

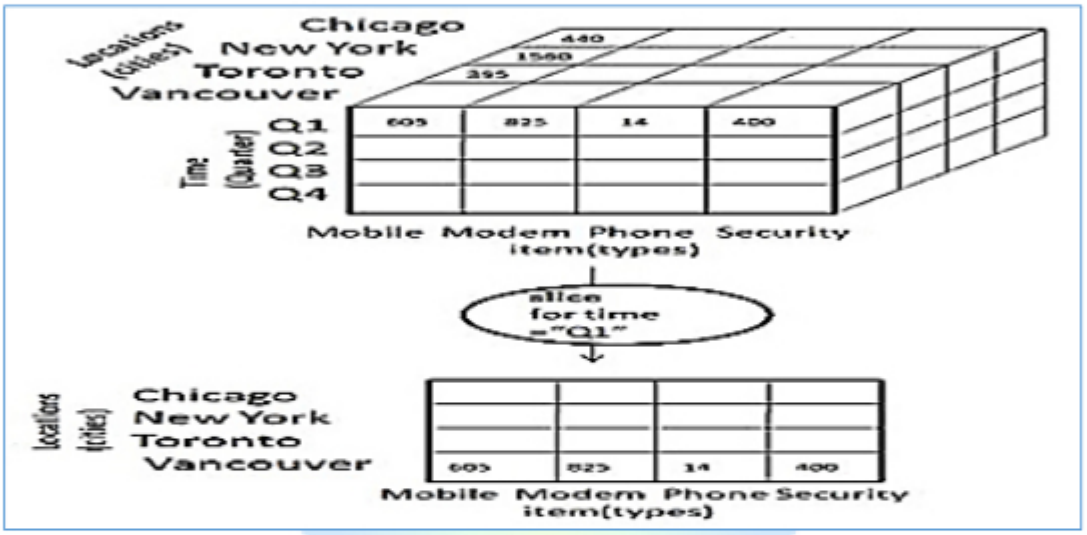


Fig: Slice

The above diagram shows how slice works.

a. Here Slice is performed for the dimension "time" using the criterion time = "Q1".
b. It will form a new sub-cube by selecting one or more dimensions.

4. Dice:

Dice selects two or more dimensions from a given cube and provides a new sub-cube. The dice operation on the cube based on the following selection criteria involves three dimensions.

a. (location = "Toronto" or "Vancouver")
b. (time = "Q1" or "Q2")
c. (item =" Mobile" or "Modem")

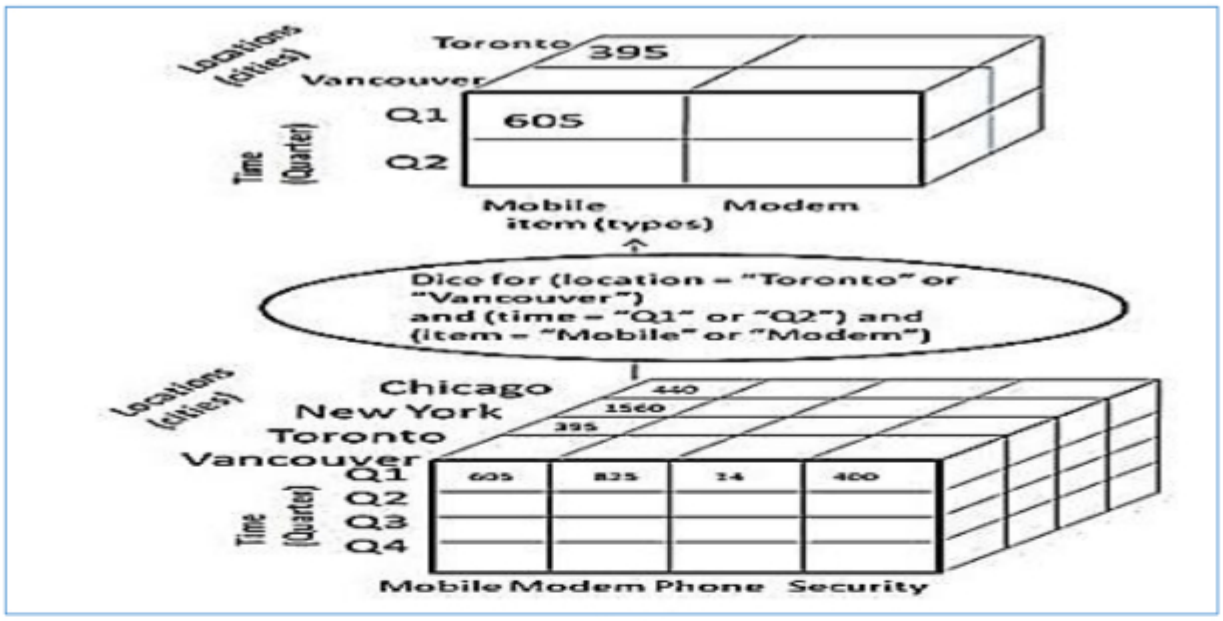


Fig: Dice

5. Pivot:

The pivot operation is also known as rotation. It rotates the data axes in view to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.

