Chapter 4

Introduction to DBMS

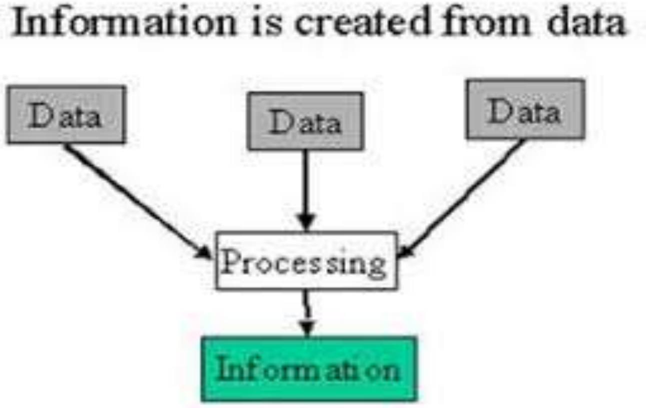
Data vs. Information

Data are simply facts or figures — *bits* of information, but not information itself. When data are processed, interpreted, organized, structured or presented so as to make them meaningful or useful, they are called **information**. Information provides context for data.

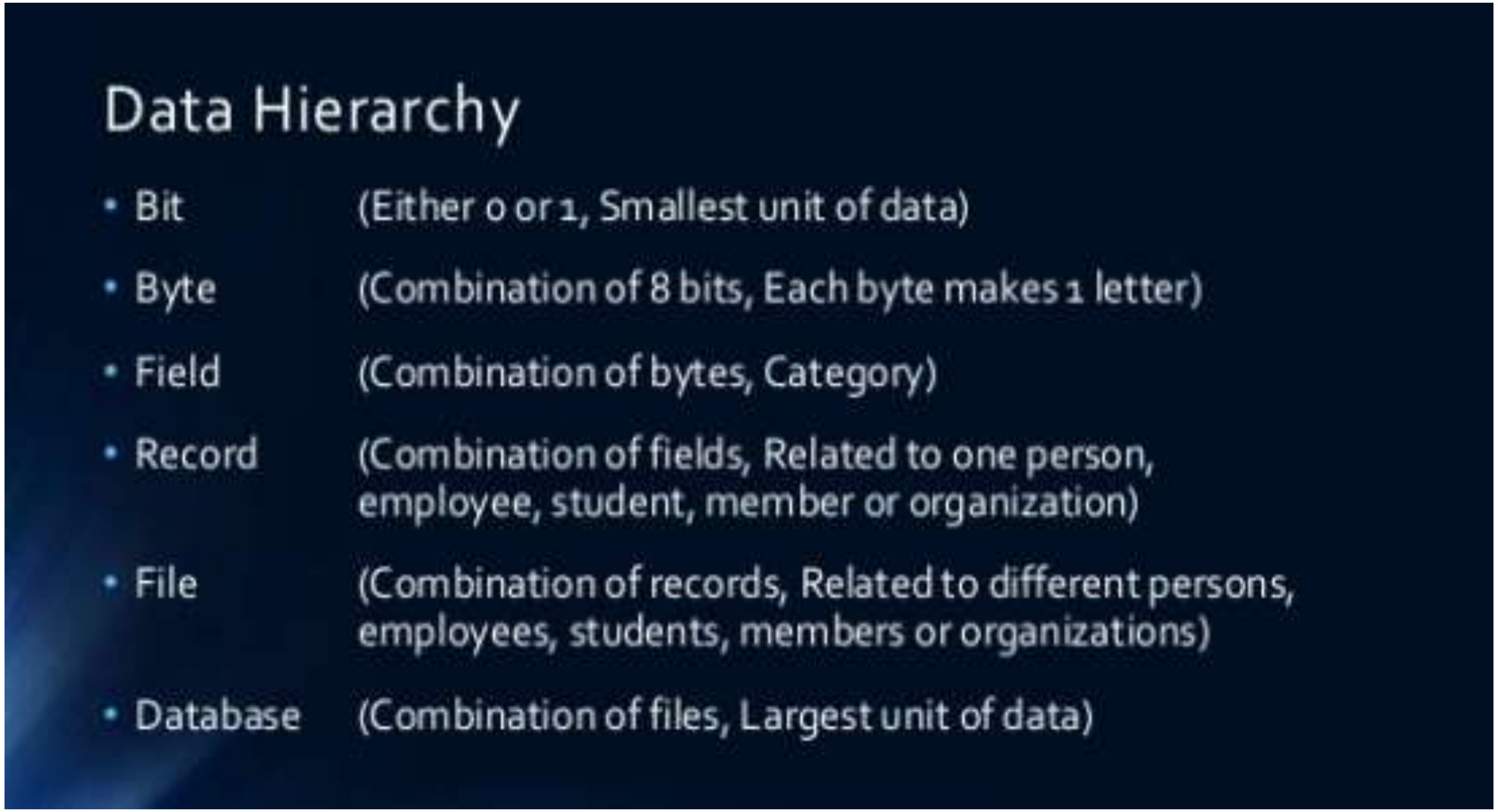
Data and information are interrelated. Data usually refers to raw data, or unprocessed data. It is the basic form of data, data that hasn't been analyzed or processed in any manner. Once the data is analyzed, it is considered as information. Information is "knowledge communicated or received concerning a particular fact or circumstance." Information is a sequence of symbols that can be interpreted as a message. It provides knowledge or insight about a certain matter.

Some differences between data and information:

- Data is used as input for the computer system. Information is the output of data.
- Data is unprocessed facts figures. Information is processed data.
- Data doesn't depend on Information. Information depends on data.
- Data is not specific. Information is specific.
- Data is a single unit. A group of data which carries news and meaning is called Information.
- Data doesn't carry a meaning. Information must carry a logical meaning.
- Data is the raw material. Information is the product.



Data Hierarchy



Data Hierarchy

- Bit (Either 0 or 1, Smallest unit of data)
- Byte (Combination of 8 bits, Each byte makes 1 letter)
- Field (Combination of bytes, Category)
- Record (Combination of fields, Related to one person, employee, student, member or organization)
- File (Combination of records, Related to different persons, employees, students, members or organizations)
- Database (Combination of files, Largest unit of data)

What Is Data Management?

- Data management is the development and execution of processes, architectures, policies, practices and procedures in order to manage the information generated by an organization.
- The effective management of data within any organization has grown in importance in recent years as organizations are subject to an increasing number of compliance regulations, large increases in storage information storage capacity and the sheer amount of data and documents being generated by organizations.
- This rate of growth is not expected to slow down as IDC(International Data Corporation) predicts the amount of information generated will increase 29 fold by 2020. These large sums of data from ERP systems, CRM systems and general business documents if often referred to as big data.

Database

- A database is a collection of information that is organized so that it can be easily accessed, managed and updated.
- Data is organized into rows, columns and tables, and it is indexed to make it easier to find relevant information. Data gets updated, expanded and deleted as new information is added. Databases process workloads to create and update themselves, querying the data they contain and running applications against it.

Database systems

- ❑ Created in the 1980s to solve problems associated with file-based data processing
- ❑ Store all organizational data in central location (a database)
- ❑ A single application called Database Management System (DBMS) performs all data-handling operations (retrieving, updating, inserting, deleting data values)
- ❑ All programs interface with the DBMS to access the database data.
- ❑ Complex database systems require a database administrator (DBA)

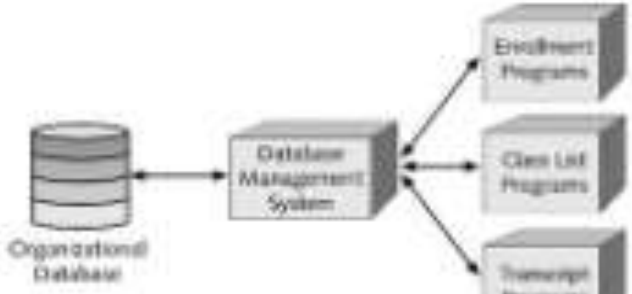
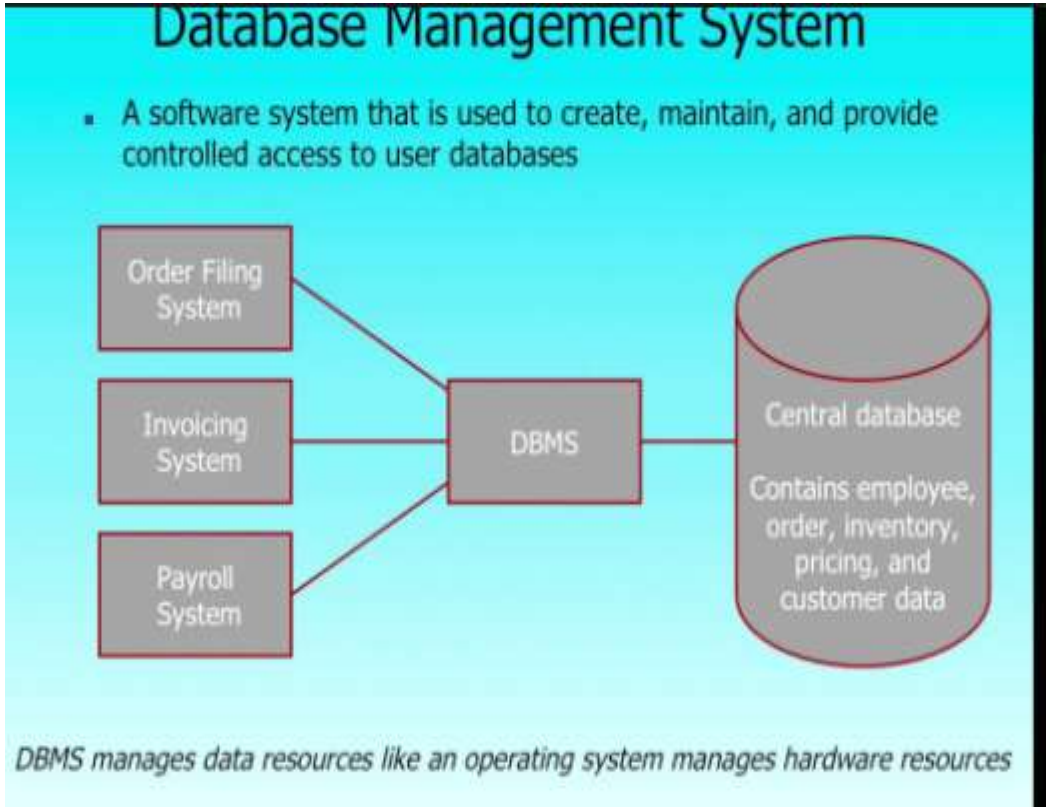


Figure 1-4 Database approach to data processing

Database management system(DBMS)

- A database management system is the software system that allows users to define, create and maintain a database and provides controlled access to the data. A Database Management System (DBMS) is basically a collection of programs that enables users to store, modify, and extract information from a database as per the requirements. DBMS is an intermediate layer between programs and the data. Programs access the DBMS, which then accesses the data. There are different types of DBMS ranging from small systems that run on personal computers to huge systems that run on mainframes.
- Examples are dbase, FoxPro, IMS and Oracle, postgres, MySQL, SQL Servers and DB2 etc.



Database Management System

- A software system that is used to create, maintain, and provide controlled access to user databases

DBMS manages data resources like an operating system manages hardware resources

Advantages of DBMS:

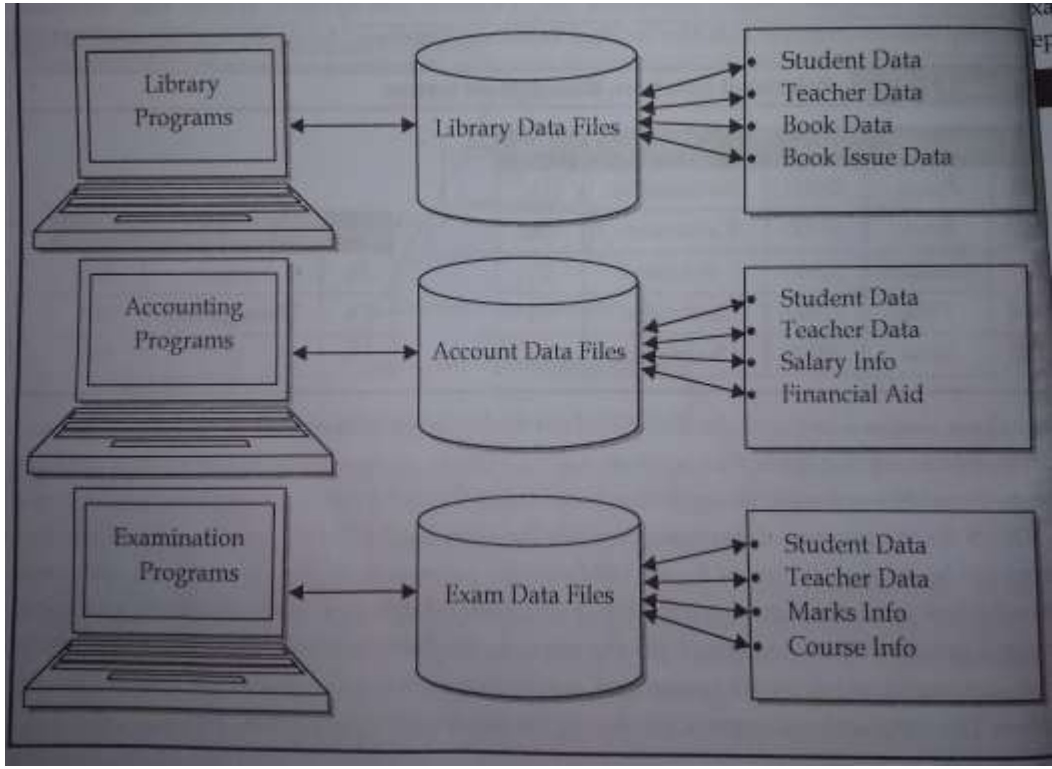
- Controls Redundancy
- Integrity can be enforced
- Inconsistency can be avoided
- Data can be shared
- Standards can be enforced
- Restricts unauthorized access
- Solves Enterprise Requirement than Individual Requirement
- Provides Backup and Recovery
- Concurrency Control

Disadvantages of DBMS

- Complexity
- Size
- Performance
- Higher impact of a failure
- Cost of DBMS
- Additional Hardware costs
- Cost of Conversion

File management system(FMS)/Flat File system

- A file management system is a type of software that manages data files in a computer system. It has limited capabilities and is designed to manage individual or group files, such as special office documents and records. It may display report details, like owner, creation date, state of completion and similar features useful in an office environment.
- A file management system is also known as a file manager.



Characteristics of File Processing System:

- It is a group of files storing data of an organization.
- Each file is independent from one another.
- Each file is called a flat file.
- Each file contained and processed information for one specific function, such as accounting or inventory.
- Files are designed by using programs written in programming languages such as COBOL, C, C++.
- The physical implementation and access procedures are written into database application; therefore, physical changes resulted in intensive rework on the part of the programmer.
- As systems became more complex, file processing systems offered little flexibility, presented many limitations, and were difficult to maintain.

