Chapter 1

INTRODUCTION TO DATABASE SYSTEMS

1.1 INTRODUCTION

An organization must have accurate and reliable data (information) for effective decision making. Data (information) is the backbone and most critical resource of an organization that enables managers and organizations to gain a competitive edge. In this age of information explosion, where people are bombarded with data, getting the right information, in the right amount, at the right time is not an easy task. So, only those organizations will survive that successfully manage information.

A database system simplifies the tasks of managing the data and extracting useful information in a timely fashion. A database system is an integrated collection of related files, along with the details of the interpretation of the data. A Data Base Management System is a software system or program that allows access to data contained in a database. The objective of the DBMS is to provide a convenient and effective method of defining, storing, and retrieving the information stored in the database.

The database and database management systems have become essential for managing business, governments, schools, universities, banks etc.

1.2 BASIC DEFINITIONS AND CONCEPTS

In an organization, the data is the most basic resource. To run the organization efficiently, the proper organization and management of data is essential. The formal definition of the major terms used in databases and database systems is defined in this section.

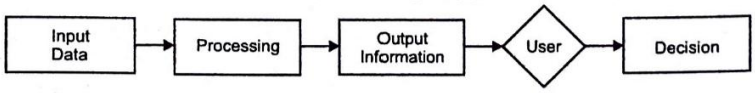
1.2.1 Data

The term data may be defined as known facts that could be recorded and stored on Computer Media. It is also defined as raw facts from which the required information is produced.

1.2.2 Information

Data and information are closely related and are often used interchangeably. Information is nothing but refined data. In other way, we can say, information is processed, organized or summarized data. According to Burch *et. al.*, "Information is data that have been put into a meaningful and useful content and communicated to a recipient who uses it to made decisions". Information consists of data, images, text, documents and voice, but always in a meaningful content. So we can say, that information is something more than mere data.

Data are processed to create information. The recipient receives the information and then makes a decision and takes an action, which may triggers other actions



In these days, there is no lack of data, but there is lack of quality information. The quality information means information that is accurate, timely and relevant, which are the three major key attributes of information.

1. **Accuracy :** It means that the information is free from errors, and it clearly and accurately reflects the meaning of data on which it is based. It also means it is free from bias and conveys an accurate picture to the recipient.
2. **Timeliness :** It means that the recipients receive the information when they need it and within the required time frame.
3. **Relevancy :** It means the usefulness of the piece of information for the corresponding persons. It is a very subjective matter. Some information that is relevant for one person might not be relevant for another and vice versa *e.g.*, the price of printer is irrelevant for a person who wants to purchase computer.

So, organization that have good information system, which produce information that is accurate, timely and relevant will survive and those that do not realize the importance of information will soon be out of business.

1.2.3 Meta Data

A meta data is the data about the data. The meta data describe objects in the database and makes easier for those objects to be accessed or manipulated. The meta data describes the database structure, sizes of data types, constraints, applications, autorisation etc., that are used as an integral tool for information resource management. There are three main types of meta data :

1. **Descriptive meta data :** It describes a resource for purpose such as discovery and identification. In a traditional library cataloging that is form of meta data, title, abstract, author and keywords are examples of meta data.
2. **Structural meta data :** It describes how compound objects are put together. The example is how pages are ordered to form chapters.
3. **Administrative meta data :** It provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of data.

1.2.4 Data Dictionary

The data dictionary contains information of the data stored in the database and is consulted by the DBMS before any manipulation operation on the database. It is an integral part of the database management systems and store meta data *i.e.*, information about the database, attribute names and definitions for each table in the database. It helps the DBA in the management of the database, user view definitions as well as their use.

Data dictionary is generated for each database and generally stores and manages the following types of information :

1. The complete information about physical database design *e.g.* storage structures, access paths and file sizes etc.
2. The information about the database users, their responsibilities and access rights of each user.
3. The complete information about the schema of the database.
4. The high level descriptions of the database transactions, applications and the information about the relationships of users to the transactions.
5. The information about the relationship between the data items referenced by the database transactions. This information is helpful in determining which transactions are affected when some data definitions are modified.