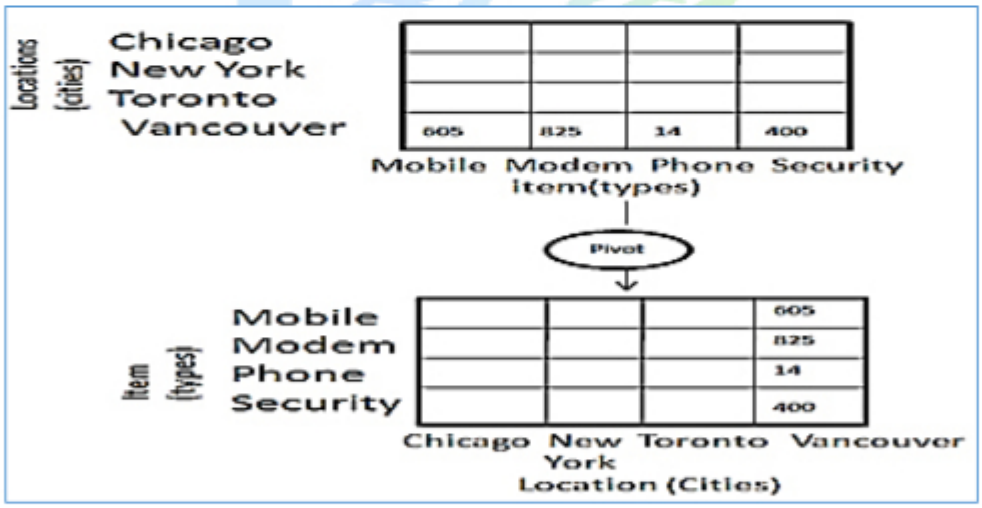


Fig: Pivot

2. Online Transaction Processing (OLTP):

Online Transaction Processing is characterized by a large number of short online transactions (like insert, update, delete). The main focus for the OLTP system is to provide fast query processing, maintain data integrity in multi-user environment and its effectiveness is measured by the number of transaction per second.

In OLTP database there is detailed and current data and the relational data model is used to represent data. OLAP is characterized by a relatively low volume of transaction. Queries are often very complex and involve aggregation.

OLAP applications are widely used in data mining technique. In OLAP databases are maintained in the multi-dimensional schema.

Characteristics of OLTP:

Following are important characteristics of OLTP:

1. OLTP uses transactions that include small amounts of data.
2. Indexed data in the database can be accessed easily.
3. OLTP has a large number of users.
4. It has fast response times
5. Databases are directly accessible to end-users
6. OLTP uses a fully normalized schema for database consistency.
7. The response time of OLTP system is short.
8. It strictly performs only the predefined operations on a small number of records.
9. OLTP stores the records of the last few days or a week.
10. It supports complex data models and tables.

Architecture of OLTP:

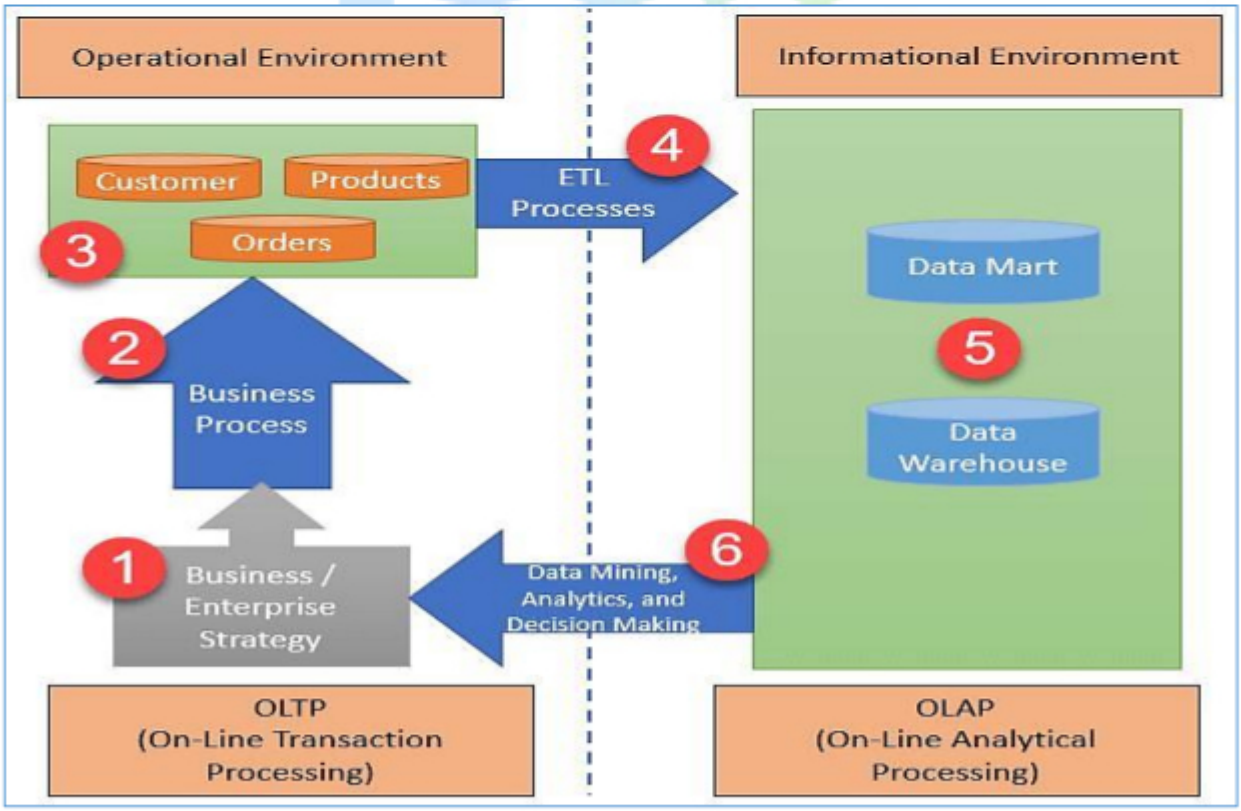


Fig: Architecture of OLTP

1. Business / Enterprise Strategy:

Enterprise strategy deals with the issues that affect the organization as a whole. In OLTP, it is typically developed at a high level within the firm, by the board of directors or the top management

2. Business Process:

OLTP business process is a set of activities and tasks that, once completed, will accomplish an organizational goal.

3. Customers, Orders, and Products:

OLTP database store information about products, orders (transactions), customers (buyers), suppliers (sellers), and employees.