Limitations of file system

- Separated and Isolated Data
- Duplication of data
- Data Dependence
- Difficulty in representing data from the user’s view
- Data Inflexibility
- Incompatible file formats
- Data Security
- Transactional Problems(Atomicity Problems)
- Concurrency problems
- Poor data modeling of real world

Advantages

- Easy to understand.
- Easy to implement.
- Low Initial investment .
- Less hardware and software requirements.
- Less Skills set are required to handle flat database systems.
- Best for small databases.
- Low overhead cost
- Not Required Dedicated staff

When flat file systems are suitable

- Simple and easy system development
- Manage few data file(2-3 data files)
- Security is not major concern
- Concurrent access is not needed

Application areas of Database Systems

- Airlines and railways
- Banking
- Education
- Telecommunications
- Credit card transactions
- E-commerce
- Health care information system & electronic patient record
- Digital libraries and digital publishing
- Finance
- Sales
- Human resources

Characteristics of Database Approach

- ***Control of data redundancy***

  In the database approach there is central repository of data that not only helps in avoiding the wastage of storage space but also helps in controlling the redundancy by data integration. It helps in avoiding the duplication of data by following techniques like normalization and key concepts. Thus the data is stored in database table at only one place from where it can be retrieved when needed, by avoiding redundancy.

- ***Data consistency***

  This is maintained by following the concept **"control of redundancy".** If the data is stored at one place in a database then while updating any information the changes will be done at only one place which is reflected at all place where ever it is present in whole database. There is no need to change at all places where that data is present. For example if an employee has a change in his address then only in employee table the address will be changed. From there it will be updated every where in database. Thus it ensures all copies of the data are kept consistent. This helps in maintaining consistency of information throughout the system without any loss or misleading of information.

- ***Improved data integrity***

Data integrity mainly refers to ensuring that data is recorded exactly as intended and when retrieved it’s in the same way as it was when it was recorded. There should not be any data loss when data is retrieved. It mainly provides the validity and consistency of stored data. The database application has various **Integrity Constraints**, which are consistency rules that the database is not permitted to violate. One of the constraints is specifying data type for every data item. Another is referential integrity constraint in which a record in one file must be related to records in other files. These constraints help in maintaining integrity of data.

- ***Improved security***

Database approach provides a protection of the data from the unauthorized users. It may take the term of user names and passwords to identify user type and their access right in the operation including retrieval, insertion, updating and deletion. Providing the facility of accessible rights in database system for various levels in an organization makes it secure for use. For example, the DBA has access to all the data in the database where as a branch manager may have access to all data that is related to only his branch office. In a similar way a sales assistant may have access to all data relating to properties but don’t have any access to sensitive data such as staff salary details.

- ***Enforcement of standards***

The integration of the database enforces the necessary standards including data formats, naming conventions, documentation standards, update procedures and access rules. It helps in maintaining standards among the user in an organization. The sharing of data within departments, exchange of information among the users on various projects become easy following the standard database on a centralized environment.

- ***Improved data accessibility and responsiveness***

  By having integration in the database approach, data accessing can be crossed departmental boundaries. This feature provides more functionality and better services to the users.

- ***Increased productivity***

  The database approach provides all the low-level file-handling routines. The provision of these functions allows the programmer to concentrate more on the specific functionality required by the users. The fourth-generation environment provided by the database can simplify the database application development.

- ***Improved maintenance***

  Database approach provides a data independence. As a change of data structure in the database will be affect the application program, it simplifies database application maintenance.

- ***Increased concurrency***

Database can manage concurrent data access effectively. It ensures no interference between users that would not result any loss of information nor loss of integrity.

- ***Multiple User Interface***

DBMS provides a variety of user interface like query language for casual users, programming language interface for application programmers, command codes for parametric users, menu-driven interface for standalone users. It provides web based GUI interface to database.

- ***Improved backup and recovery services***

Modern database management system provides facilities to minimize the amount of processing that can be lost following a failure by using the transaction approach.

Functions of DBMS

Data Storage Management: It provides a mechanism for management of permanent storage of the data. The internal schema defines how the data should be stored by the storage management mechanism and the storage manager interfaces with the operating system to access the physical storage.

Data Manipulation Management: A DBMS furnishes users with the ability to retrieve, update and delete existing data in the database.

Data Definition Services: The DBMS accepts the data definitions such as external schema, the conceptual schema, the internal schema, and all the associated mappings in source form.

Data Dictionary/System Catalog Management: The DBMS provides a data dictionary or system catalog function in which descriptions of data items are stored and which is accessible to users.

Database Communication Interfaces: The end-user's requests for database access are transmitted to DBMS in the form of communication messages.

Authorization / **Security Management**: The DBMS protects the database against unauthorized access, either international or accidental. It furnishes mechanism to ensure that only authorized users an access the database.

Backup and Recovery Management: The DBMS provides mechanisms for backing up data periodically and recovering from different types of failures. This prevents the loss of data,

Concurrency Control Service: Since DBMSs support sharing of data among multiple users, they must provide a mechanism for managing concurrent access to the database. DBMSs ensure that the database kept in consistent state and that integrity of the data is preserved.

Transaction Management: A transaction is a series of database operations, carried out by a single user or application program, which accesses or changes the contents of the database. Therefore, a DBMS must provide a mechanism to ensure either that all the updates corresponding to a given transaction are made or that none of them is made.

Database Access and Application Programming Interfaces: All DBMS provide interface to enable applications to use DBMS services. They provide data access via Structured Query Language (SQL). The DBMS query language contains two components: (a) a Data Definition Language (DDL) and (b) a Data Manipulation Language (DML).

Database models

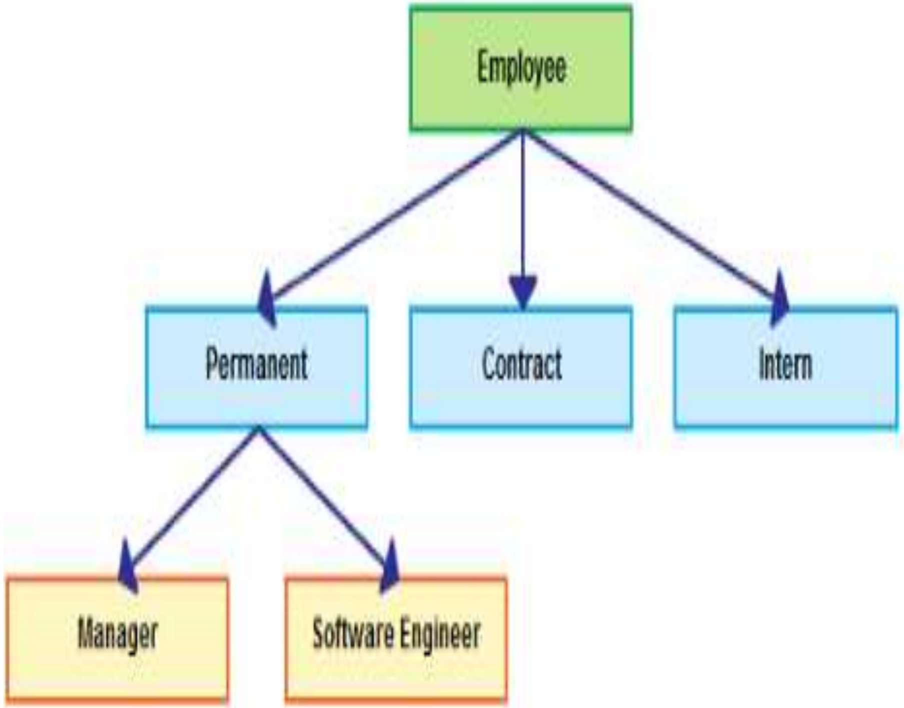
- Data models define how the logical structure of a database is modeled. Data Models are fundamental entities to introduce abstraction in a DBMS. Data models define how data is connected to each other and how they are processed and stored inside the system.
- A **Data Model** is a logical structure of Database. It describes the design of database to reflect entities, attributes, relationship among data, constrains etc.
- The very first data model could be flat data-models, where all the data used are to be kept in the same plane. Earlier data models were not so scientific, hence they were prone to introduce lots of duplication and update anomalies.

Types of database models

- Hierarchical Model
- Network Model
- Entity-Relational Model
- Relational Model

Hierarchical Model

- In **hierarchical model**, data is organized into a tree like structure with each record is having one parent record and many children.
- A hierarchical model represents the data in a tree-like structure in which there is a single parent for each record. To maintain order there is a sort field which keeps sibling nodes into a recorded manner. These types of models are designed basically for the early mainframe database management systems, like the Information Management System (IMS) by IBM.
- This model structure allows the one-to-one and a one-to-many relationship between two/ various types of data. This structure is very helpful in describing many relationships in the real world; table of contents, any nested and sorted information.
- The main drawback of this model is that, it can have only one to many relationships between nodes.



Advantages of Hierarchical Model

- Easy to understand
- Conceptual Simplicity
- Data Independence
- Performance is better than relational data model

Disadvantages of Hierarchical Model

- Difficult to access values at lower level
- Complex implementation
- This model may not be flexible to accommodate the dynamic needs of an organization
- Deletion of parent node result in deletion of child node forcefully
- Extra space is required for the storage of pointers

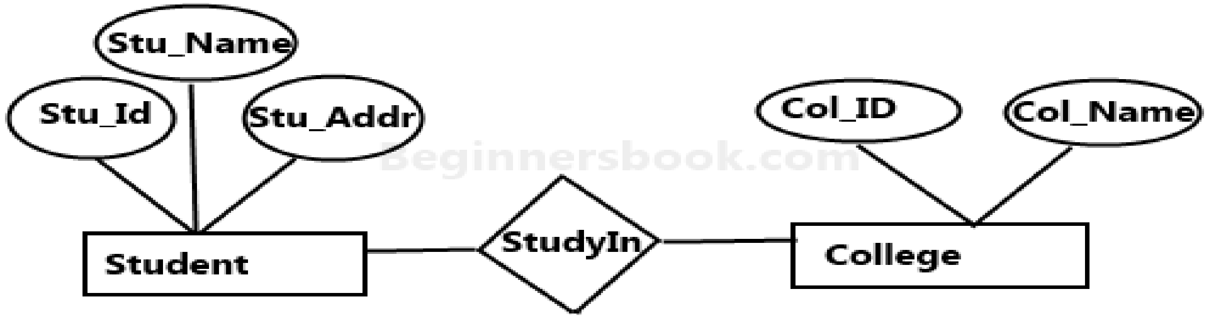
Entity-Relational Model

- Entity-Relationship (ER) Model is based on the notion of real-world entities and relationships among them. While formulating real-world scenario into the database model, the ER Model creates entity set, relationship set, general attributes and constraints.
- ER Model is best used for the conceptual design of a database.

Here are the geometric shapes and their meaning in an E-R Diagram :-

- **Rectangle:** Represents Entity sets.
- **Ellipses/Oval**: Attribues
- **Diamonds:** Relationship Set
- **Lines:** They link attributes to Entity Sets and Entity sets to Relationship Set
- **Double Ellipses:** Multivalued Attributes
- **Dashed Ellipses:** Derived Attributes
- **Double Rectangles:** Weak Entity Sets
- **Double Lines:** Total participation of an entity in a relationship set

A sample E-R Diagram:

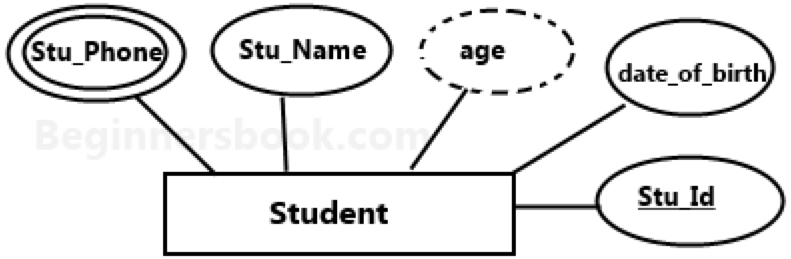


Sample E-R Diagram

- **Entity** − An entity in an ER Model is a real-world entity having properties called **attributes**.
- Every **attribute** is defined by its set of values called **domain**. For example, in a school database, a student is considered as an entity. Student has various attributes like name, age, class, etc.
- **Relationship** − The logical association among entities is called ***relationship***. Relationships are mapped with entities in various ways. Mapping cardinalities define the number of association between two entities.
- Mapping cardinalities −
  - one to one
  - one to many
  - many to one
  - many to many

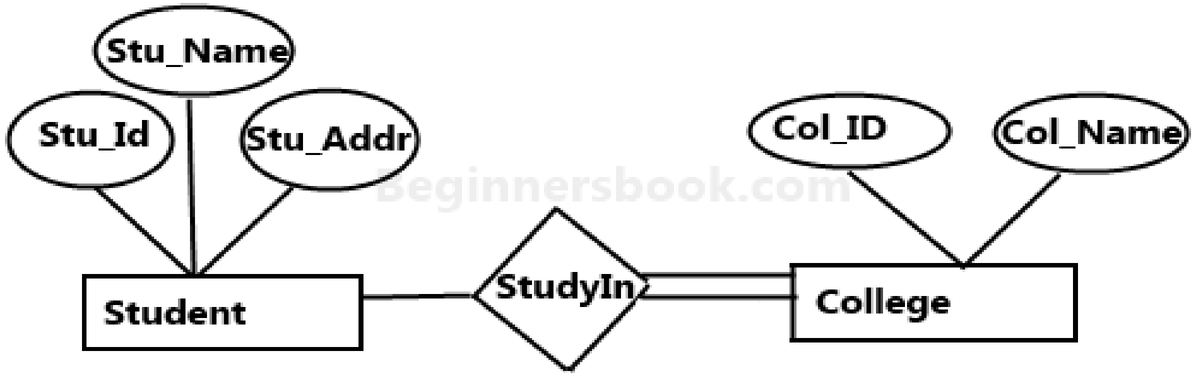
Multivalued Attributes: An attribute that can hold multiple values is known as multivalued attribute. We represent it with double ellipses in an E-R Diagram. E.g. A person can have more than one phone numbers so the phone number attribute is multivalued.
Derived Attribute: A derived attribute is one whose value is dynamic and derived from another attribute. It is represented by dashed ellipses in an E-R Diagram. E.g. Person age is a derived attribute as it changes over time and can be derived from another attribute (Date of birth).

E-R diagram with multivalued and derived attributes:



Total Participation of an Entity set:

A Total participation of an entity set represents that each entity in entity set must have at least one relationship in a relationship set. For example: In the below diagram each college must have at-least one associated Student.