The data dictionaries are of two types : Active data dictionary and passive data dictionary.

1. **Active Data Dictionary :** It is managed automatically by the database management system (DBMS) and are always consistent with the current structure and definition of the database. Most of the RDBMS's maintain active data dictionaries.
2. **Passive Data Dictionary :** It is used only for documentation purposes and the data about fields, files and people are maintained into the dictionary for cross references. It is generally managed by the users of the system and is modified whenever the structure of the database is changed. The passive dictionary may not be consistent with the structure of the database, since modifications are performed manually by the user. It is possible that passive dictionaries may contain information about organisational data that is not computerized as these are maintained by the users.

1.2.5 Database

A database is a collection of interrelated data stored together with controlled redundancy to serve one or more applications in an optimal way. The data are stored in such a way that they are independent of the programs used by the people for accessing the data. The approach used in adding the new data, modifying and retrieving the existing data from the database is common and controlled one.

It is also defined as a collection of logically related data stored together that is designed to meet information requirements of an organization. We can also define it as an electronic filling system.

The example of a database is a telephone directory that contains names, addresses and telephone numbers of the people stored in the computer storage.

Databases are organized by fields, records and files. These are described briefly as follows :

1.2.5.1 Fields

It is the smallest unit of the data that has meaning to its users and is also called data item or data element. Name, Address and Telephone number are examples of fields. These are represented in the database by a value.

1.2.5.2 Records

A record is a collection of logically related fields and each field is possessing a fixed number of bytes and is of fixed data type. Alternatively, we can say a record is one complete set of fields and each field have some value. The complete information about a particular phone number in the database represents a record. Records are of two types **fixed length records** and **variable length records.**

1.2.5.3 Files

A file is a collection of related records. Generally, all the records in a file are of same size and record type but it is not always true. The records in a file may be of fixed length or variable length depending upon the size of the records contained in a file. The telephone directory containing records about the different telephone holders is an example of file. More detail is available in chapter 3.

1.2.6 Components of a Database

A Database consists of four components as shown in Figure 1.1.

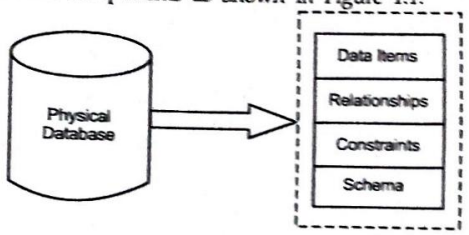


FIGURE 1.1. Components of Database.

1. **Data item :** It is defined as a distinct piece of information and is explained in the previous section.
2. **Relationships :** It represents a correspondence between various data elements.
3. **Constraints :** These are the predicates that define correct database states.
4. **Schema :** It describes the organization of data and relationships within the database. The schema consists of definitions of the various types of record in the database, the data-items they contain and the sets into which they are grouped. The storage structure of the database is described by the *storage schema*. The *conceptual schema* defines the stored data structure. The *external schema* defines a view of the database for particular users.

1.2.7 Database Management System (DBMS)

DBMS is a program or group of programs that work in conjunction with the operating system to create, process, store, retrieve, control and manage the data. It acts as an interface between the application program and the data stored in the database.

Alternatively, it can be defined as a computerized record-keeping system that stores information and allows the users to add, delete, modify, retrieve and update that information.

The DBMS performs the following five primary functions :

1. **Define, create and organise a database :** The DBMS establishes the logical relationships among different data elements in a database and also defines schemas and subschemas using the DDL.
2. **Input data :** It performs the function of entering the data into the database through an input device (like data screen, or voice activated system) with the help of the user.
3. **Process data :** It performs the function of manipulation and processing of the data stored in the database using the DML.
4. **Maintain data integrity and security :** It allows limited access of the database to authorised users to maintain data integrity and security.
5. **Query database :** It provides information to the decision makers that they need to make important decisions. This information is provided by querying the database using SQL.

