Topic One: Introduction to Database Systems

At the end of this topic, you should be able to understand:

- The difference between data and information
- What a database is, the various types of databases, and why they are valuable assets for
- decision making
- The importance of database design
- How modern databases evolved from file systems
- About flaws in file system data management
- The main components of the database system

- The database is now such an integral part of our day-to-day life that often we are not aware
- we are using one.
- **Data** are raw facts. The word *raw* indicates that the facts have not yet been processed to reveal their meaning.
- **Information** is the result of processing raw data to reveal its meaning. Data processing can be as simple **as organizing data to reveal patterns** or as complex as **making forecasts** or **drawing inferences using statistical modeling**. To reveal meaning, information requires *context.*
- In this "**information age**," production of accurate, relevant, and timely information is the key to **good decision making**. In turn, good decision making is the key to **business survival** in a global market.
- Data are the foundation of information, which is the bedrock of **knowledge**—that is, the body of information and facts about a specific subject.
- **Knowledge** implies familiarity, awareness, and understanding of information as it applies to an environment.
- A key characteristic of knowledge is that "new" knowledge can be derived from "old" knowledge.

Therefore:

- Data constitute the building blocks of information.
- Information is produced by processing data.
- Information is used to reveal the meaning of data.
- Accurate, relevant, and timely information is the key to good decision making.
- Good decision making is the key to organizational survival in a global environment.

- Timely and useful information requires accurate data. Such data must be properly generated and stored in a format that is easy to access and process. And, like any basic resource, the data environment must be managed carefully.
- **Data management** is a discipline that focuses on the proper generation, storage, and retrieval of data.
- Given the crucial role that data play, it should not surprise you that data management is a core activity for any business, government agency, service organization, or charity.

So what is Database?

- *Database* to be a collection of related data and the *Database Management System* (DBMS) to be the software that manages and controls access to the database.
- A *database application* is simply a program that interacts with the database at some point in its execution.
- *A database system* to be a collection of application programs that interact with the database along with the DBMS and database itself.

Practical Application of Databases

- Some of the areas where databases are employed include

Sector	Use of Databases
Banking	**For customer information, account activities, payment, deposits, loans among others**
Airlines	**For reservations and schedule information**
Universities	**For student information, course registrations, colleges and grades**
Telecommunication	**It helps to keep call records, monthly bills, maintaining balances**
Finance	**For storing information about stock, sales, and purchase of financial instruments like stocks and bonds**
Sales	**Use for storing customer, product & Sales information**
Manufacturing	**It is used for the management of supply chain and for tracking production of items, inventories status in the warehouse**
HR Management	**For information about employees, salaries, payroll, deductions, generation of paychecks.**

Traditional File-Based Systems

- Although the file-based approach is largely obsolete, there are good reasons for studying it.
  - Understanding the problems inherent in file-based systems may prevent us from repeating these problems in database systems.
  - If you wish to convert a file-based system to a database system, understanding how the file system works will be extremely useful, if not essential.

File-Based Approach

- **File-based approach** is a collection of application programs that perform services for the end-users such as the production of reports. Each program defines and manages its own data.
- It is a decentralized approach was taken, where each department, with the assistance of **Data Processing** (DP) staff, stored and controlled its own data.
- Since each department manages its information quite clearly that there is a significant amount of duplication of data in these departments, and this is generally true of file-based systems.

Some terminologies employed in the **file-based** approach include:

- A file is simply a collection of **records**, which contains **logically related data**.

- Each **record** contains a logically connected set of one or more **fields**,
- Where each field represents some characteristic of the real-world object that is being modeled.

TABLE 1.2 **Basic File Terminology**

TERM	DEFINITION
Data	"Raw" facts, such as a telephone number, a birth date, a customer name, and a year-to-date (YTD) sales value. Data have little meaning unless they have been organized in some logical manner.
Field	A character or group of characters (alphabetic or numeric) that has a specific meaning. A field is used to define and store data.
Record	A logically connected set of one or more fields that describes a person, place, or thing. For example, the fields that constitute a record for a customer might consist of the customer's name, address, phone number, date of birth, credit limit, and unpaid balance.
File	A collection of related records. For example, a file might contain data about the students currently enrolled at Gigantic University.

Limitations of the File-Based Approach

❖ **Separation and isolation of data:** When data is isolated in separate files, it is more difficult to access data that should be available.