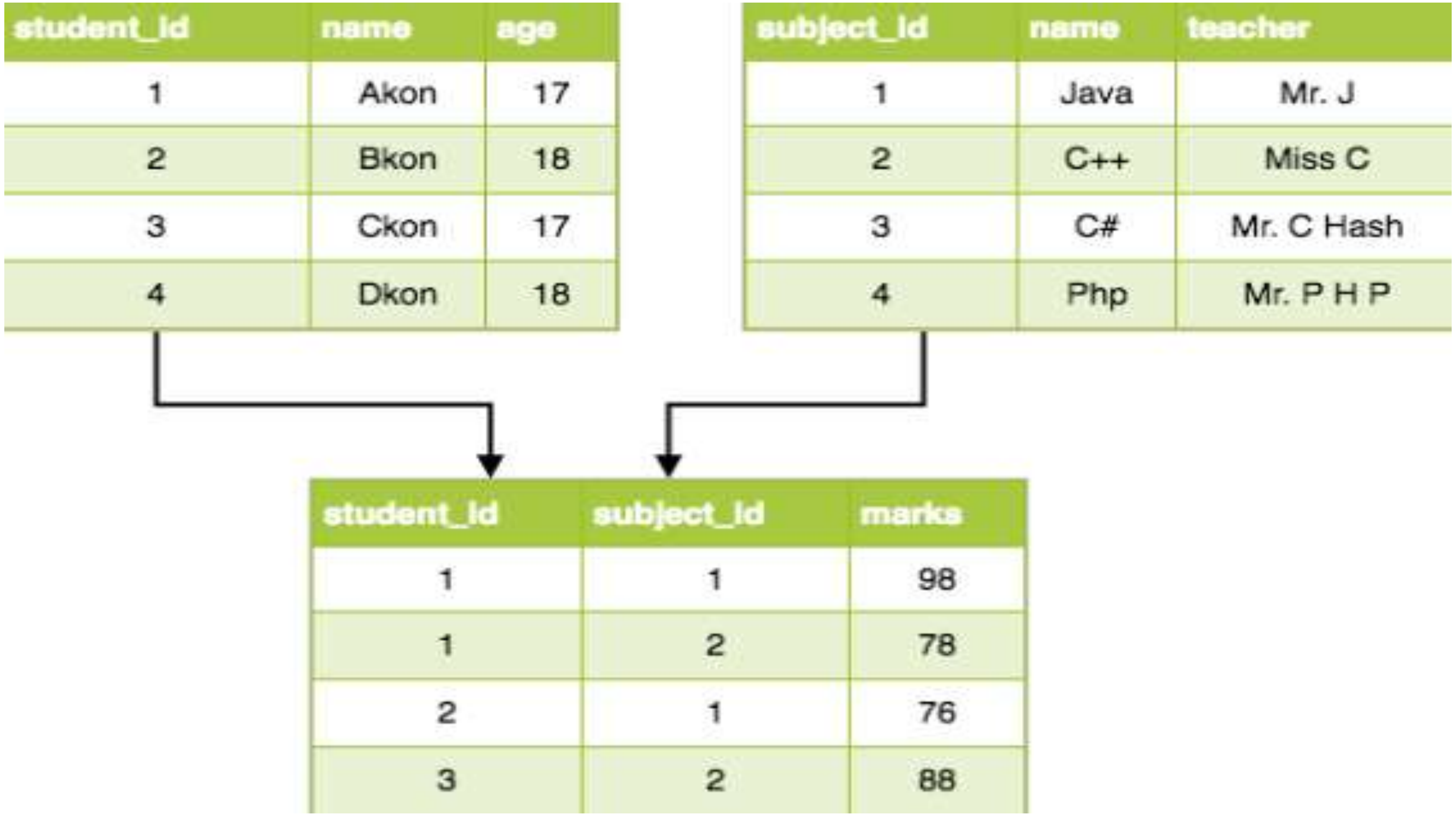
E-R Digram with total participation of College entity set in StudyIn relationship Set - This indicates that each college must have atleast one associated Student.

Relational model

- In relational model, the data and relationships are represented by collection of inter-related tables. Each table is a group of column and rows, where column represents attribute of an entity and rows represents records.
- Relational data model is the primary data model, which is used widely around the world for data storage and processing. This model is simple and it has all the properties and capabilities required to process data with storage efficiency.

Basic Concepts

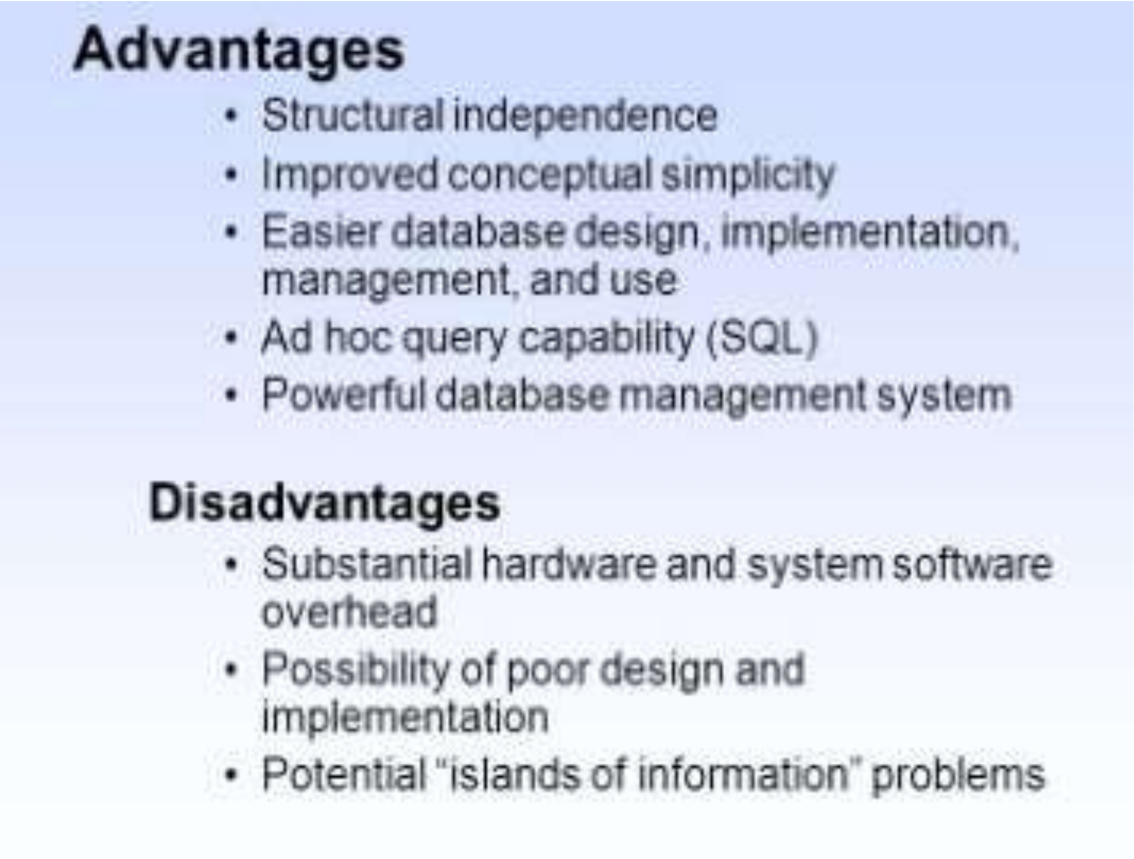
- **Tables** − In relational data model, relations are saved in the format of Tables. This format stores the relation among entities. A table has rows and columns, where rows represents records and columns represent the attributes.
- **Tuple** − A single row of a table, which contains a single record for that relation is called a tuple.
- **Relation instance** − A finite set of tuples in the relational database system represents relation instance. Relation instances do not have duplicate tuples.
- **Relation schema** − A relation schema describes the relation name (table name), attributes, and their names.
- **Relation key** − Each row has one or more attributes, known as relation key, which can identify the row in the relation (table) uniquely.
- **Attribute domain** − Every attribute has some pre-defined value scope, known as attribute domain.



student_id	name	age
1	Akon	17
2	Bkon	18
3	Ckon	17
4	Dkon	18

subject_id	name	teacher
1	Java	Mr. J
2	C++	Miss C
3	C#	Mr. C Hash
4	Php	Mr. P H P

student_id	subject_id	marks
1	1	98
1	2	78
2	1	76
3	2	88



Advantages

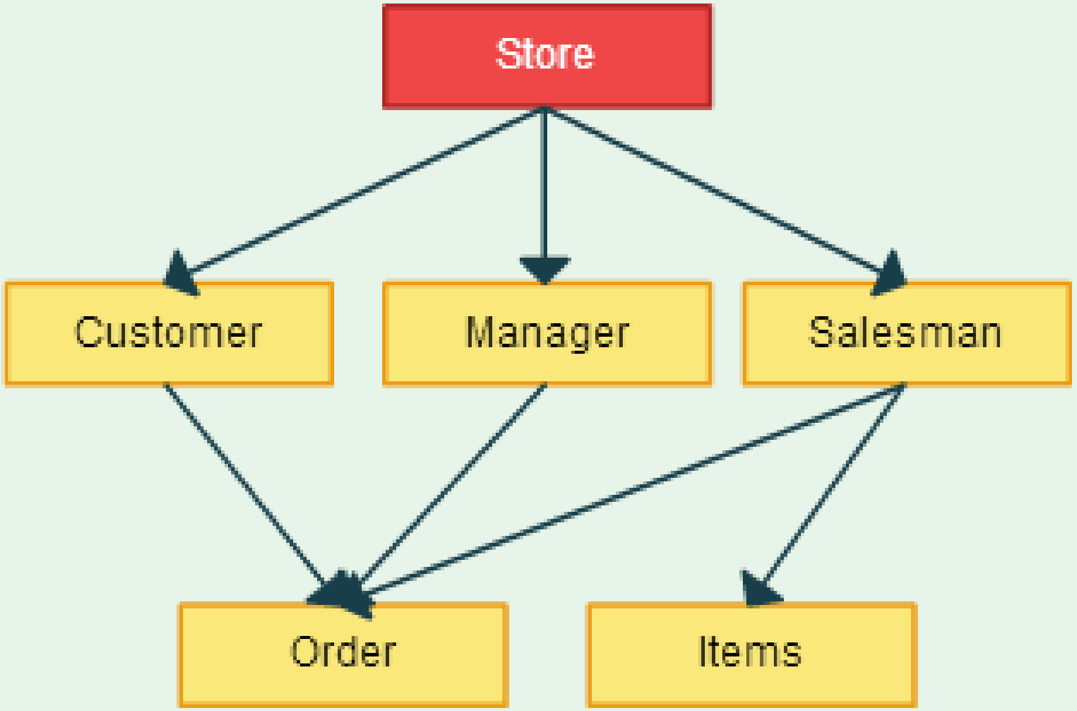
- Structural independence
- Improved conceptual simplicity
- Easier database design, implementation, management, and use
- Ad hoc query capability (SQL)
- Powerful database management system

Disadvantages

- Substantial hardware and system software overhead
- Possibility of poor design and implementation
- Potential "islands of information" problems

Network Model

- A network database model is a database model that allows multiple records to be linked to the same owner file.
- It is an extension of hierarchical database model.
- The model can be seen as an upside down tree where the branches are the member information linked to the owner, which is the bottom of the tree.
- The multiple linkages which this information allows the network database model to be very flexible.
- In addition, the relationship that the information has in the network database model is defined as many-to-many relationship because one owner file can be linked to many member files and vice versa.
- It describes data and relations between data by using graph rather than tree like structure.



Database design done using network model. In the network model a node can have multiple parent nodes.

ADVANTAGES OF NETWORK MODEL

1.) Conceptual simplicity-Just like the hierarchical model, the network model is also conceptually simple and easy to design.

2.) Capability to handle more relationship types-The network model can handle the one to many and many to many relationships which is real help in modeling the real life situations.

3.) Ease of data access-The data access is easier and flexible than the hierarchical model.

4.) Data integrity- The network model does not allow a member to exist without an owner.

5.) Data independence- The network model is better than the hierarchical model in isolating the programs from the complex physical storage details.

6.) Database standards

DISADVANTAGE OF NETWORK MODEL

1.) System complexity- All the records are maintained using pointers and hence the whole database structure becomes very complex.

2.) Operational Anomalies- The insertion, deletion and updating operations of any record require large number of pointers adjustments.

3.) Absence of structural independence-structural changes to the database is very difficult.

Structured Query Language(SQL)

- SQL is a standard language for accessing and manipulating databases.
- SQL became a standard of the American National Standards Institute (ANSI) in 1986, and of the International Organization for Standardization (ISO) in 1987
- RDBMS is the basis for SQL, and for all modern database systems such as MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.
- The data in RDBMS is stored in database objects called tables. A table is a collection of related data entries and it consists of columns and rows.

What Can SQL do?

- SQL can execute queries against a database
- SQL can retrieve and insert data from a database
- SQL can update records in a database
- SQL can delete records from a database
- SQL can create new databases
- SQL can create new tables in a database
- SQL can create stored procedures in a database
- SQL can create views in a database
- SQL can set permissions on tables, procedures, and views

Applications of SQL (Structured Query Language)

- **Data Integration Scripts**

The main application of SQL is to write data integration scripts by the *database administrators* and *developers.*

- **Analytical Queries**

The data analysts use structured query language for setting and running analytical queries on a regular basis.

- **Retrieve Information**

Another popular application of this language is to retrieve the subsets of information within a database for analytics applications and transaction processing. The most commonly used SQL elements are select, insert, update, add, delete, create, truncate and alter.

- **Other Important Applications**

The SQL is used for modification of the index structures and database table. Additionally, the users can add, update and delete the rows of the data by using this language.

Advantages of SQL

- **No coding needed**

  It is very easy to manage the database systems without any need to write the substantial amount of code by using the standard SQL.

- **Well defined standards**

  Long established are used by the SQL databases that is being used by ISO and ANSI. There are no standards adhered by the non-SQL databases.

- **Portability**

  SQL can be used in the program in PCs, servers, laptops, and even some of the mobile phones.

- **Interactive Language**

  This domain language can be used for communicating with the databases and receive answers to the complex questions in seconds.

- **Multiple data views**

  With the help of SQL language, the users can make different views of database structure and databases for the different users.

Disadvantages of SQL

- **Difficult Interface**

  SQL has a complex interface that makes it difficult for some users to access it.

- **Partial Control**

  The programmers who use SQL doesn’t have a full control over the database because of the hidden business rules.

- **Implementation**

  Some of the databases go to the proprietary extensions to standard SQL for ensuring the vendor lock-in.

- **Cost**

The operating cost of some SQL versions makes it difficult for some programmers to access it.

- **SQL Queries** SQL (pronounced "ess-que-el") is the acronym for Structured Query Language. SQL is the standard language for communicating with relational database management systems. SQL statements are used to retrieve data from the database as well as perform tasks such as adding updating and deleting the data. Some common relational database management systems that use SQL are: Oracle, Sybase, MS SQL Server,MS Access, MySQL, etc.
- **Basic Query Structure** SELECT field1 [,"field2",etc]
  FROM table
  [WHERE "condition"]
  [GROUP BY "field"]
  [ORDER BY "field"]
  [ ] = optional

  The **SELECT** statement is used to query the database and retrieve the fields that you specify. You can select as many fields (column names) as you want, or use the asterisk symbol "*" to select all fields.