1.2.8 Components of DBMS

A DBMS has three main components. These are Data Definition Language (DDL), Data Manipulation Language and Query Facilities (DML/SQL) and software for controlled access of Database as shown in Figure 1.2 and are defined as follows :

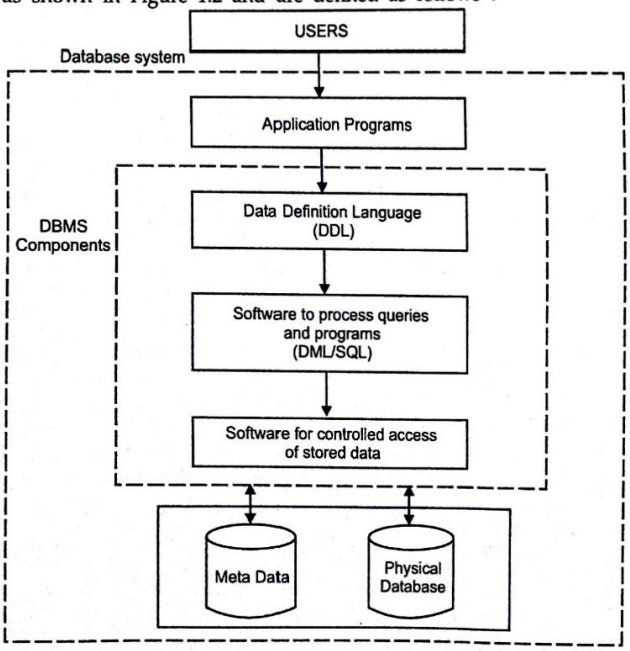


FIGURE 1.2. Components of DBMS.

1.2.8.1 Data Definition Language (DDL)

It allows the users to define the database, specify the data types, data structures and the constraints on the data to be stored in the database. More about DDL in section 1.5.

1.2.8.2 Data Manipulation Language (DML) and Query Language

DML allows users to insert, update, delete and retrieve data from the database. SQL provides general query facility. More about DML and SQL in section 1.5.

1.2.8.3 Software for Controlled Access of Database

This software provides the facility of controlled access of the database by the users, concurrency control to allow shared access of the database and a recovery control system to restore the database in case of ardware or software failure.

NOTE *The DBMS software together with the database is called a Database System.*

1.3 TRADITIONAL FILE SYSTEM VERSUS DATABASE SYSTEMS

Conventionally, the data were stored and processed using traditional file processing systems. In these traditional file systems, each file is independent of other file, and data in different files can be integrated only by writing individual program for each application. The data and the application programs that uses the data are so arranged that any change to the data requires modifying all the programs that uses the data. This is because each file is hard-coded with specific information like data type, data size etc. Some time it is even not possible to identify all the programs using that data and is identified on a trial-and-error basis.

A file processing system of an organization is shown in Figure 1.3. All functional areas in the organization creates, processes and disseminates its own files. The files such as inventory and payroll generate separate files and do not communicate with each other.

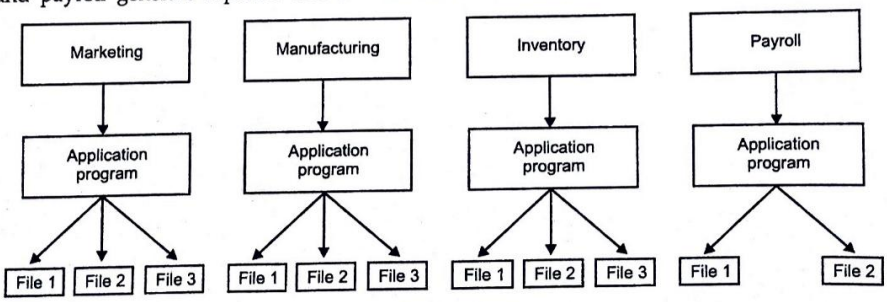


FIGURE 1.3. Traditional file system.