4. ETL Processes:

It separates the data from various RDBMS source systems, then transforms the data (like applying concatenations, calculations, etc.) and loads the processed data into the Data Warehouse system.

5. Data Mart and Data warehouse:

A data mart is a structure/access pattern specific to data warehouse environments. It is used by OLAP to store processed data.

6. Data Mining, Analytics, and Decision Making:

Data stored in the data mart and data warehouse can be used for data mining, analytics, and decision making.

This data helps us to discover data patterns, analyze raw data, and make analytical decisions for our organization's growth.

Difference between OLAP and OLTP:

Data Warehouse (OLAP)	Operational Database (OLTP)
Involves historical processing of information.	Involves day-to-day processing.

OLAP systems are used by knowledge workers such as executives, managers and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
Useful in analyzing the business.	Useful in running the business.
It focuses on Information out.	It focuses on Data in.
Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity-Relationship Model.
Contains historical data.	Contains current data.
Provides summarized and consolidated data.	Provides primitive and highly detailed data.
Provides summarized and multidimensional view of data.	Provides a detailed and flat relational view of data.
The number of users is in hundreds.	Number of users is in thousands.
The number of records accessed is in millions.	Number of records accessed is in tens.
Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
Highly flexible.	Provides high performance.

Data Marts:

Companies often build an enterprise-wide data warehouse, where a central data warehouse serves the entire organization or they create smaller, decentralized data warehouses called data marts.

A data mart is a subset of a data warehouse, in which a summarized or highly focused portion of the organization's data is placed in a separate database for a specific population of users.

For example: A company may develop marketing and sales data marts to deal with customer information. A data mart typically focuses on a single subject area or line of business. So, it usually can be constructed more rapidly at a lower cost than an enterprise-wide data warehouse.

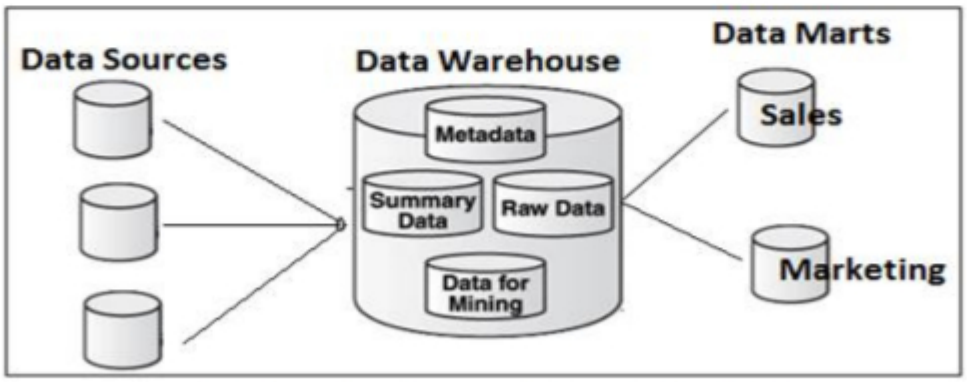


Fig: Graphical Representation of Data Marts

Points to Remember about Data Marts:

1. Windows-based or Linux based servers are used to implement data marts. They are implemented on low-cost servers.
2. The implementation cycle of a data mart is measured in short periods i.e. in weeks rather than months or years.
3. The life cycle of data marts may be complex in the long run, if their planning and design are not organization-wide.
4. Data marts are small in size.
5. Data marts are customized by the department.
6. The source of a data mart is departmentally structured data warehouse.
7. Data marts are flexible.

Metadata:

Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata.

For example, the index of a book serves as metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to detailed data. In terms of a data warehouse, we can define metadata as follows:

a. Metadata is the road-map to a data warehouse.
b. Metadata in a data warehouse defines the warehouse objects.
c. Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Categories of Metadata:

Meta data can be broadly categorized into three categories:

1. Business Metadata:

It has the data ownership information, business definition and changing policies.

2. Technical Metadata:

It includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.

3. Operational Metadata:

It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied to it.

Role of Metadata:

Metadata has a very important role in a data warehouse. The role of metadata in a warehouse is different from the warehouse data, yet it plays an important role. The various roles of metadata are explained below:

a. Metadata acts as a directory.
b. This directory helps the decision support system to locate the contents of the data warehouse.
c. Metadata helps in the decision support system for mapping of data when data is transformed from the operational environment to the data warehouse environment.
d. Metadata helps in summarization between current detailed data and highly summarized data.
e. Metadata also helps in summarization between lightly detailed data and highly summarized data.
f. Metadata is used for query tools.
g. Metadata is used in extraction and cleansing tools.
h. Metadata is used in reporting tools.
i. Metadata is used in transformation tools.
j. Metadata plays an important role in loading functions.

Data Warehouse Schema:

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 a relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema.

Mostly a data warehouse stores multi-dimensional data. For the storage and management of multi-dimensional data, a relational data model is not sufficient. For this different multi-dimensional schema is used and two common multi-dimensional schemas are:

1. Star Schema:

The star schema architecture is the simplest data warehouse schema. It is called star schema because the diagram seems like a star with points radiating from a center. In this schema, tables are categorized as fact tables and dimension tables.

The center of the star consists of the fact table and the points of the star are the dimension table. Usually, fact tables and star schema are in 3NF, *whereas* dimension tables are De-

normalized. It is the simplest architecture but most commonly used nowadays and is recommend by Oracle.

Fact Tables:

A fact table has two types of columns as foreign keys to dimension tables and measure attributes that contain numeric fact. A fact table can contain data on detail or aggregated level.

Dimension Tables:

A dimension is a structure usually composed of one or more hierarchies that categorizes data. The primary keys of each of the dimension tables are part of the composite primary keys of the fact table.