  The **FROM** statement specifies the table names that will be queried to retrieve the desired data.

  The **WHERE** clause (optional) specifies which data values or rows will be returned or displayed, based on the criteria you specify.

  The **GROUP BY** clause (optional) organizes data into groups.

  The **ORDER BY** clause (optional) sorts the data by the field specified.

Database Languages: DDL, DML, DCL

- Language that are used to interact with database management system are known as database Language.
- A DBMS must provide appropriate languages and interfaces for each category of users to express database queries and updates.
- Database Languages are used to create and maintain database on computer.
- Types of database languages are Data Definition Language (DDL), Data Manipulation Language (DML) and Data Control Language (DCL) .

Data Definition Language (DDL)

- DDL is used for specifying the database structure or schema(Eg. Create,delete or modify)
- It is a type of language that allows the DBA or user to depict and name those entities, attributes, and relationships that are required for the application along with any associated integrity and security constraints
- Here are the lists of tasks that come under DDL:

✓ CREATE - used to create objects in the database

✓ ALTER - used to alters the structure of the database

✓ DROP - used to delete objects from the database

✓ TRUNCATE - used to remove all records from a table, including all spaces allocated for the records are removed

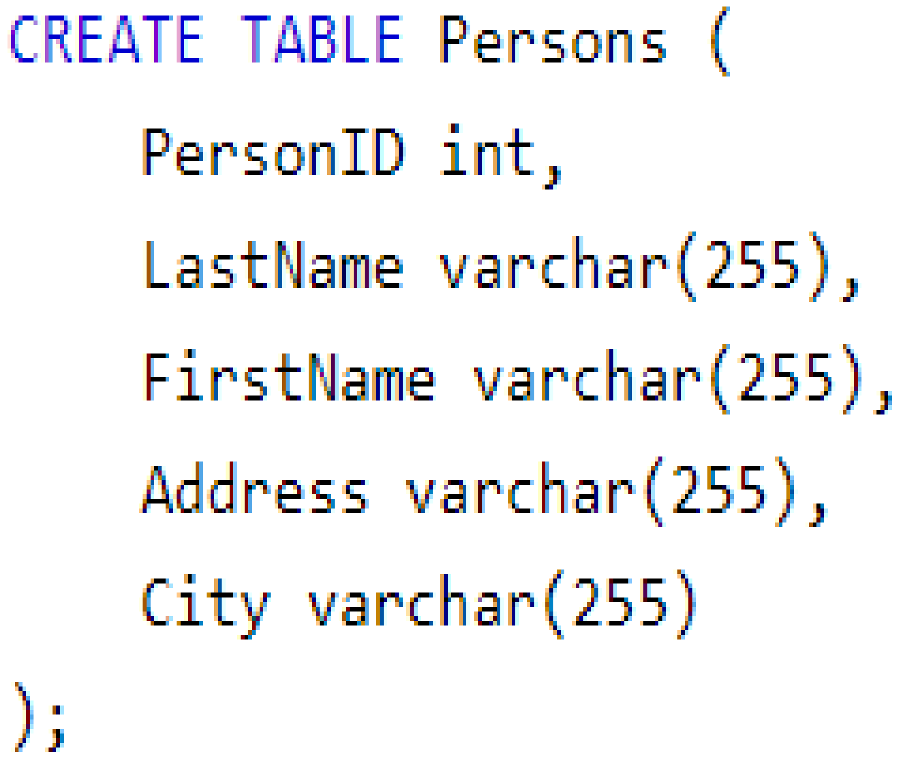
✓ COMMENT - used to add comments to the data dictionary

✓ RENAME - used to rename an object

Example of DDL statement

Syntax for create query

CREATE TABLE *table_name* (
  column1 datatype,
  column2 datatype,
  column3 datatype,
  ....
);



```
CREATE TABLE Persons (
    PersonID int,
    LastName varchar(255),
    FirstName varchar(255),
    Address varchar(255),
    City varchar(255)
);
```

Data Manipulation Language(DML)

- It is a language that provides a set of operations to support the basic data manipulation operations on the data held in the databases.
- It allows users to insert, update, delete and retrieve data from the database.
- DML is used for accessing and manipulating data in a database
- Basic DML Functions are:

✓ To read records from table(s) – **SELECT**

✓ To insert record(s) into the table(s) – **INSERT**

✓ Update the data in table(s) – **UPDATE**

✓ Delete all the records from the table – **DELETE**

Example of DML statement

Syntax

INSERT INTO *table_name*
VALUES (*value1*, *value2*, *value3*, ...);

Example

INSERT INTO Customers VALUES (‘Cardinal’, ’Stavanger’, ’Norway’);

Syntax

INSERT INTO *table_name* (*column1*, *column2*, *column3*, ...)
VALUES (*value1*, *value2*, *value3*, ...);

Example

INSERT INTO Customers (CustomerName, City, Country)
VALUES ('Cardinal', 'Stavanger', 'Norway');

Types of DML

- Procedural DMLs: a user specifies what data are required and how to get those data. these are normally low level data manipulation languages. Relational algebra is an example of procedural DML.
- Nonprocedural DMLs: a user specifies what data are needed without specifying how to get those data. these are normally high level DML. SQL is an example of nonprocedural DML.
- SELECT * from student

WHERE roll_no=4;

Data Control Language (DCL)

- DCL statements control access to data and the database using statements such as GRANT and REVOKE.
- A privilege can either be granted to a User with the help of GRANT statement. The privileges assigned can be SELECT, ALTER, DELETE, EXECUTE, INSERT, INDEX etc. In addition to granting of privileges, you can also revoke (taken back) it by using REVOKE command.

✓To grant access to user – GRANT

✓To revoke access from user – REVOKE

- In practical data definition language, data manipulation language and data control languages are not separate language; rather they are the parts of a single database language such as SQL.

- GRANT privilege_name
  ON object_name
  TO {user_name |PUBLIC |role_name}
  [WITH GRANT OPTION];
- ***privilege_name*** is the access right or privilege granted to the user. Some of the access rights are ALL, EXECUTE, and SELECT.
- ***object_name*** is the name of an database object like TABLE, VIEW, STORED PROC and SEQUENCE.
- ***user_name*** is the name of the user to whom an access right is being granted.
- ***user_name*** is the name of the user to whom an access right is being granted.
- ***PUBLIC*** is used to grant access rights to all users.
- ***ROLES*** are a set of privileges grouped together.
- ***WITH GRANT OPTION*** - allows a user to grant access rights to other users.
- GRANT SELECT ON employee TO user1; This command grants a SELECT permission on employee table to user1.You should use the WITH GRANT option carefully because for example if you GRANT SELECT privilege on employee table to user1 using the WITH GRANT option, then user1 can GRANT SELECT privilege on employee table to another user, such as user2 etc. Later, if you REVOKE the SELECT privilege on employee from user1, still user2 will have SELECT privilege on employee table.

Database design

- A database is a large repository of facts, designed in such a way that processing the facts into information is easy. If the phone book was structured in a less convenient way, such as with names and numbers placed in chronological order according to when the numbers were issued, converting the data into information would be much more difficult.

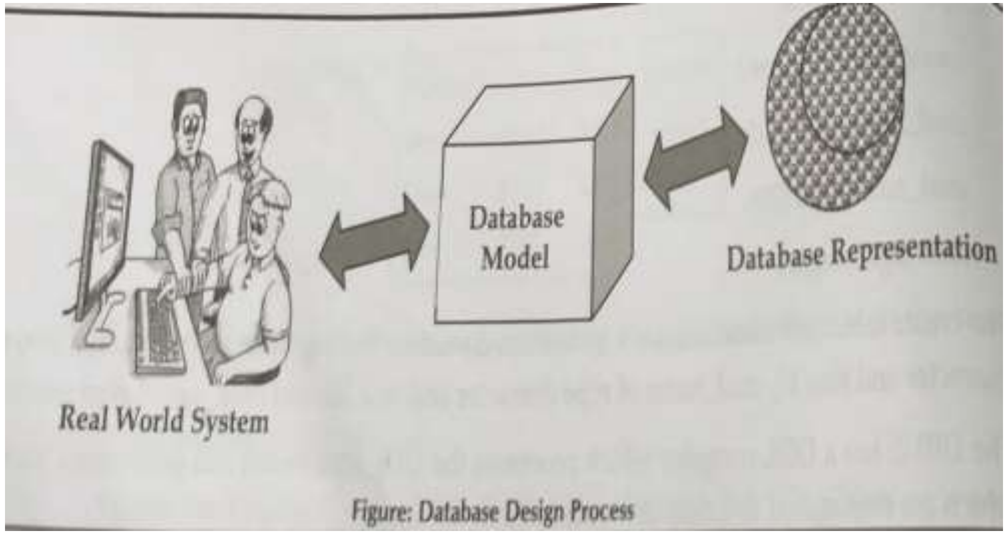


Figure: Database Design Process

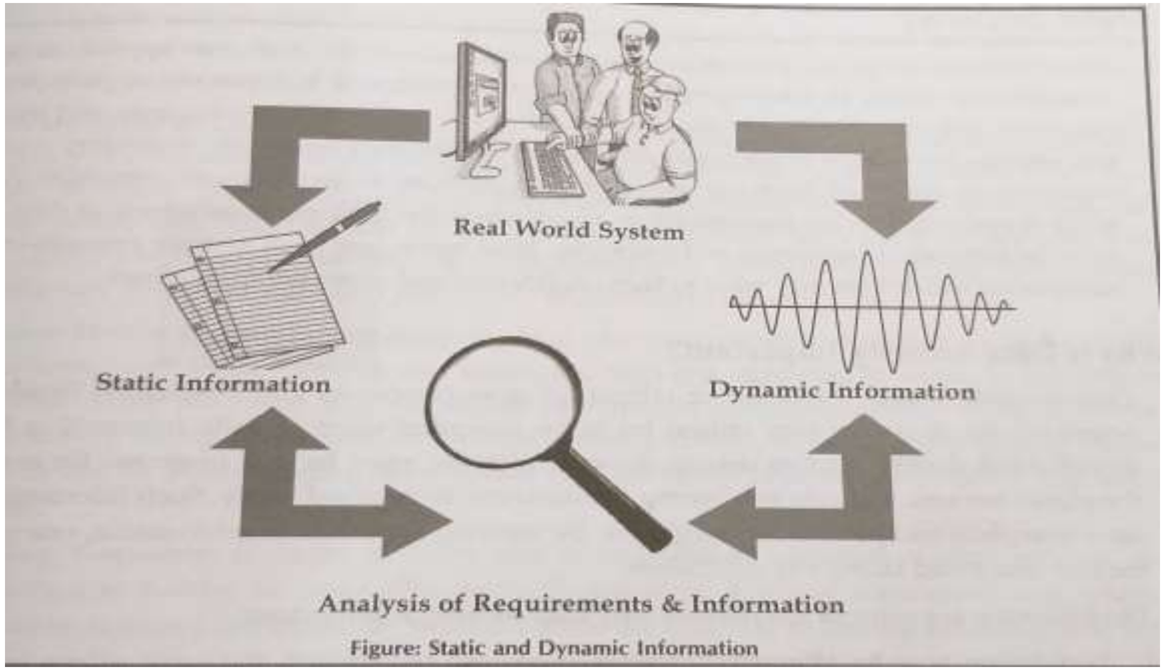


Figure: Static and Dynamic Information

Data Security

- Data security is a set of standards and technologies that protect data from intentional or accidental destruction, modification or disclosure. Data security can be applied using a range of techniques and technologies, including administrative controls, physical security, logical controls, organizational standards, and other safeguarding techniques that limit access to unauthorized or malicious users or processes.
- Database of an enterprises consist of huge amount of data collected from different business activities in a course of time. The data hence collected carries a significant importance to the enterprise. Those data should be carefully stored so that the database or certain part of the database does not get lost, altered, or accessed by unauthorized person. Therefore to protect the database from such events database security plays a vital role.
- Database security concerns the use of a broad range of information security controls to protect databases against compromises of their confidentiality, integrity and availability. Database security helps to control the access of the database objects, formulate the rules regarding modification of data, auditing of the data, ensure the privacy, etc. For designing a good database security, we need to consider three main objectives:

- ➢ **Secrecy**
- ➢ **Integrity**
- ➢ **Availability**

Secrecy: The data should only be accessed by the authorized user. The access to data by unauthorized user should always be denied. Sometime a user may be denied of the whole data and activity in a given database or only a part of data and certain activities associated with the database.

E.g.: *A Bank customer should not be allowed to check the details of other customers but can have authority to check their own detail.*

Integrity: Modification of data may be allowed to only authorize user. The accidental modification or intentional modification should always be protected. The user should be allowed to modify only those part of data that they need.

E.g.: *A customers should not be allowed to change the balance of his/her account through database application.*

Availability: Security should not restrict the authorized user to perform their actions on the part of the database available for them i.e. authorized user should not be denied access.

E.g.: *All the authorized employee should be able to change the data and information about the customer in their bank.*

Why do we need Database Security?

Database consists of all the information that is generated by an enterprise. The data in database are very important asset for any organization, it might be confidential and sensitive. In order to protect such data from intruders and from being modified or deleted accidentally database security is very essential. We have a belief that most of the security breaches are caused by the hackers but in fact 80% of the data loss is caused by insider, this is also a reason why we need database security and provided only necessary access to each user within an organization.

Some of general reasons for need of database security are:

i. To maintain data consistency when large number of user are accessing the same data in shared data environment.

ii. To control and restrict the intentional or accidental modification of the sensitive data by the insider.

iii. To protect data from hackers who tries to intrude the database through internet.

iv. Hackers have developed specialized software to illegally enter the system and extract the financial data. Therefore database security is needed to secure the monetary transactions performed through credit card using internet.

Essential steps Every Business Must Take to secure data

- Establish Strong Password
- Strong firewall
- Antivirus protection
- Secure system
- Secure mobile phones
- Backup regularly
- Monitor well
- Surf safely

Data Securing Technologies

- **Disk encryption**

  Disk encryption refers to encryption technology that encrypts data on a hard disk drive. Disk encryption typically takes form in either software or hardware . Disk encryption is often referred to as on-the-fly encryption (OTFE) or transparent encryption.
- **Data encryption**: Data encryption applies a code to every individual piece of data and will not grant access to encrypted data without an authorized key being given
- **Data masking**: Masking specific areas of data can protect it from disclosure to external malicious sources, and also internal personnel who could potentially use the data. For example, the first 12 digits of a credit card number may be masked within a database.
- **Data erasure**: There are times when data that is no longer active or used needs to be erased from all systems. For example, if a customer has requested for their name to be removed from a mailing list, the details should be deleted permanently.

- **Software versus hardware-based mechanisms for protecting data**

  Software-based security solutions encrypt the data to protect it from theft. However, a malicious program or a hacker could corrupt the data in order to make it unrecoverable, making the system unusable. Hardware-based security solutions can prevent read and write access to data and hence offer very strong protection against tampering and unauthorized access.