No doubt such an organization was simple to operate and had better local control but the data of the organization is dispersed throughout the functional sub-systems. These days, databases are preferred because of many disadvantages of traditional file systems.

1.3.1 Disadvantages of Traditional File System

A traditional file system has the following disadvantages:

1. **Data Redundancy :** Since each application has its own data file, the same data may have to be recorded and stored in many files. For example, personal file and payroll file, both contain data on employee name, designation etc. The result is unnecessary duplicate or redundant data items. This redundancy requires additional or higher storage space, costs extra time and money, and requires additional efforts to keep all files upto-date.
2. **Data Inconsistency :** Data redundancy leads to data inconsistency especially when data is to be updated. Data inconsistency occurs due to the same data items that appear in more than one file do not get updated simultaneously in each and every file. For example, an employee is promoted from Clerk to Superintendent and the same is immediately updated in the payroll file may not necessarily be updated in provident fund file. This results in two different designations of an employee at the same time. Over the period of time, such discrepencis degrade the quality of information contain in the data file that affects the accuracy of reports.
3. **Lack of Data Integration :** Since independent data file exists, users face difficulty in getting information on any ad hoc query that requires accessing the data stored in many files. In such a case complicated programs have to be developed to retrieve data from every file or the users have to manually collect the required information.
4. **Program Dependence :** The reports produced by the file processing system are program dependent, which means if any change in the format or structure of data and records in the file is to be made, the programs have to modified correspondingly. Also, a new program will have to be developed to produce a new report.
5. **Data Dependence :** The Applications/programs in file processing system are data dependent *i.e.*, the file organization, its physical location and retrieval from the storage media are dictated by the requirements of the particular application. For example, in payroll application, the file may be organised on employee records sorted on their last name, which implies that accessing of any employee's record has to be through the last name only.
6. **Limited Data Sharing :** There is limited data sharing possibilities with the traditional file system. Each application has its own private files and users have little choice to share the data outside their own applications. Complex programs required to be written to obtain data from several incompatible files.
7. **Poor Data Control :** There was no centralised control at the data element level, hence a traditional file system is decentralised in nature. It could be possible that the data field may have multiple names defined by the different departments of an organization and depending on the file it was in. This situation leads to different meaning of a data field in different context or same meaning for different fields. This causes poor data control.
8. **Problem of Security :** It is very difficult to enforce security checks and access rights in a traditional file system, since application programs are added in an adhoc manner.

9. **Data Manipulation Capability is Inadequate :** The data manipulation capability is very limited in traditional file systems since they do not provide strong relationships between data in different files.
10. **Needs Excessive Programming :** An excessive programming effort was needed to develop a new application program due to very high interdependence between program and data in a file system. Each new application requires that the developers start from the scratch by designing new file formats and descriptions and then write the file access logic for each new file.

1.3.2 Database Systems or Database System Environment

The DBMS software together with the Database is called a database system. In other words, it can be defined as an organization of components that define and regulate the collection, storage, management and use of data in a database. Furthermore, it is a system whose overall purpose is to record and maintain information. A database system consists of four major components as shown in Figure 1.4.