The attributes of dimension tables also known as dimension attributes and are normally descriptive attributes having textual values. Dimension tables are generally small in size than fact table.

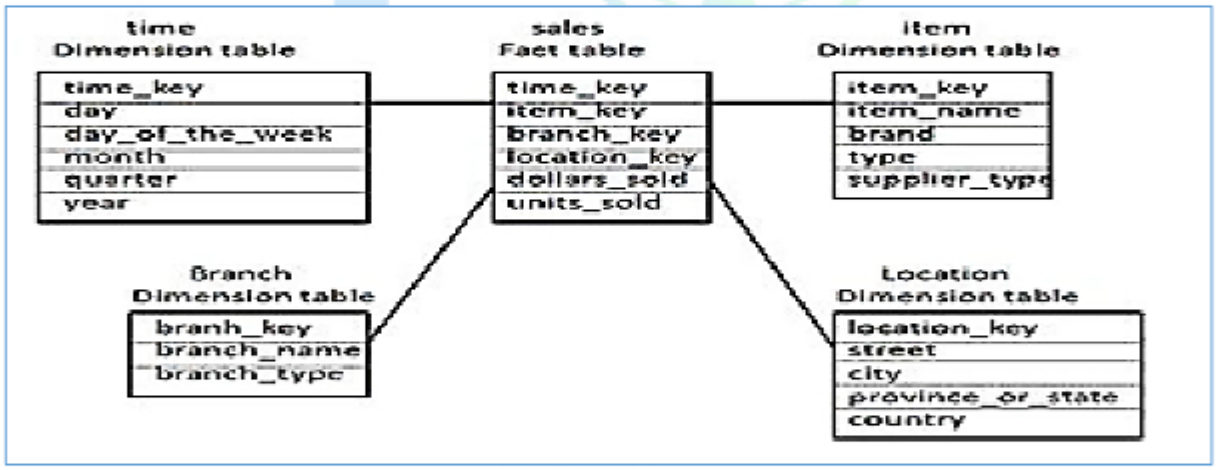


Fig: Star Schema

Characteristics of Star Schema:

1. **Simple Structure**: It is a very simple structure that is easy to understand.

2. **Great Query Effectiveness:** Small numbers of tables are joined so query execution performance is better.

3. **Relatively Long Time to Load Data into the Dimension Table:** De-normalization of dimension tables and redundancy of data can cause the size of the table to be large.

4. **The Most Commonly Used The Data Warehouse Implementation:** It is widely supported by a large number of business intelligence tools and recommended by Oracle.

2. Snowflake Schema:

The snowflake schema is an extension of star schema where each point of the star explodes into many points. In this schema, the dimension table is normalized into multiple tables each representing a level in a dimension hierarchy.

A snowflake schema is a logical arrangement of tables in a multi-dimensional database such that the ER-diagram seems like a snowflake shape. This schema is represented by centralized fact tables which are connected to multiple dimension tables.

Snow flaking is a method of normalizing the dimension table of a star schema. When it is completely normalized resultant structure seems like a snowflake with the fact table in the middle.

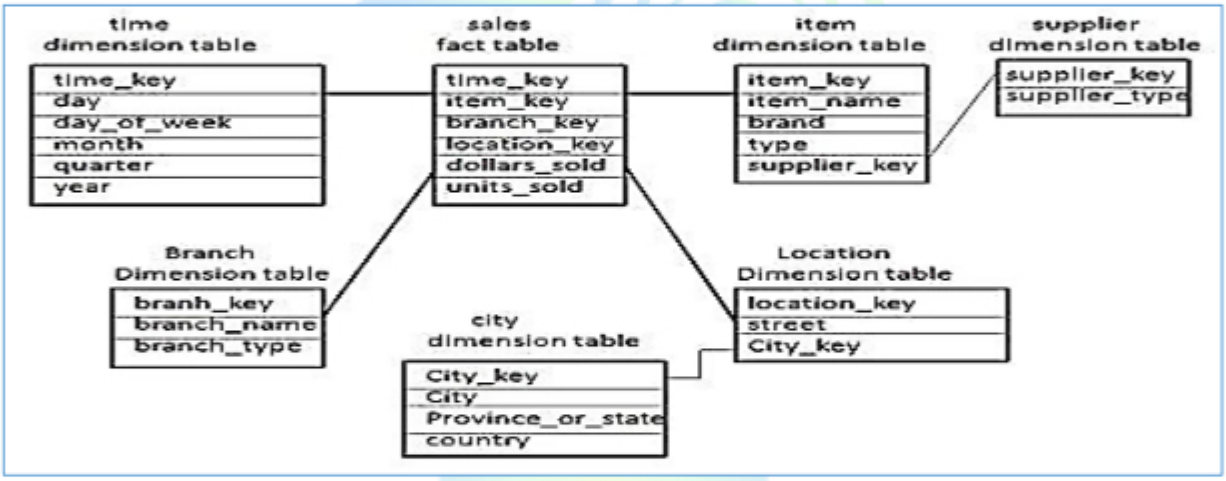


Fig: Snowflake Schema

The Snowflake Schema Provides Some Advantages Over the Star Schema Including:

a. Some OLAP multi-dimensional database modelling tools are optimized for snowflake schema.
b. Normalizing attributes results in the saving of storage space.
c. The primary disadvantage of this schema is the additional level of complexities due to the joining while writing queries.

3. 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.