  Hardware based security or assisted computer security offers an alternative to software-only computer security. Security tokens such as those using PKCS#11 may be more secure due to the physical access required in order to be compromised. Access is enabled only when the token is connected and correct PIN is entered (see two-factor authentication). However, dongles can be used by anyone who can gain physical access to it. Newer technologies in hardware-based security solves this problem offering full proof security for data.

- **Data resilience**: By creating backup copies of data, organizations can recover data should it be erased or corrupted accidentally or stolen during a data breach.

Database Administrator

- The person who has the central control over a database system is called Database Administrator (DBA).
- A database administrator (DBA) directs or performs all activities related to maintaining a successful database environment. Responsibilities include designing, implementing, and maintaining the database system; establishing policies and procedures pertaining to the management, security, maintenance, and use of the database management system; and training employees in database management and use

The database administrator has the following functions in a database system.

- **Schema Definition:** The database administrator creates the original database schema by executing a set of data definition statements in DDL.
- **Storage structure an access method definition.**
- **Schema and physical or organization modification:** The database administrator performs the changes to the schema according to the needs of organizations or physical needs to improve the database performance.
- **Provide the granting of authorization to access data:** The database administrator can decide the which parts of the database can be accessed by a user, by using the different types of authorization methods.
- **Database maintenance:** The database maintenance includes the following processes.

➢ Regular backing up of the database.

➢ Ensuring the disk space for performing the required operations.

➢ Monitoring the jobs running on the database.

- **Installing and upgrading the DBMS Servers: -** DBA is responsible for installing a new DBMS server for the new projects. He is also responsible for upgrading these servers as there are new versions comes in the market or requirement. If there is any failure in upgradation of the existing servers, he should be able revert the new changes back to the older version, thus maintaining the DBMS working. He is also responsible for updating the service packs/ hot fixes/ patches to the DBMS servers.
- **Design and implementation: -** Designing the database and implementing is also DBA's responsibility. He should be able to decide proper memory management, file organizations, error handling, log maintenance etc. for the database.
- **Performance tuning: -** Since database is huge and it will have lots of tables, data, constraints and indices, there will be variations in the performance from time to time. Also, because of some designing issues or data growth, the database will not work as expected. It is responsibility of the DBA to tune the database performance. He is responsible to make sure all the queries and programs works in fraction of seconds.
- **Migrate database servers: -** Sometimes, users using oracle would like to shift to SQL server or Netezza. It is the responsibility of DBA to make sure that migration happens without any failure, and there is no data loss.
- **Backup and Recovery: -** Proper backup and recovery programs needs to be developed by DBA and has to be maintained him. This is one of the main responsibilities of DBA. Data/objects should be backed up regularly so that if there is any crash, it should be recovered without much effort and data loss.
- **Security: -** DBA is responsible for creating various database users and roles, and giving them different levels of access rights.
- **Documentation: -** DBA should be properly documenting all his activities so that if he quits or any new DBA comes in, he should be able to understand the database without any effort. He should basically maintain all his installation, backup, recovery, security methods. He should keep various reports about database performance.

Database Application architectures

- A Database Management system is not always directly available for users and applications to access and store data in it. A Database Management system can be **centralised**(all the data stored at one location), **decentralised**(multiple copies of database at different locations) or **hierarchical**, depending upon its architecture.

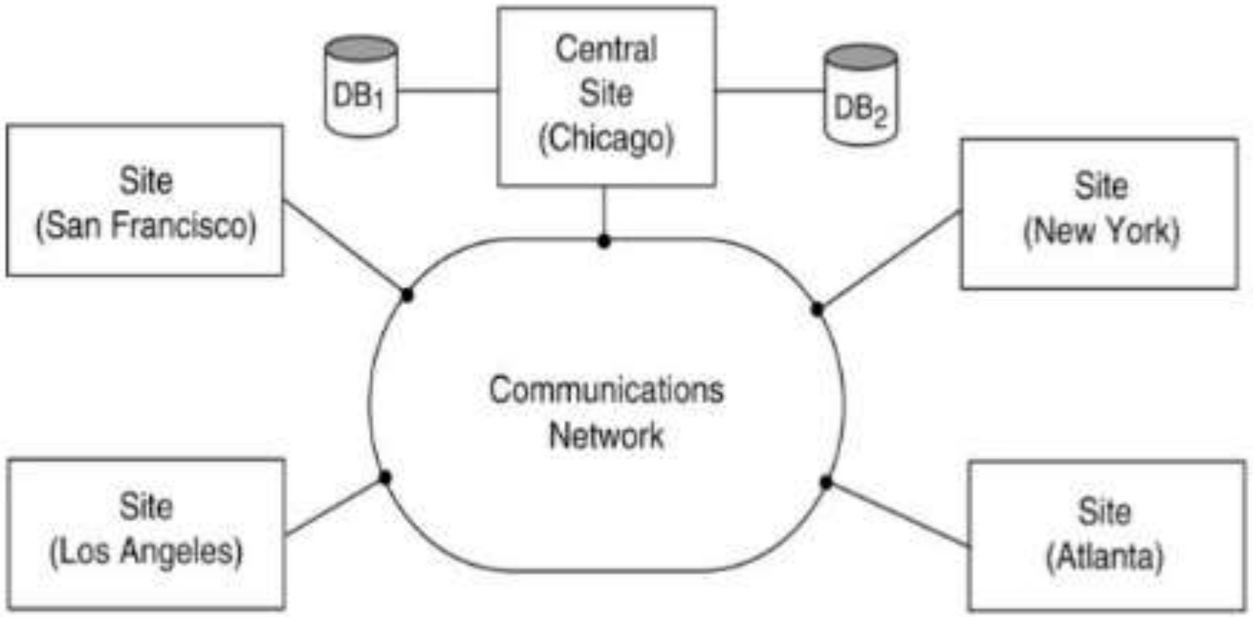
Centralized Architecture

- A **centralized database** (sometimes abbreviated **CDB**) is a database that is located, stored, and maintained in a single location. This location is most often a central computer or database system, for example a desktop or server CPU, or a mainframe computer. In most cases, a centralized database would be used by an organization (e.g. a business company) or an institution (e.g. a university.) Users access a centralized database through a computer network which is able to give them access to the central CPU, which in turn maintains to the database itself.

Centralized Databases

- Highly dependent on network connectivity.
- Slower the internet connection, longer time is needed to access database
- Bottlenecks can occur as a result of high traffic
- If no fault-tolerant setup is arranged and hardware failure occurs, all data within database is lost.
- Once data is lost, there's no way of recovering the data back.

Centralized database



Advantages

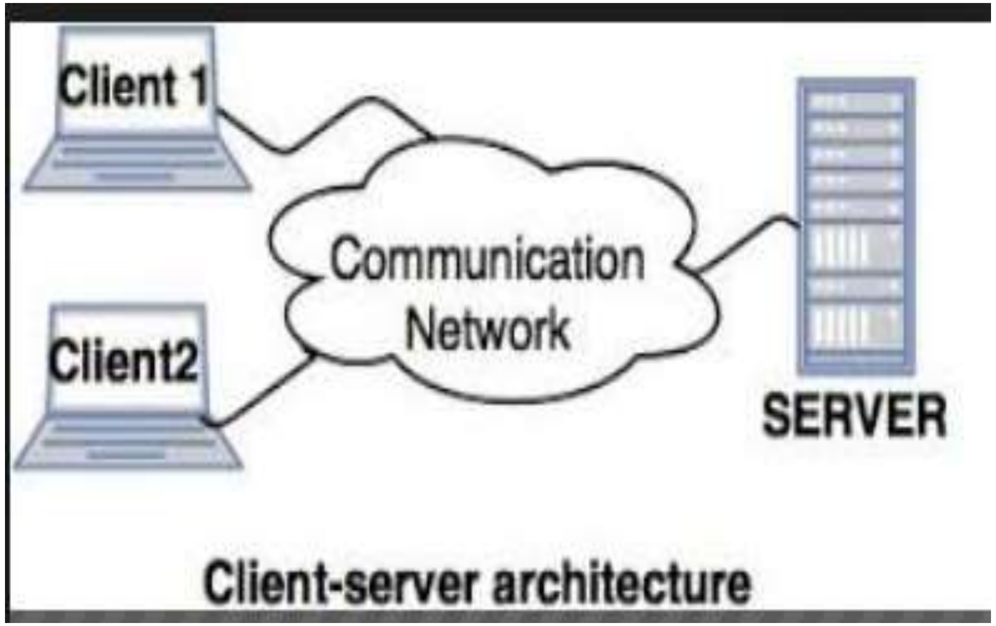
- **Data integrity** is maximized and **data redundancy** is minimized, as the single storing place of all the data also implies that a given set of data only has one primary record. This aids in the maintaining of data as accurate and as consistent as possible and enhances data reliability.
- Generally bigger **data security**, as the single data storage location implies only a one possible place from which the database can be attacked and sets of data can be stolen or tampered with.
- Better **data preservation** than other types of databases due to often-included fault-tolerant setup.
- Easier for using by the end-user due to the **simplicity** of having a single database design.
- Generally easier **data portability** and **database administration**.
- More **cost effective** than other types of database systems as labor, power supply and maintenance costs are all minimized.
- **Data kept** in the same location is easier to be changed, re-organized, mirrored, or analyzed.
- All the **information can be accessed** at the same time from the same location.
- **Updates** to any given set of data are immediately received by every end-user.

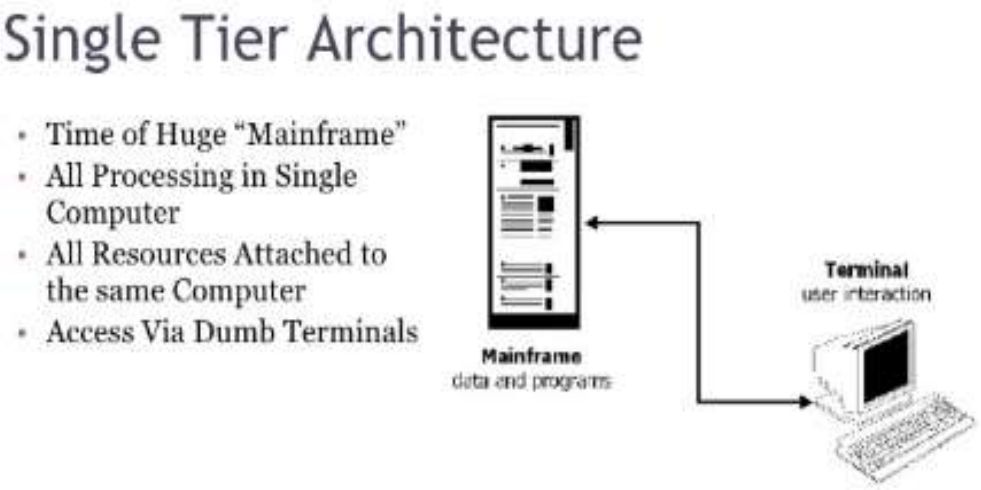
Disadvantages

- Centralized databases are highly dependent on **network connectivity**. The slower the internet connection is, the longer the database access time needed will be.
- **Bottlenecks** can occur as a result of high traffic.
- **Limited access** by more than one person to the same set of data as there is only one copy of it and it is maintained in a single location. This can lead to major decreases in the general efficiency of the system.
- If there is no **fault-tolerant setup** and **hardware failure** occurs, all the data within the database will be lost.
- Since there is minimal to no **data redundancy**, if a set of data is unexpectedly lost it is very hard to retrieve it back, in most cases it would have to be done manually.

Client/Server Database Systems

- Client/Server architecture of database system has two logical components namely client, and server.
- Clients are generally personal computers or workstations whereas server is large workstations, mini range computer system or a mainframe computer system.
- The applications and tools of DBMS run on one or more client platforms, while the DBMS software's reside on the server.
- The server computer is called backend and the client’s computer is called front end.
- These server and client computers are connected into a network.
- The applications and tools act as clients of the DBMS, making requests for its services. The DBMS, in turn, processes these requests and returns the results to the client(s).
- Client/Server architecture handles the Graphical User Interface (GUI) and does computations and other programming of interest to the end user.
- The server handles parts of the job that are common to many clients, for example, database access and updates.



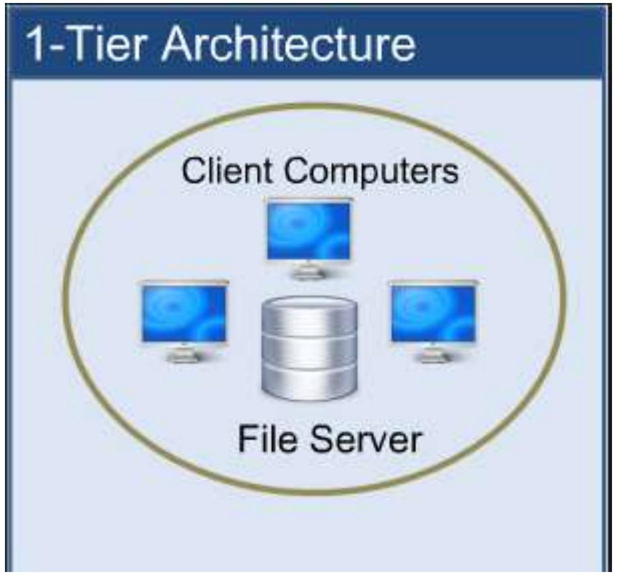


Single Tier Architecture

- Time of Huge "Mainframe"
- All Processing in Single Computer
- All Resources Attached to the same Computer
- Access Via Dumb Terminals

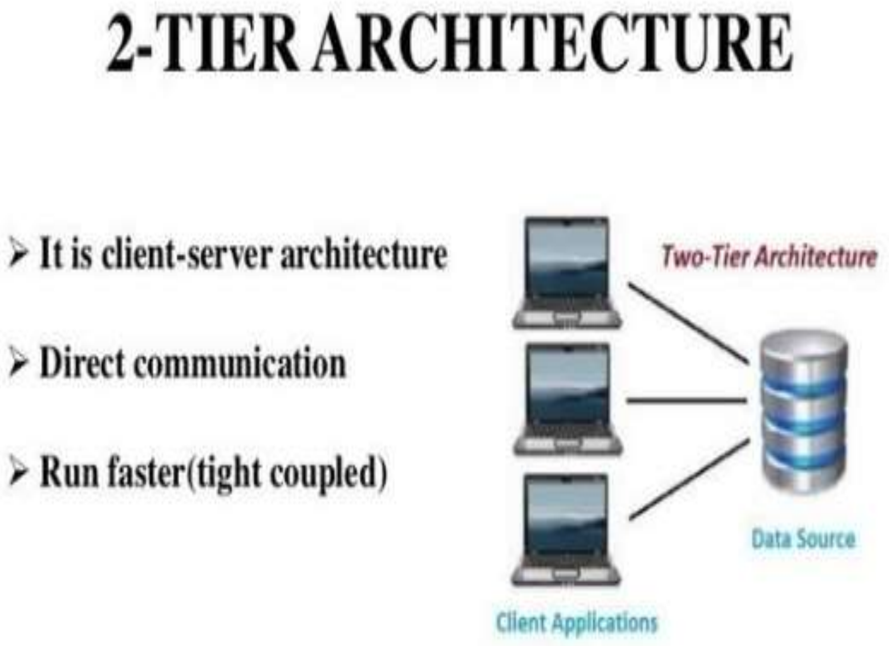
1-tier DBMS

- **In 1-tier architecture**, the DBMS is the only entity where the user directly sits on the DBMS and uses it. Any changes done here will directly be done on the DBMS itself. It does not provide handy tools for end-users. Database designers and programmers normally prefer to use single-tier architecture.
- Generally such a setup is used for local application development, where programmers communicate directly with the database for quick response.