1. Data 2. Hardware 3. Software 4. Users

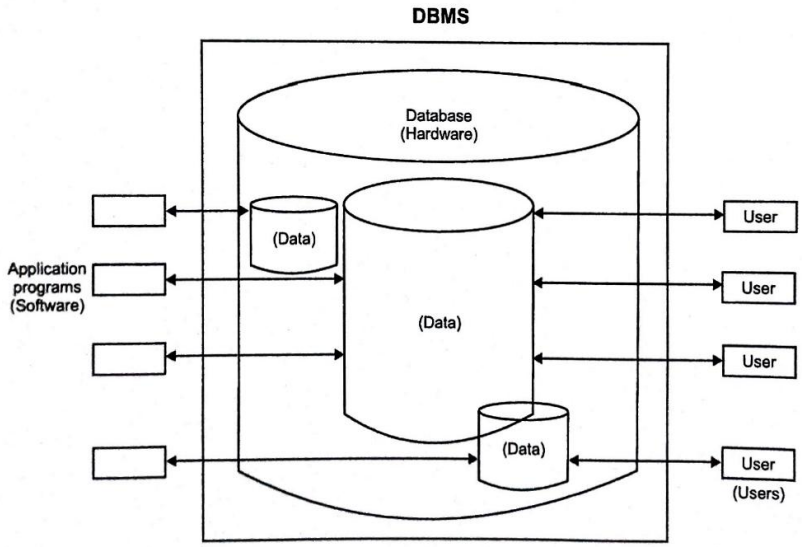


FIGURE 1.4. Database system.

1. Data : The whole data in the system is stored in a single database. This data in the database are both shared and integrated. Sharing of data means individual pieces of data in the database is shared among different users and every user can access the same piece of data but may be for different purposes. Integration of data means the database can be function of several distinct files with redundancy controlled among the files.

2. Hardware : The hardware consists of the secondary storage devices like disks, drums and so on, where the database resides together with other devices. There is two types of hardware. The first one, *i.e.*, processor and main memory that supports in running the DBMS. The second one is the secondary storage devices, *i.e.*, hard disk, magnetic disk etc., that are used to hold the stored data.

3. Software : A layer or interface of software exists between the physical database and the users. This layer is called the DBMS. All requests from the users to access the database are handled by the DBMS. Thus, the DBMS shields the database users from hardware details. Furthermore, the DBMS provides the other facilities like accessing and updating the data in the files and adding and deleting files itself.

4. Users : The users are the people interacting with the database system in any way. There are four types of users interacting with the database systems. These are Application Programmers, online users, end users or naive users and finally the Database Administrator (DBA). More about users in section 1.4.

1.3.3 Advantages of Database Systems (DBMS's)

The database systems provide the following advantages over the traditional file system:

1. **Controlled redundancy :** In a traditional file system, each application program has its own data, which causes duplication of common data items in more than one file. This duplication/redundancy requires multiple updations for a single transaction and wastes a lot of storage space. We cannot eliminate all redundancy due to technical reasons. But in a database, this duplication can be carefully controlled, that means the database system is aware of the redundancy and it assumes the responsibility for propagating updates.
2. **Data consistency :** The problem of updating multiple files in traditional file system leads to inaccurate data as different files may contain different information of the same data item at a given point of time. This causes incorrect or contradictory information to its users. In database systems, this problem of inconsistent data is automatically solved by controlling the redundancy.
3. **Program data independence :** The traditional file systems are generally data dependent, which implies that the data organization and access strategies are dictated by the needs of the specific application and the application programs are developed accordingly. However, the database systems provide an independence between the file system and application program, that allows for changes at one level of the data without affecting others. This property of database systems allow to change data without changing the application programs that process the data.
4. **Sharing of data :** In database systems, the data is centrally controlled and can be shared by all authorized users. The sharing of data means not only the existing applications programs can also share the data in the database but new application programs can be developed to operate on the existing data. Furthermore, the requirements of the new application programs may be satisfied without creating any new file.
5. **Enforcement of standards :** In database systems, data being stored at one central place, standards can easily be enforced by the DBA. This ensures standardised data formats

to facilitate data transfers between systems. Applicable standards might include any or all of the following—departmental, installation, organizational, industry, corporate, national or international.