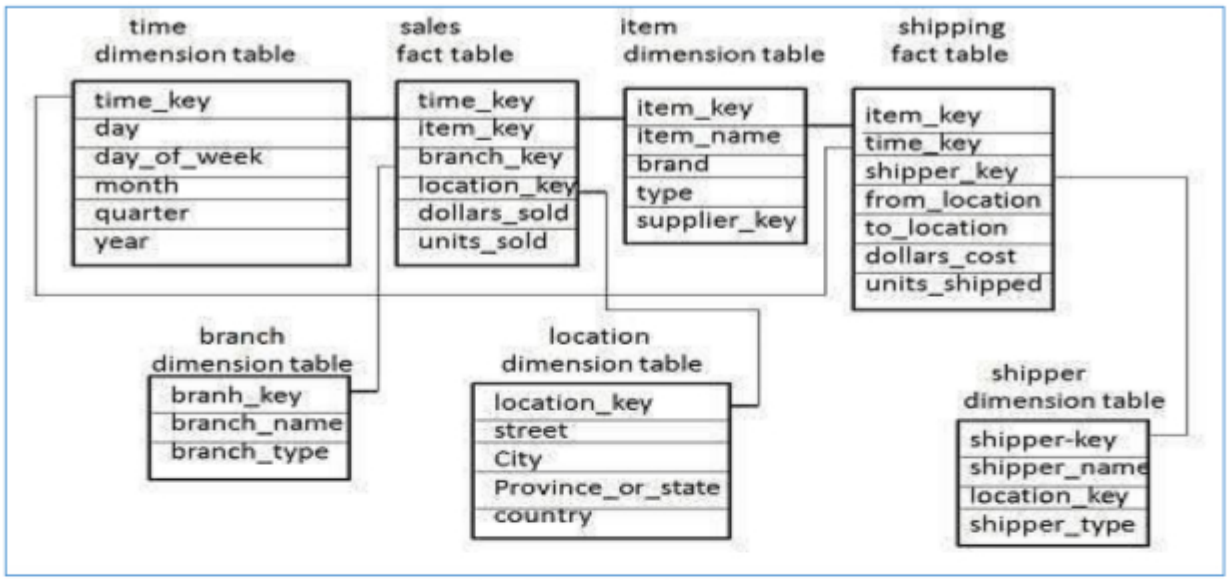


Fig: Fact Constellation Schema

The sales fact table is the same as that in the star schema. The shipping fact table has 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.

Data Mining:

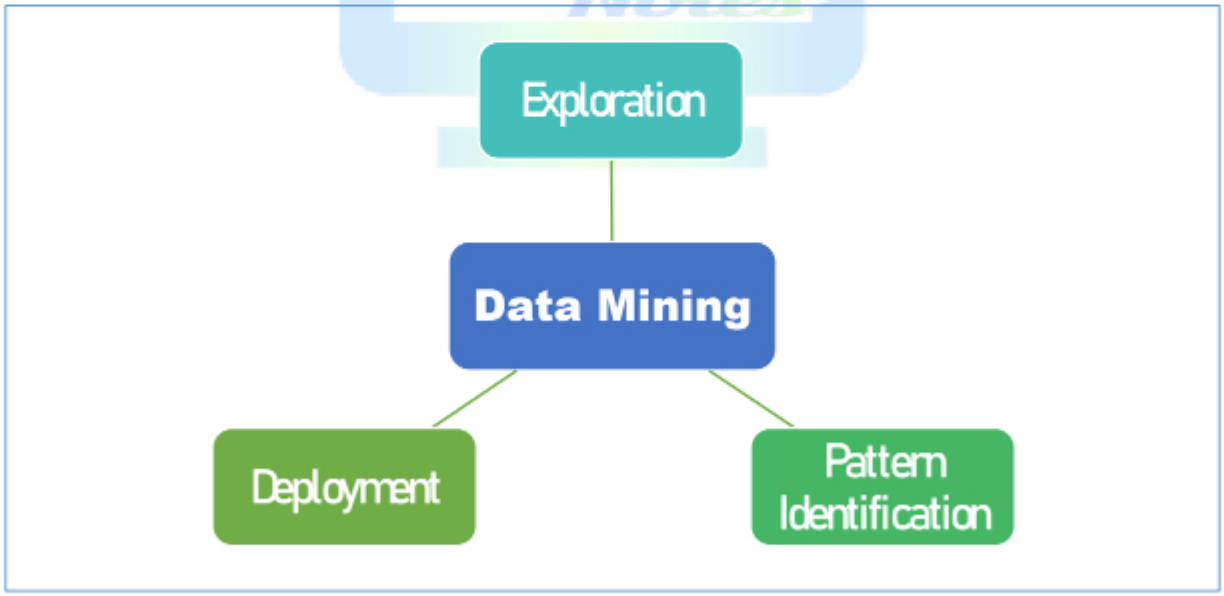


Fig: Data Mining

Data mining is one of the knowledge discovery tools that is used to analyze the huge volume of data contained in a data warehouse or data marts and to reveal the hidden patterns, trends or relationships in data.

Data mining is more discoveries driven. It provides insights into the corporate data that cannot be optioned with OLAP by finding patterns and relationships in data and using inference rules to predict future behavior. The patterns and rules are used to guide decision making and forecast the effects of those decisions.

So, data mining is the process of extracting useful information and patterns from huge data. Data mining includes collection, extraction, analysis and statistics of data.

Data mining involves three basic steps and they are:

1. **Exploration:** In this step, the data are cleared and converted into a suitable form. The nature of data is also determined.
2. **Pattern Identification:** The next step is to choose the pattern, which will make the best prediction.
3. **Deployment:** The identified patterns are used to get the desired outcome.

Advantages:

1. Automated prediction of trends and behaviors.
2. It can analyze a huge volume of data in a very short period.
3. Automated discovery of hidden patterns.
4. It provides lots of models to understand complex data easily.
5. It helps in better prediction and improves decision making.
6. It helps to determine the behavior of the customer and to customize the products and services according to customer's preference.
7. It is helpful to predict future trends. It also helps to increase company revenue.