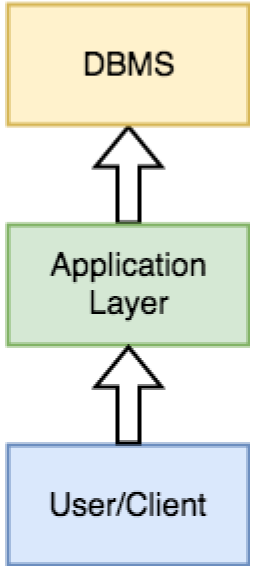


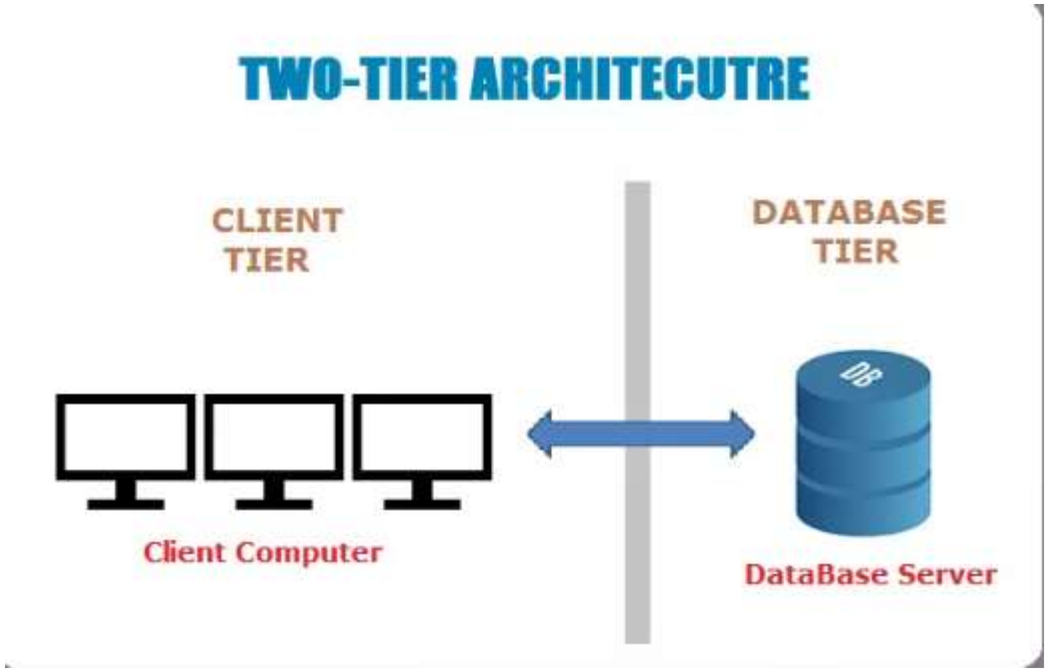
2-tier DBMS Architecture

- 2-tier DBMS architecture includes an **Application layer** between the user and the DBMS, which is responsible to communicate the user's request to the database management system and then send the response from the DBMS to the user.
- An application interface known as **ODBC**(Open Database Connectivity) provides an API that allow client side program to call the DBMS. Most DBMS vendors provide ODBC drivers for their DBMS.
- Such an architecture provides the DBMS extra security as it is not exposed to the End User directly. Also, security can be improved by adding security and authentication checks in the Application layer too.



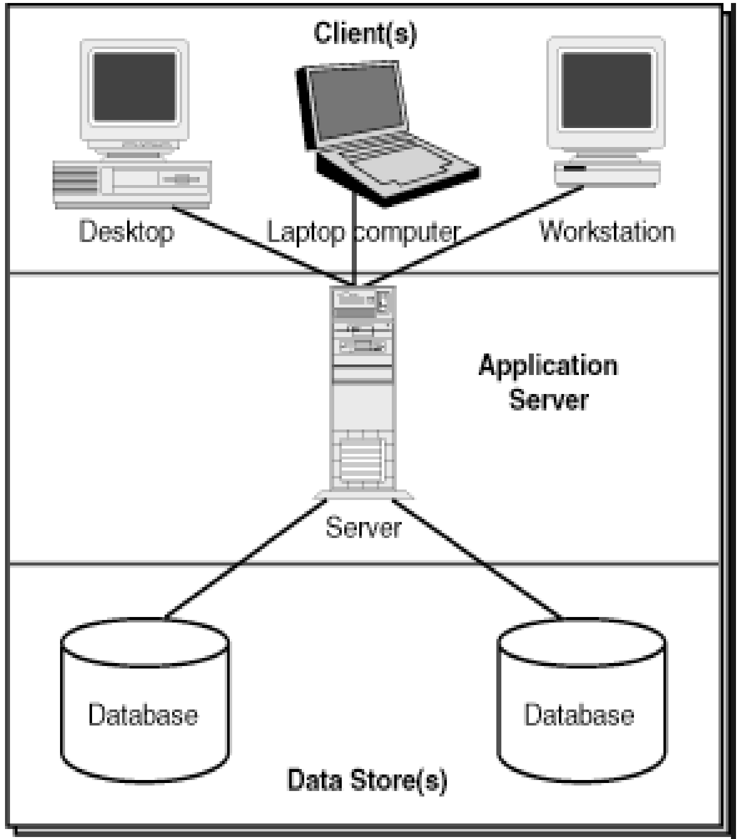
- It is client-server architecture
- Direct communication
- Run faster(tight coupled)

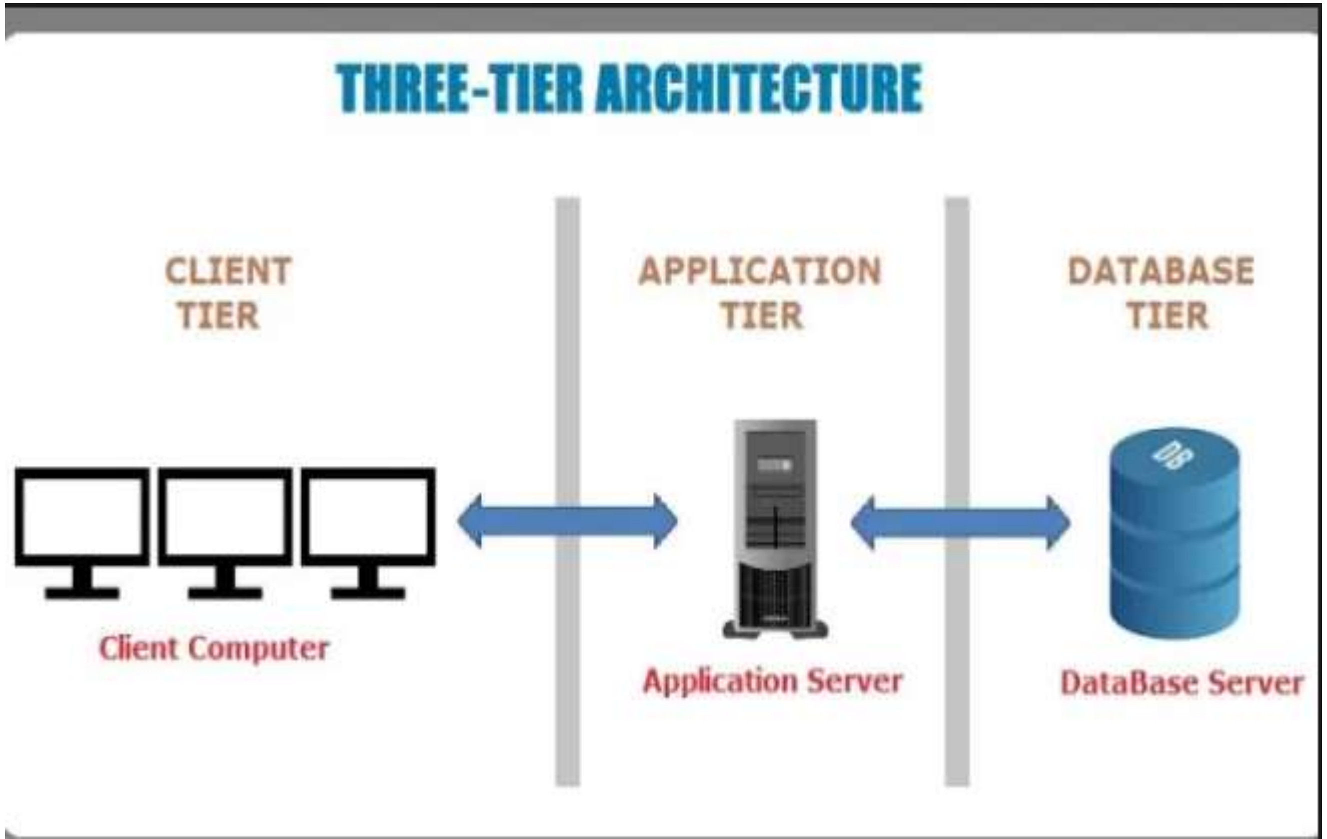




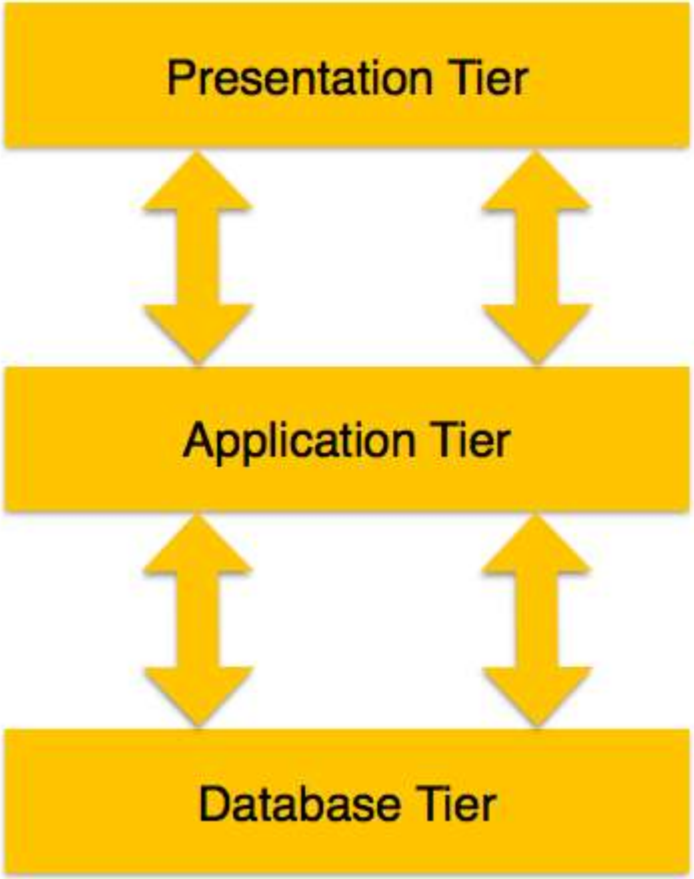
3-tier Architecture

- 3-tier DBMS architecture is the most commonly used architecture for web applications.it adds intermediate layer known as application server(or web server) between the client and database server.
- A 3-tier architecture separates its tiers from each other based on the complexity of the users and how they use the data present in the database. It is the most widely used architecture to design a DBMS.





- **Database (Data) Tier** – At this tier, the database resides along with its query processing languages. We also have the relations that define the data and their constraints at this level.
- **Application (Middle) Tier** – At this tier reside the application server and the programs that access the database. For a user, this application tier presents an abstracted view of the database. End-users are unaware of any existence of the database beyond the application. At the other end, the database tier is not aware of any other user beyond the application tier. Hence, the application layer sits in the middle and acts as a mediator between the end-user and the database.
- **User (Presentation) Tier** – End-users operate on this tier and they know nothing about any existence of the database beyond this layer. At this layer, multiple views of the database can be provided by the application. All views are generated by applications that reside in the application tier.



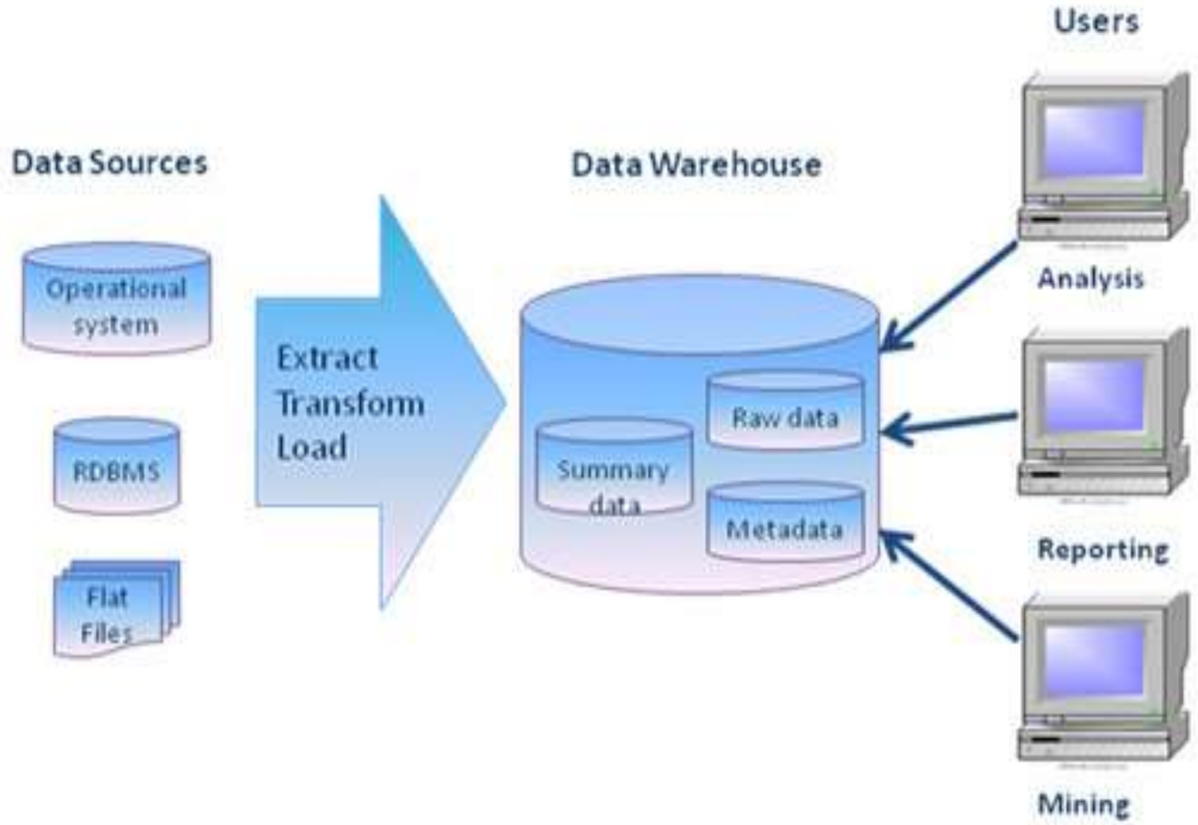
Data dictionary

A **data dictionary**, or metadata (data about data) repository, as defined in the ***IBM Dictionary of Computing***, is a "centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format."The term can have one of several closely related meanings pertaining to databases and database management systems (DBMS):

- A document describing a database or collection of databases
- An integral component of a DBMS that is required to determine its structure
- A piece of middleware that extends the native data dictionary of a DBMS

Data Warehouse

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.
- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.