6. **Improved data integrity** : Data integrity means that the data contained in the database is both accurate and consistent. The centralized control property allow adequate checks can be incorporated to provide data integrity. One integrity check that should be incorporated in the database is to ensure that if there is a reference to certain object, that object must exist.
7. **Improved security** : Database security means protecting the data contained in the database from unauthorised users. The DBA ensures that proper access procedures are followed, including proper authentical schemes for access to the DBMS and additional checks before permitting access to sensitive data. The level of security could be different for various types of data and operations.
8. **Data access is efficient** : The database system utilizes different sophisticated techniques to access the stored data very efficiently.
9. **Conflicting requirements can be balanced** : The DBA resolves the conflicting requirements of various users and applications by knowing the overall requirements of the organization. The DBA can structure the system to provide an overall service that is best for the organization.
10. **Improved backup and recovery facility** : Through its backup and recovery subsystem, the database system provides the facilities for recovering from hardware or software failures. The recovery subsystem of the database system ensures that the database is restored to the state it was in before the program started executing, in case of system crash.
11. **Minimal program maintenance** : In a traditional file system, the application programs with the description of data and the logic for accessing the data are built individually. Thus, changes to the data formats or access methods results in the need to modify the application programs. Therefore, high maintenance effort are required. These are reduced to minimal in database systems due to independence of data and application programs.
12. **Data quality is high** : The quality of data in database systems are very high as compared to traditional file systems. This is possible due to the presence of tools and processes in the database system.
13. **Good data accessibility and responsiveness** : The database systems provide query languages or report writers that allow the users to ask ad hoc queries to obtain the needed information immediately, without the requirement to write application programs (as in case of file system), that access the information from the database. This is possible due to integration in database systems.
14. **Concurrency control** : The database systems are designed to manage simultaneous (concurrent) access of the database by many users. They also prevents any loss of information or loss of integrity due to these concurrent accesses.
15. **Economical to scale** : In database systems, the operational data of an organization is stored in a central database. The application programs that work on this data can be

built with very less cost as compared to traditional file system. This reduces overall costs of operation and management of the database that leads to an economical scaling.

16. **Increased programmer productivity** : The database system provides many standard functions that the programmer would generally have to write in file system. The availability of these functions allow the programmers to concentrate on the specific functionality required by the users without worrying about the implementation details. This increases the overall productivity of the programmer and also reduces the development time and cost.

1.3.4 Disadvantages of Database Systems

In contrast to many advantages of the database systems, there are some disadvantages as well. The disadvantages of a database system are as follows :

1. **Complexity increases** : The data structure may become more complex because of the centralised database supporting many applications in an organization. This may lead to difficulties in its management and may require professionals for management.
2. **Requirement of more disk space** : The wide functionality and more complexity increase the size of DBMS. Thus, it requires much more space to store and run than the traditional file system.
3. **Additional cost of hardware** : The cost of database system's installation is much more. It depends on environment and functionality, size of the hardware and maintenance costs of hardware.
4. **Cost of conversion** : The cost of conversion from old file-system to new database system is very high. In some cases the cost of conversion is so high that the cost of DBMS and extra hardware becomes insignificant. It also includes the cost of training manpower and hiring the specialized manpower to convert and run the system.
5. **Need of additional and specialized manpower** : Any organization having database systems, need to be hire and train its manpower on regular basis to design and implement databases and to provide database administration services.
6. **Need for backup and recovery** : For a database system to be accurate and available all times, a procedure is required to be developed and used for providing backup copies *to all its users when damage occurs.*
7. **Organizational conflict** : A centralised and shared database system requires a consensus on data definitions and ownership as well as responsibilities for accurate data maintenance.
8. **More installational and management cost** : The big and complete database systems are more costly. They require trained manpower to operate the system and has additional annual maintenance and support costs.

1.4 DBMS USERS

The users of a database system can be classified into various categories depending upon their interaction and degree of expertise of the DBMS.