Uses of Data Mining:

1. Associations:

Associations are occurrences (linked to a single event) for example in an online e-commerce website while purchasing particular products other associated accessories are automatically displayed to promote the sales. These types of information help managers to make a better decision and increasing the profitability.

2. Sequences:

In sequences, events are linked over time. For example: if a house is purchased then the possibility of purchasing furniture, kitchen accessories, etc. is much higher in the coming days.

3. Classification:

It recognizes patterns that describe the group to which an item belongs by examining existing items that had been classifying and by inferring a set of rules. Classifications are also useful in business decision making.

4. Clusters:

Clustering works similarly to classification when no groups have been defined yet. Data mining tools can discover different grouping within data, which can be helpful in various decision making.

5. Forecasts:

It uses a series of existing values to forecast what other values will be. For example, forecasting may find patterns in data to help managers to estimate the future value of continuous variables such as sales figures.

Data Mining Tools:

Data mining tools are the software tools that are used to query information in a data warehouse. These data mining tools support the concept of OLAP to support decision making tasks. There are four major types of data mining tools in a data warehouse environment.

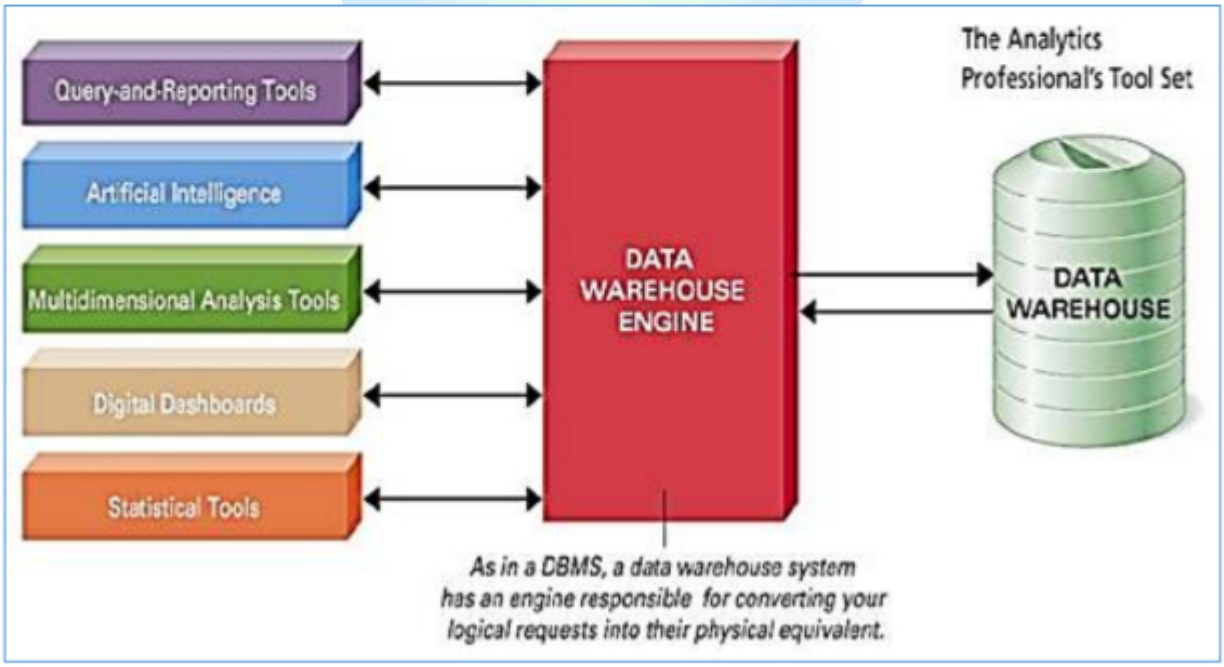


Fig: Data Mining Tools

1. Query and Reporting Tools:

Query and reporting tools are similar to QBE tools, SQL and report generators in the typical database environment. These types of tools are used to generate simple queries and reports.

2. Intelligent Agents:

Intelligent agents utilize various artificial intelligence tools such as neural networks and fuzzy logic to form the basis of information discovery and building business intelligence in OLAP.

3. Multidimensional Analysis (MDA) Tools:

It is slice and dice technique that allows viewing multidimensional information from different perspectives.

4. Statistical Tools:

Those tools apply various mathematical models to the information stored in data warehouse perspectives.

Data Mining Process:

Data mining is an iterative process that typically involves the following phases:

1. Problem Definition:

A data mining project starts with an understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective.

The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required.

2. Data Exploration:

Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data.

A frequent exchange with the data mining experts and business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

3. Data Preparation:

Domain experts build the data model for the modelling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value.

In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modelling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

4. Modelling:

Data mining experts select and apply various mining functions because we can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model.

In the modelling phase, a frequent exchange with the domain experts from the data preparation phase are required.

The modelling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modelling phase is completed, a model of high quality has been built.

5. Evaluation:

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modelling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

a. Does the model achieve the business objective?
b. Have all business issues been considered?

At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

6. Deployment:

Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets. Intelligent Miner products assist us to follow this process. We can apply the functions of the Intelligent Miner products independently, iteratively, or in combination.

The following figure shows the phases of the **Cross-Industry Standard Process** for data mining (CRISP-DM) process model.