Users
Data Sources
Data Warehouse
Operational system
Extract Transfom Load
RDBMS
Flat Files
Raw data
Summary data
Metadata
Analysis
Reporting
Mining

Data Warehouse Features/Characteristics

- **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. A data warehouse does not focus on the ongoing operations, rather it focuses on modeling and analysis of data for decision making.
- **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.
- **Time Variant** - The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.
- **Non-volatile** - Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in operational database is not reflected in the data warehouse.

Note: A data warehouse does not require transaction processing, recovery, and concurrency controls, because it is physically stored and separate from the operational database.

Applications of Data warehousing

Data warehouses are widely used in the following fields

- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing
- Insurance fraud analysis
- Logistic management

Components of Data Warehouse

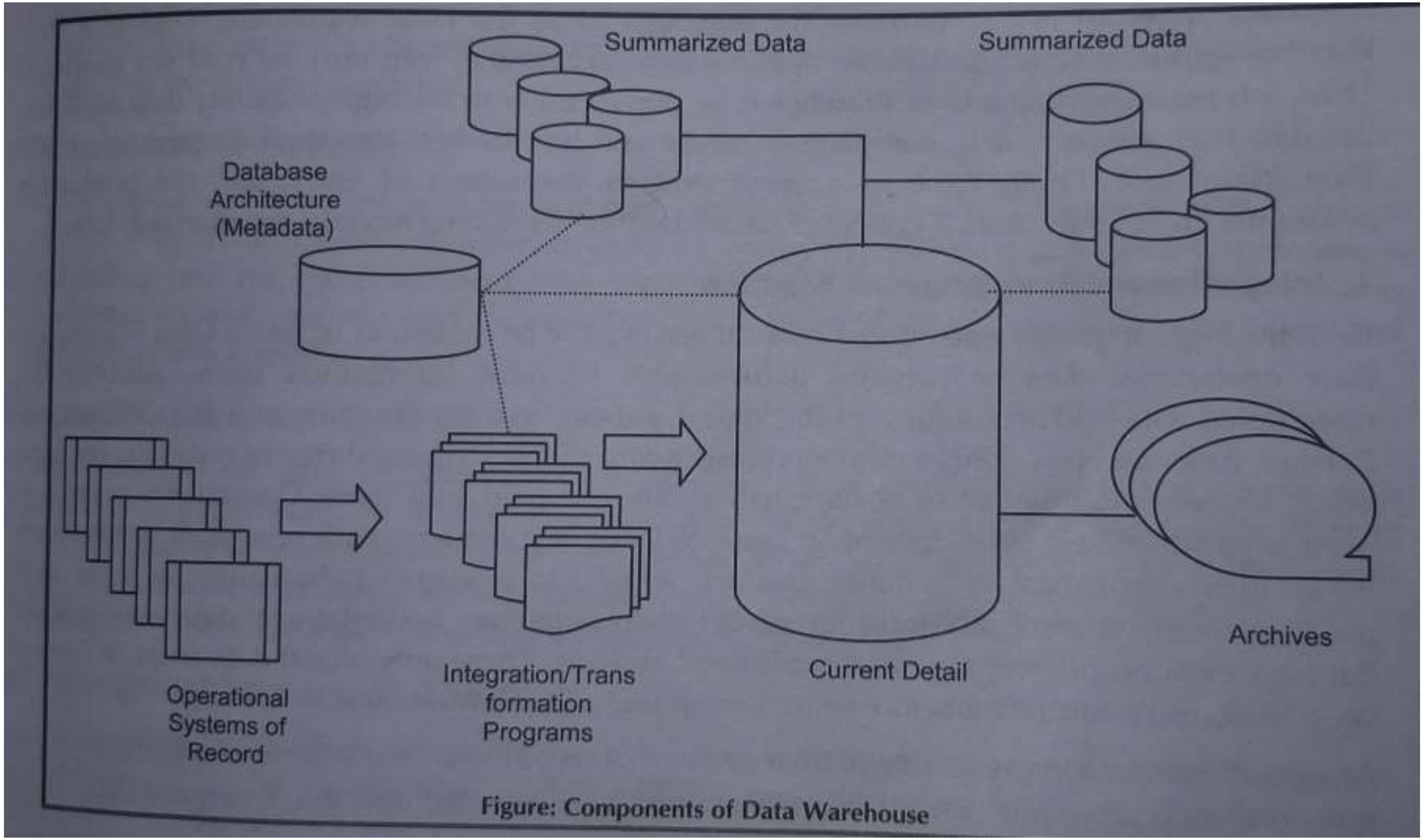


Figure: Components of Data Warehouse

Advantages of Data Warehouse:

- Data warehouse allows business users to quickly access critical data from some sources all in one place.
- Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
- Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
- Data warehouse helps to reduce total turnaround time for analysis and reporting.
- Restructuring and Integration make it easier for the user to use for reporting and analysis.
- Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
- Data warehouse stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.

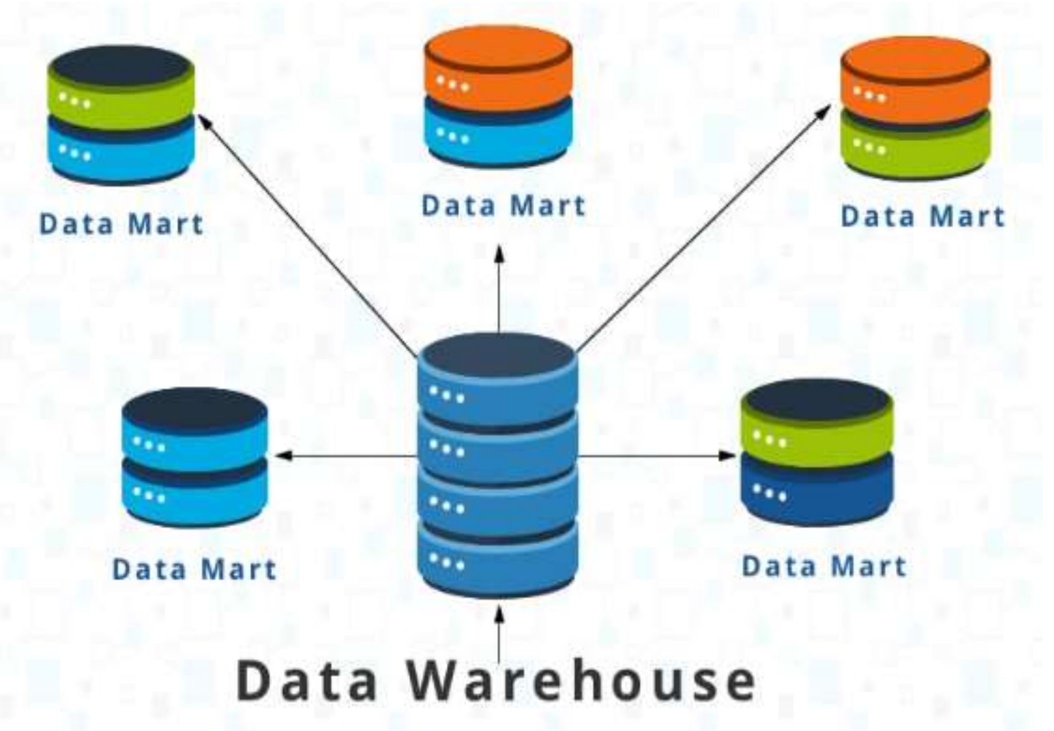
Disadvantages of Data Warehouse:

- Not an ideal option for unstructured data.
- Creation and Implementation of Data Warehouse is surely time confusing affair.
- Data Warehouse can be outdated relatively quickly
- Difficult to make changes in data types and ranges, data source schema, indexes, and queries.
- The data warehouse may seem easy, but actually, it is too complex for the average users.
- Despite best efforts at project management, data warehousing project scope will always increase.
- Sometime warehouse users will develop different business rules.
- Organizations need to spend lots of their resources for training and Implementation purpose.

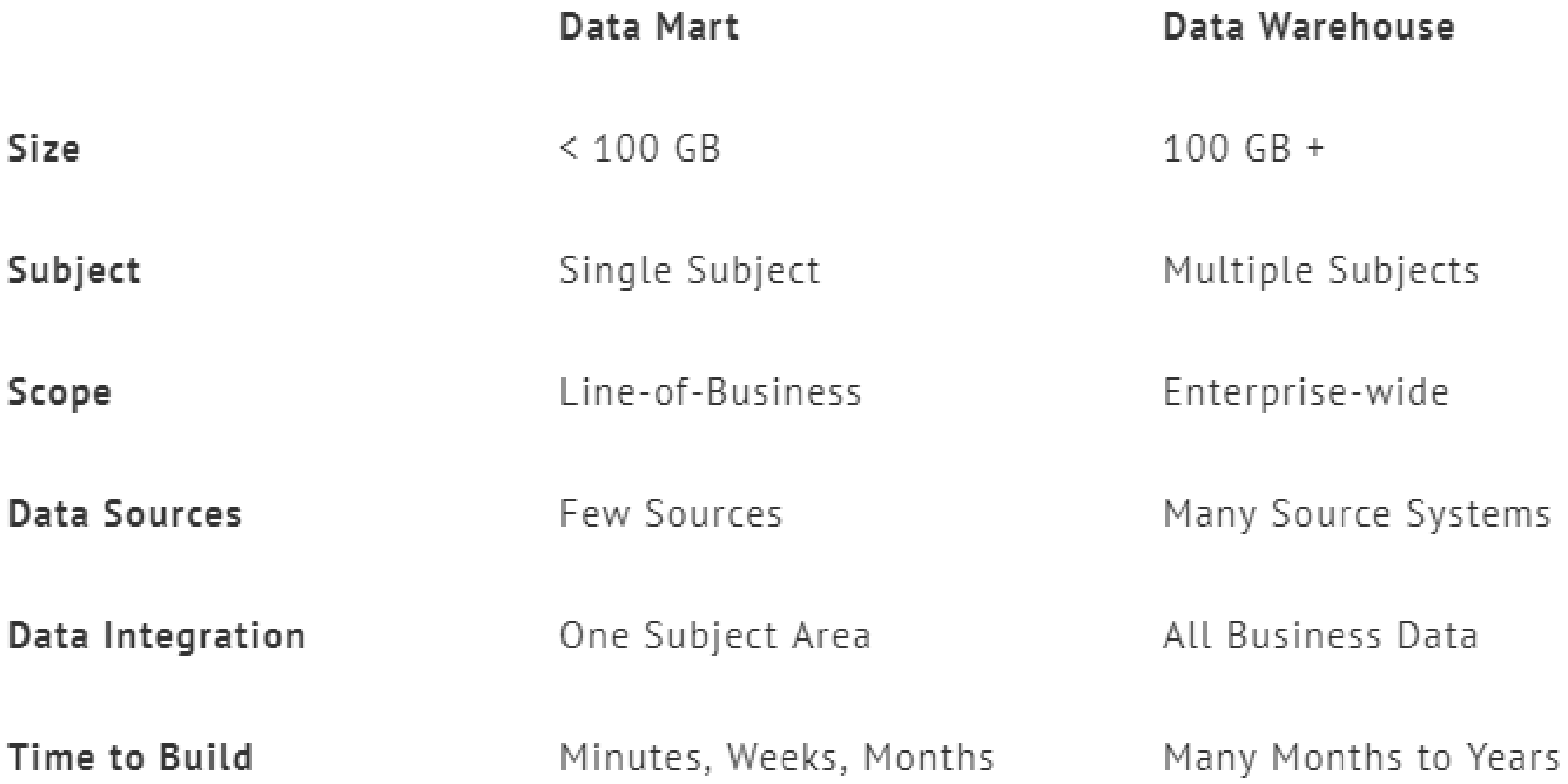
SN.	Data Warehouse (OLAP)	Operational Database(OLTP)
1	It involves historical processing of information.	It involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	It is used to analyze the business.	It is used to run the business.
4	It focuses on Information out.	It focuses on Data in.
5	It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
6	It focuses on Information out.	It is application oriented.
7	It contains historical data.	It contains current data.
8	It provides summarized and consolidated data.	It provides primitive and highly detailed data.
9	It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
10	The number of users is in hundreds.	The number of users is in thousands.
11	The number of records accessed is in millions.	The number of records accessed is in tens.
12	The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
13	These are highly flexible.	It provides high performance.

Data Mart

- The data mart is a subset of the data warehouse and is usually oriented to a specific business line or team. Whereas data warehouses have an enterprise-wide depth, the information in data marts pertains to a single department. In some deployments, each department or business unit is considered the *owner* of its data mart including all the *hardware*, *software* and *data*.
- A data mart is a subject-oriented database that is often a partitioned segment of an enterprise data warehouse. The subset of data held in a data mart typically aligns with a particular business unit like sales, finance, or marketing. Data marts accelerate business processes by allowing access to relevant information in a data warehouse or operational data store within days, as opposed to months or longer. Because a data mart only contains the data applicable to a certain business area, it is a cost-effective way to gain actionable insights quickly.