1.4.1 End Users or Naive Users

The end users or naive users use the database system through a menu-oriented application program, where the type and range of response is always displayed on the screen. The user need not be aware of the presence of the database system and is instructed through each step. A user of an ATM falls in this category.

1.4.2 Online Users

These type of users communicate with the database directly through an online terminal or indirectly through an application program and user interface. They know about the existence of the database system and may have some knowledge about the limited interaction they are permitted.

1.4.3 Application Programmers

These are the professional programmers or software developers who develop the application programs or user interfaces for the naive and online users. These programmers must have the knowledge of programming languages such as Assembly, C, C++, Java, or SQL, etc., since the application programs are written in these languages.

1.4.4 Database Administrator

Database Administrator (DBA) is a person who have complete control over database of any enterprise. DBA is responsible for overall performance of database. He is free to take decisions for database and provides technical support. He is concerned with the Back-End of any project. Some of the main responsibilities of DBA are as follows :

1. **Deciding the conceptual schema or contents of database :** DBA decides the data fields, tables, queries, data types, attributes, relations, entities or you can say that he is responsible for overall logical design of database.
2. **Deciding the internal schema of structure of physical storage :** DBA decides how the data is actually stored at physical storage, how data is represented at physical storage.
3. **Deciding users :** DBA gives permission to users to use database. Without having proper permission, no one can access data from database.
4. **Deciding user view :** DBA decides different views for different users.
5. **Granting of authorities :** DBA decides which user can use which portion of database. DBA gives authorities or rights to data access. User can use only that data on which access right is granted to him.
6. **Deciding constraints :** DBA decides various constraints over database for maintaining consistency and validity in database.
7. **Security :** Security is the major concern in database. DBA takes various steps to make data more secure against various disasters and unauthorized access of data.
8. **Monitoring the performance :** DBA is responsible for overall performance of database. DBA regularly monitors the database to maintain its performance and try to improve it.

9. **Backup :** DBA takes regular backup of database, so that it can be used during system failure. Backup is also used for checking data for consistency.
10. **Removal of dump and maintain free space :** DBA is responsible for removing unnecessary data from storage and maintain enough free space for daily operations. He can also increase storage capacity when necessary.
11. **Checks :** DBA also decides various security and validation checks over database to ensure consistency.
12. **Liaisioning with users :** Another task of the DBA is to liaisioning with users and ensure the availability of the data they require and write the necessary external schemas.

1.5 DATABASE OR DBMS LANGUAGES

The DBMS provides different languages and interfaces for each category of users to express database queries and updations. When the design of the database is complete and the DBMS is chosen to implement it, the first thing to be done is to specify the conceptual and internal schemas for the database and the corresponding mappings. The following five languages are available to specify different schemas.

1. Data Definition Language (DDL)
2. Storage Definition Language (SDL)
3. View Definition Language (VDL)
4. Data Manipulation Language (DML)
5. Fourth-Generation Language (4-GL)

1.5.1 Data Definition Language (DDL)

It is used to specify a database conceptual schema using set of definitions. It supports the definition or declaration of database objects. Many techniques are available for writing DDL. One widely used technique is writing DDL into a text file. More about DDL in chapter 7.

1.5.2 Storage Definition Language (SDL)

It is used to specify the internal schema in the database. The storage structure and access methods used by the database system is specified by the specified set of SDL statements. The implementation details of the database schemas are implemented by the specified SDL statements and are usually hidden from the users.

1.5.3 View Definition Language (VDL)

It is used to specify user's views and their mappings to the conceptual schema. But generally, DDL is used to specify both conceptual and external schemas in many DBMS's. There are two views of data the **logical view**—that is perceived by the programmer and **physical view**—data stored on storage devices.

1.5.4 Data Manipulation Language (DML)

It provides a set of operations to support the basic data manipulation operations on the data held in the database. It is used to query, update or retrieve data stored in a database. The part of DML that provide data retrieval is called query language.