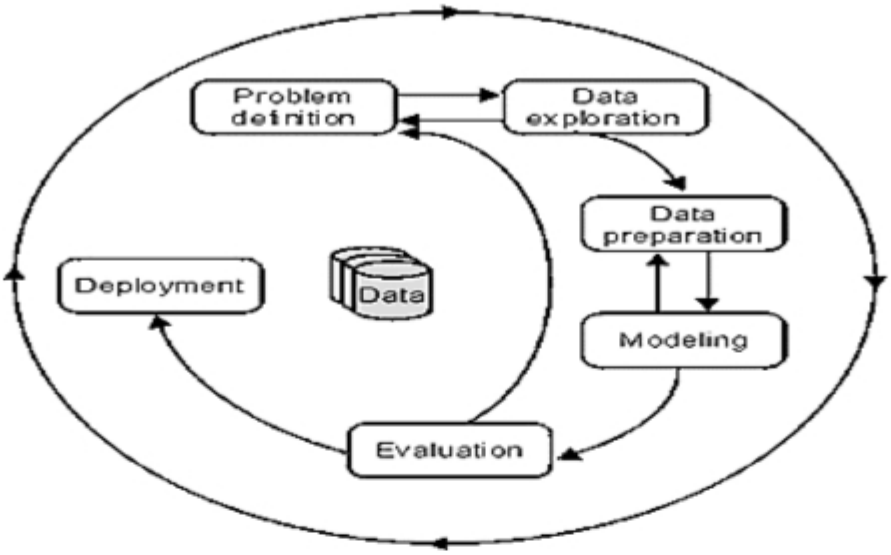


Fig: The CRISP-DM process model

Integrated Modeling (IM) helps us to select the input data, explore the data, transform the data, and mine the data. With IM Visualization we can display the data mining results to analyze and interpret them. With IM Scoring, we can apply the model that we have created with Integrated Modeling.

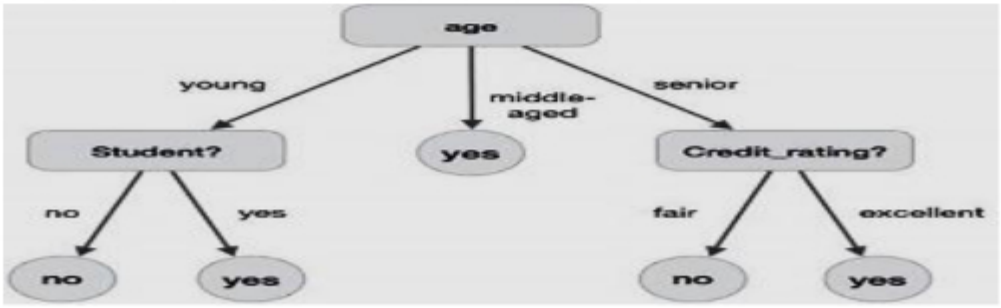
Classification of Data Mining Algorithm:

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme. In this scheme, the main focus is on data mining design and on developing efficient and effective algorithms for mining the available datasets.

1. Decision Tree Induction Algorithm:

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.



The benefits of having a decision tree are:

a. It does not require any domain knowledge.
b. It is easy to comprehend.
c. The learning and classification steps of a decision tree are simple and fast.

Decision Tree Induction Algorithm:

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm, there is no backtracking; the trees are constructed in a top-down recursive divide-and-conquer manner.

```
Generating a Decision Tree Form Training Tuples of Data Partition D
Algorithm : Generate_decision_tree

Input:
Data partition, D, which is a set of training tuples and their associated
class labels.
attribute_list, the set of candidate attributes.
Attribute selection method, a procedure to determine the
splitting criterion that best partitions that the data
tuples into individual classes. This criterion includes a
splitting_attribute and either a splitting point or splitting subset.
Output:
 A Decision Tree
```

2. Bayesian Classifier Algorithm:

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem:

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities:

a. Posterior Probability [P(H/X)]
b. Prior Probability [P(H)]

Where X is data tuple and H is some hypothesis. According to Bayes' Theorem, P(H/X)= P(X/H)P(H) / P(X)

Bayesian Belief Network:

Bayesian Belief Networks specify joint conditional probability distributions. They are also known as Belief Networks, Bayesian Networks, or Probabilistic Networks.

a. A Belief Network allows class conditional independencies to be defined between subsets of variables.
b. It provides a graphical model of a causal relationship in which learning can be performed.
c. We can use a trained Bayesian Network for classification.

Two components define a Bayesian Belief Network:

a. Directed acyclic graph
b. A set of conditional probability tables

3. Rule-Based Algorithm:

Rule-based classifier makes use of a set of IF-THEN rules for classification. We can express a rule in the following form:

IF condition THEN conclusion
Let us consider a rule R1,

```
R1: IF age = youth AND student = yes
THEN buy_computer = yes
```

Points To Remember:

a. The IF part of the rule is called rule antecedent or precondition.
b. The THEN part of the rule is called rule consequent.
c. The antecedent part the condition consists of one or more attribute tests and these tests are logically ANDed.
d. The consequent part consists of class prediction.

Note: We can also write rule R1 as follows:

```
R1: (age = youth) ^ (student = yes))(buys computer = yes)
```

If the condition holds true for a given tuple, then the antecedent is satisfied.

4. Genetic Algorithm:

A genetic algorithm is an artificial intelligence system that mimics the evolutionary, survival-of-the-fittest process to generate increasingly better solutions to a problem. Genetic algorithms are best suited to decision-making environments in which thousands or even millions of solutions are possible. It uses the evolutionary concepts of selection, crossover, and mutation to generate new many more solutions or strategies.

a. **Selection:** It is the process of choosing good solutions i.e. survival-of-the-fittest.
b. **Crossover:** It is the process of combining portions of good solutions in the hope of creating an even better outcome, and

c. **Mutation:** It is the process of randomly changing parts of a solution and evaluating the success or failure of the outcome.

Data Mining Techniques:

1. Statistical Data Mining Techniques:

Some of the statistical data mining techniques are as follows:

A. Regression:

Regression methods are used to predict the value of the response variable from one or more predictor variables were the variables are numeric. Listed below are the forms of regression:

a. Linear
b. Multiple
c. Weighted
d. Polynomial
e. Nonparametric
f. Robust

B. Generalized Linear Model:

Generalized linear model includes:

a. Logistic Regression
b. Poisson Regression

C. Analysis of Variance:

This technique analyzes:

a. Experimental data for two or more populations described by a numeric response variable.
b. One or more categorical variables (factors).