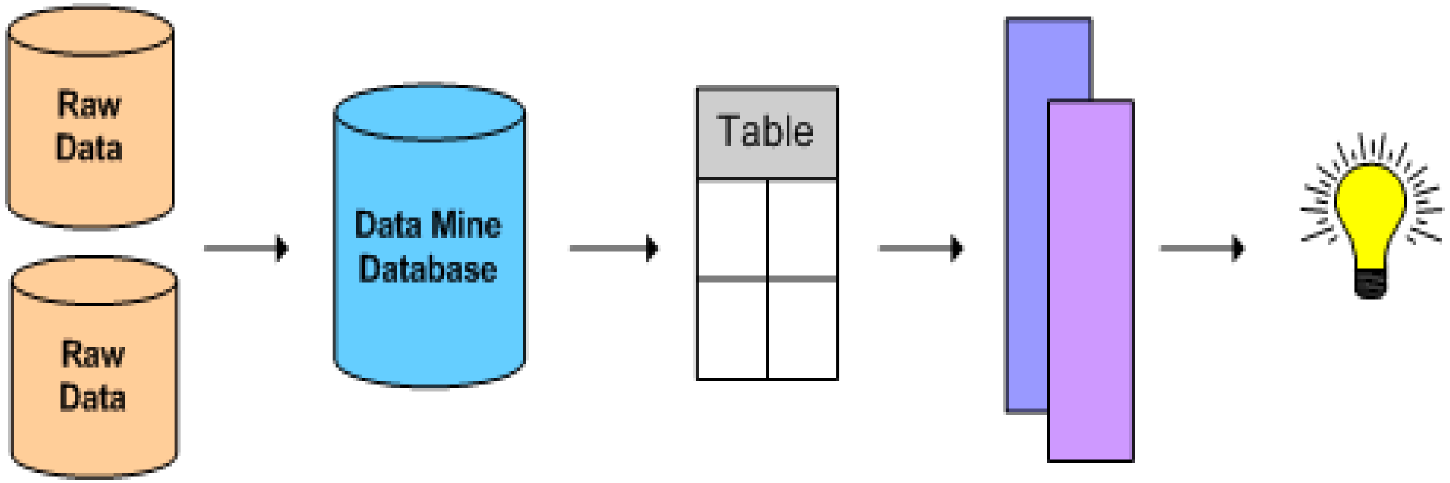
Data Mart Vs. Data Warehouse



	Data Mart	Data Warehouse
Size	< 100 GB	100 GB +
Subject	Single Subject	Multiple Subjects
Scope	Line-of-Business	Enterprise-wide
Data Sources	Few Sources	Many Source Systems
Data Integration	One Subject Area	All Business Data
Time to Build	Minutes, Weeks, Months	Many Months to Years

Data mining

- **Data mining**, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.
- Data mining is the process of discovering hidden, valuable knowledge by analyzing a large amount of data. Also, we have to store that data in different databases.
- Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line
- Data mining techniques are used in a many research areas, including mathematics, cybernetics, genetics and marketing. Web mining, a type of data mining used in customer relationship management (CRM), takes advantage of the huge amount of information gathered by a Web site to look for patterns in user behavior.



Extraction of data from repositories

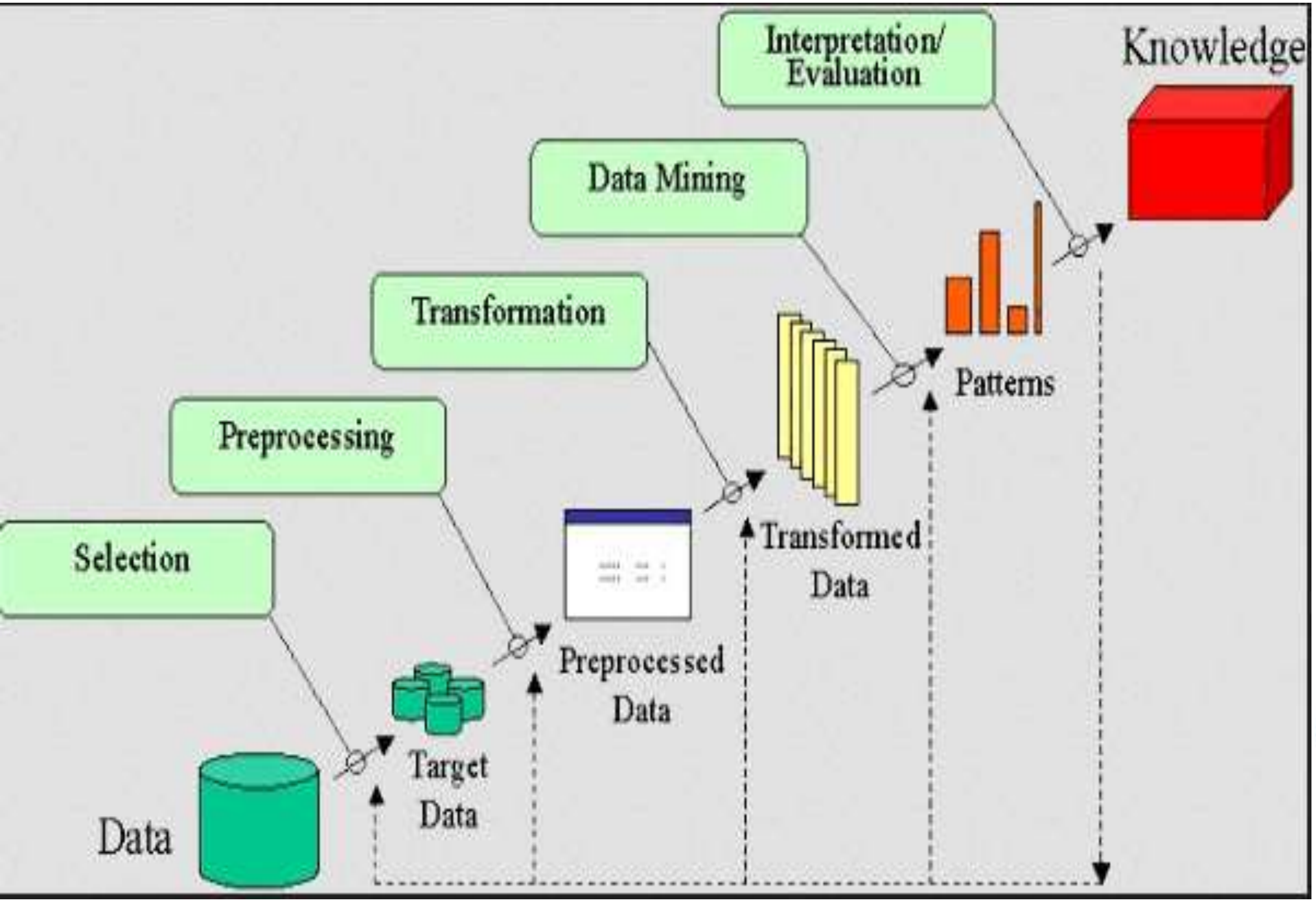
Data cleansing and loading into data mining database

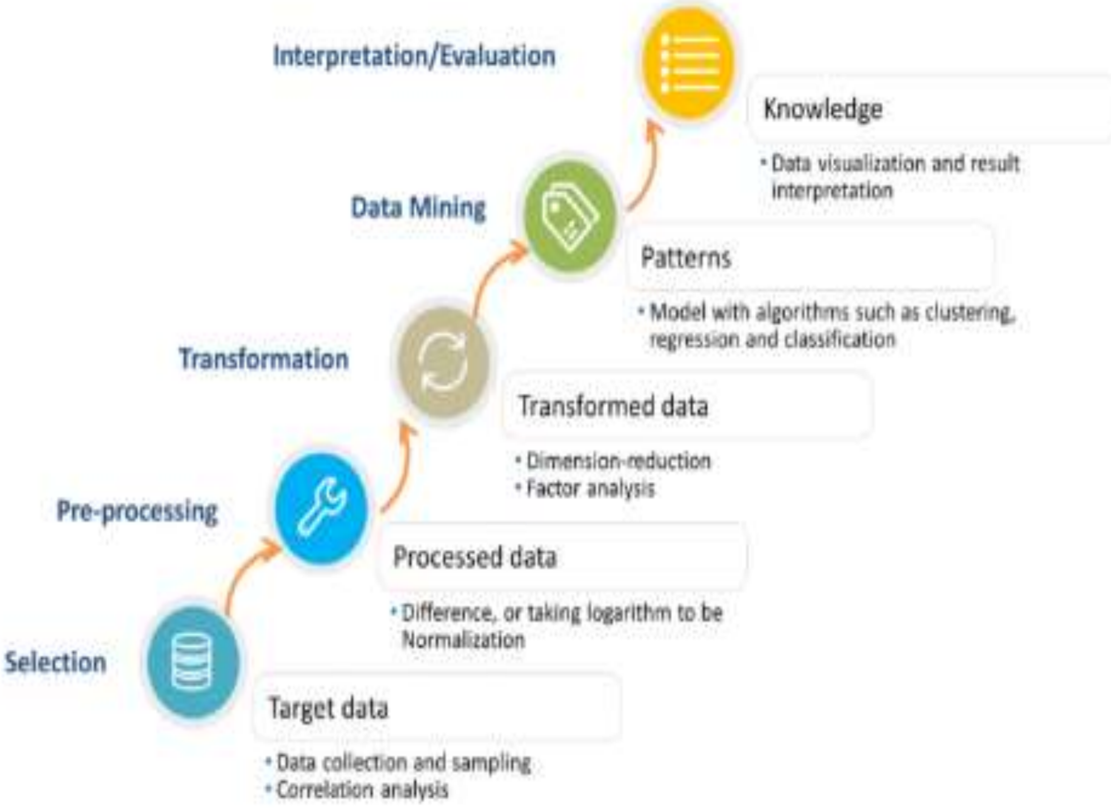
Data transformation

Pattern discovery using algorithms such as *clustering*, *regression* and *classification*

Data visualization and interpretation of results

How does Data Mining Works?/Stages in Data Mining





Interpretation/Evaluation
Knowledge
• Data visualization and result interpretation
Data Mining
Patterns
• Model with algorithms such as clustering, regression and classification
Transformation
Transformed data
• Dimension-reduction
• Factor analysis
Pre-processing
Processed data
• Difference, or taking logarithm to be Normalization
Selection
Target data
• Data collection and sampling
• Correlation analysis

Data Cleaning

- In the data mining process, data gets cleaned, as data in the real world is noisy, inconsistent, and incomplete.
- Data cleaning includes a number of techniques, such as filling in the missing values and combined compute.

Data Integration

- In this process, data in integrated from different data sources, as data is in different formats in different locations. We can store data in a database, text files, spreadsheets, documents, data cubes, and so on. Although, data integration is complex because normally data doesn't match the different sources.
- We use metadata to reduce errors in the data integration process. Another issue faced is data redundancy. In this case, the same data might be available in different tables in the same database. Data integration tries to reduce redundancy as much as possible without affecting the reliability of the data.

Data Selection

- This is the process by which data relevant to the analysis is retrieved from the database. This process requires large volumes of historical data for analysis, as usually the data repository with integrated data contains much more data than actually required. From the available data, data of interest needs to be selected and stored.

Data Transformation

- In this process, we have to transform and consolidate the data into different forms that's suitable for mining. Normally this process includes normalization, aggregation, generalization, etc.
- **For example**, a data set available as "-5, 37, 100, 89, 78" can be transformed as "-0.05, 0.37, 1.00, 0.89, 0.78". Here, data becomes more suitable for data mining. After data integration, the available data is ready for data mining.

Data Mining

- In this process, we have applied methods to extract patterns from the data. Also, this mining includes several tasks, such as classification, prediction, clustering, time series analysis, and so on.

Pattern Evaluation

- Pattern evaluation identifies the truly interesting patterns that represent knowledge based on different types of interesting measures. A pattern is considered to be interesting if it is potentially useful and easily understandable. Further, it validates some hypothesis that someone wants to confirm new data with some degree of certainty.

Knowledge Representation

- Knowledge representation is the means by which to represent data to the user in an appealing way. This can also include information that's mined from the data. To generate output, different techniques need to be applied.

Objectives of Data Mining

- **Sequence or path analysis**: Finding patterns where one event leads to another, later event.
- **Classifications**: Finding whether certain facts fall into predefined groups.
- **Clustering** : Finding groups of related facts not previously known.
- **Forecasting** :discovering patterns in data that can lead to reasonable predictions

Uses of Data Mining

- **Decision making**: Analyzing the mined data in better way will help you in making better decision irrespective of sectors.
- **CRM**: Customer relation management is another way of success in your business, by acquiring all round data of customers interest will end you with better customer relation of course to your business.
- **Research analysis**: The researchers can find any similar data from the database that might bring any change in the research.
- **Business analysis**: Business analysis is the most needed practice to know status of your business and its competitors and the domain where improvement is needed.
- **Financial Data Analysis**
- **Retail Industry**
- **Telecommunication Industry**
- **Biological Data Analysis**
- **Other Scientific Applications**
- **Intrusion Detection**
- **Fraud Detection**

Benefits or Advantages of Data Mining Techniques:

It is helpful to predict future trends:

- Most of the working nature of the data mining systems carries on all the informational factors of the elements and their structure.
- One of the common benefits that can be derived with these data mining systems is that they can be helpful while predicting future trends. And that is quite possible with the help of technology and behavioral changes adopted by the people.

It signifies customer habits:

- For example, while working in the marketing industry one can understand all the matters of customer behaviour and their habits. And that is possible with the help of data mining systems.
- As these data mining systems handle all the information acquiring techniques. It is helpful in keeping the track of customer habits and their behavior.

Helps in decision making:

- There are some people who make use of these data mining techniques to help them with some kind of decision making.
- Nowadays, all the information about anything can be determined easily with the help of technology and similarly, with the help of such technology one can make a precise decision about something unknown and unexpected.

Increase company revenue:

- As it has been explained earlier that data mining is a process wherein which it involves some sort of technology to acquire
- some information about anything possible. And this type of technology makes things easier for their profit earning ratio.
- As people can collect information about the marketed products online, which eventually reduces the cost of the product and their services.

It depends upon market-based analysis:

- Data mining process is a system where in which all the information has been gathered on the basis of market information.
- Nowadays, technology plays a crucial role in everything and that casualty can be seen in these data mining systems. Therefore, all the information collected through these data mining is basically from marketing analysis.

Quick fraud detection:

- Most parts of the data mining process is basically from information gathered with the help of marketing analysis. With the help of such marketing analysis, one can also find out those fraudulent acts and products available in the market.
- Moreover, with the help of it one can understand the importance of accurate information.

Limitations or Disadvantages of Data Mining Techniques:

It violates user privacy:

- It is a known fact that data mining collects information about people using some market-based techniques and information technology. And these data mining process involves several numbers of factors.
- But while involving those factors, data mining system violates the privacy of its user and that is why it lacks in the matters of safety and security of its users. Eventually, it creates Mis-communication between people.

Additional irrelevant information:

- The main functions of the data mining systems creates a relevant space for beneficial information.
- But the main problem with these information collection is that there is a possibility that the collection of information process can be little overwhelming for all.
- Therefore, it is very much essential to maintain a minimum level of limit for all the data mining techniques.

Misuse of information:

- As it has been explained earlier that in the data mining system the possibility of safety and security measure are really minimal. And that is why some can misuse this information to harm others in their own way.
- Therefore, the data mining system needs to change its course of working so that it can reduce the ratio of misuse of information through the mining process.

An accuracy of data:

- Most of the time while collecting information about certain elements one used to seek help from their clients, but nowadays everything has changed. And now the process of information collection made things easy with the mining technology and their methods.
- One of the most possible limitations of this data mining system is that it can provide accuracy of data with its own limits.

END