D. Mixed Effect Models:

These models are used for analyzing grouped data. These models describe the relationship between a response variable and some co-variants in the data grouped according to one or more factors.

E. Factor Analysis:

Factor analysis is used to predict a categorical response variable. This method assumes that independent variables follow a multivariate normal distribution.

F. Time Series Analysis:

Following are the methods for analyzing time-series data:

a. Autoregression methods.
b. Long memory time series modelling.

2. Visual Data Mining Techniques:

Visual data mining uses data and/or knowledge visualization techniques to discover implicit knowledge from large data sets. Visual data mining can be viewed as an integration of the following disciplines:

a. Data visualization
b. Data mining

Visual data mining is closely related to the following:

a. Computer graphics
b. Multimedia systems
c. Human-computer interaction
d. Pattern recognition
e. High-performance computing

Data Mining Result Visualization:

Data mining result visualization is the presentation of the results of data mining in visual forms. These visual forms could be scattered plots, box plots, etc.

Data Mining Processing Visualization:

Data mining process visualization presents the several processes of data mining. It allows users to see how the data is extracted. It also allows the users to see from which database or data warehouse the data is cleaned, integrated, preprocessed, and mined.

3. Audio Data Mining Techniques:

Audio data mining makes use of audio signals to indicate the patterns of data or the features of data mining results. By transforming patterns into sound and musing, we can listen to pitches and tunes, instead of watching pictures, to identify anything interesting.

Implementation of Data Warehouse and Data Mining:

Data Warehouse Applications:

Data Warehouses owing to their potential have deep-rooted applications in every industry which use historical data for prediction, statistical analysis, and decision making. Listed below are the applications of Data warehouses across innumerable industry backgrounds.

1. Banking Industry:

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

Most banks also use warehouses to manage the resources available on deck effectively. Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.

Analysis of card holder's transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity. Apart from all these, there is also scope for co-branding.

2. Finance Industry:

Similar to the applications seen in banking, mainly revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

3. Consumer Goods Industry:

They are used for prediction of consumer trends, inventory management, market and advertising research. In-depth analysis of sales and production is also carried out. Apart from these, information is exchanged from business partners and clientele.

4. Government and Education:

The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.

The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers, and also their entire criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.

Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management. The entire financial department of most universities depends on data warehouses, inclusive of the Financial Aid department.

5. Healthcare:

One of the most important sector which utilizes data warehouses are the Healthcare sector. All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.

6. Hospitality Industry:

A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services. They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.

7. Insurance:

As the saying goes in the insurance services the sector, "Insurance can never be bought, it can be only be sold", the warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants. The design of tailor-made customer offers and promotions are also possible through warehouses.

8. Manufacturing and Distribution Industry:

This industry is one of the most important sources of income for any state. A manufacturing organization has to take several make-or-buy decisions can influence the

future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyze current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.

They also use them for product shipment records, records of product portfolios, identify profitable product lines, analyze previous data and customer feedback to evaluate the weaker product lines and eliminate them.

For the distributions, the supply chain management of products operates through data warehouses.

9. The Retailers:

Retailers serve as middlemen between producers and consumers. They need to maintain records of both the parties to ensure their existence in the market.

They use warehouses to track items, their advertising promotions, and the consumers buying trends. They also analyze sales to determine fast selling and slow-selling product lines and determine their shelf space through a process of elimination.

10. Services Sector:

Data warehouses find themselves to be of use in the service sector for maintenance of financial records, revenue patterns, customer profiling, resource management, and human resources.

11. Telephone Industry:

The telephone industry operates over both offline and online data burdening them with a lot of historical data which has to be consolidated and integrated.

Apart from those operations, analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.

12. Transportation Industry:

In the transportation industry, data warehouses record customer data enabling traders to experiment with target marketing where the marketing campaigns are designed by keeping customer requirements in mind.

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

Data Mining Applications:

Data Mining is primarily used today by companies with a strong consumer focus retail, financial, communication, and marketing organizations, to "drill down" into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits.

With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments.

1. Future Healthcare:

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics.

Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

2. Market Basket Analysis:

Market basket analysis is a modelling technique based upon a theory that if we buy a certain group of items we are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behavior of a buyer.

This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

3. Education:

There is a newly emerging field, called Educational Data Mining concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning.

Data mining can be used by an institution to take accurate decisions and also to predict the results of the student.

With the results, the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

4. Manufacturing Engineering:

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in a complex manufacturing process.

Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

5. CRM:

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer-focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyses the information.

This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

6. Fraud Detection:

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time-consuming and complex. Data mining aids in providing meaningful patterns and turning data into information.

Any information that is valid and useful is knowledge. A perfect fraud detection system should protect the information of all the users. A supervised method includes a collection of sample records.

These records are classified as fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

7. Intrusion Detection:

Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection.

Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

8. Lie Detection:

Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also.

This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model, processes can be created according to the necessity.

9. Customer Segmentation:

Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers.

The market is always about retaining customers. Data mining allows finding a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction.

10. Financial Banking:

With computerized banking everywhere a huge amount of data is supposed to be generated with new transactions.

Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts.

The managers may find this information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

11. Corporate Surveillance:

Corporate surveillance is the monitoring of a person or group's behavior by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies.

It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

12. Research Analysis:

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research.

Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualization and visual data mining provide us with a clear view of the data.

13. Criminal Investigation:

Criminology is a process that aims to identify crime characteristics. Actually, crime analysis includes exploring and detecting crimes and their relationships with criminals.

The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques.

Text-based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

14. Bio Informatics:

Data Mining approaches seem ideally suited for Bioinformatics since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience.

Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein subcellular location prediction.