❖ **Duplication of data**

- Owing to the decentralized approach taken by each department, the file-based approach encouraged, if not necessitated, the uncontrolled duplication of data.
- Uncontrolled duplication of data is undesirable for several reasons, including:
  - Duplication is wasteful. It costs time and money to enter the data more than once.
  - It takes up additional storage space, again with associated costs. Often, the duplication of data can be avoided by sharing data files.
  - Perhaps more importantly, duplication can lead to loss of data integrity; in other words, the data is no longer consistent.

❖ **Structural and Data dependence**

- The physical structure and storage of the data files and records are defined in the application code.
- This means that changes to an existing structure are difficult to make.
- A file system exhibits **structural dependence**, which means that access to a file is dependent on its structure.
- The file system application programs are affected by change in the file structure.
- **Data independence** exists when it is possible to make changes in the data storage characteristics without affecting the application program's ability to access the data.

❖ **Fixed queries/proliferation of application programs**

- The file-based systems are very dependent upon the application developer, who has to write any queries or reports that are required. In some organizations, the type of query or report that could be produced was fixed.
- There was no facility for asking unplanned (that is, spur-of-the-moment or ad hoc) queries either about the data itself or about which types of data were available.

❖ **Incompatible file formats**

- Because the structure of files is embedded in the application programs, the structures are dependent on the application programming language.
- For example, the structure of a file generated by a COBOL program may be different from the structure of a file generated by a 'C' program. The direct incompatibility of such files makes them difficult to process jointly.

❖ **Lack of Design and Data-Modeling Skills**

- A new problem that has evolved with the use of personal productivity tools (such as spreadsheet and desktop databases) is that users typically lack the knowledge of proper design and data-modeling skills.
- Data-modeling skills are also a vital part of the design process. It is important that the design that is created be properly documented.
- Design documentation is necessary to facilitate communication among the database designer, the end user, and the developer.

❖ **There was no provision for security or integrity;**

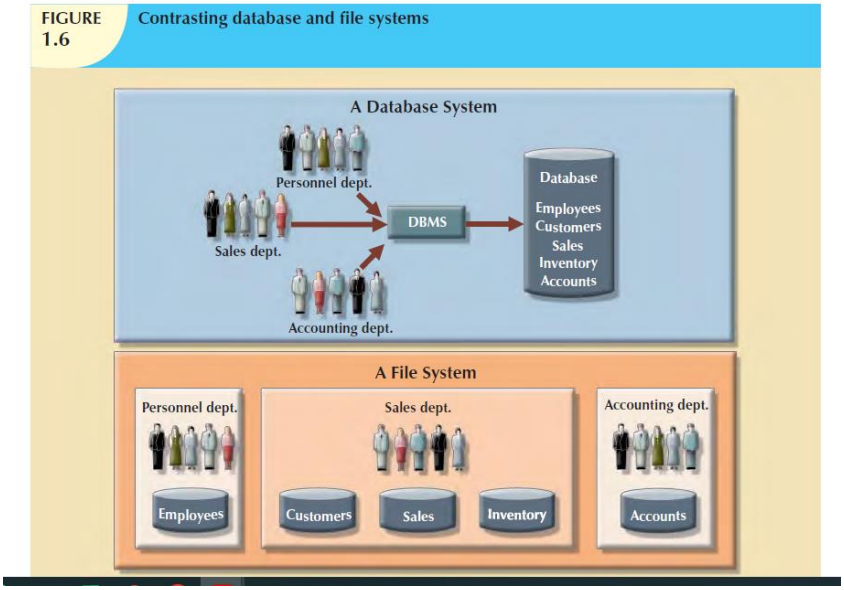
❖ Recovery, in the event of a hardware or software failure, was limited or non-existent;

❖ Access to the files was restricted to one user at a time – there was no provision for shared access by staff in the same department.

Database Approach

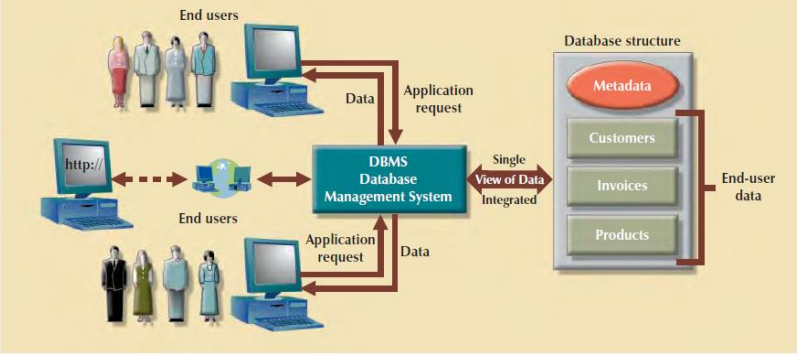
All the above limitations of the file-based approach can be attributed to two factors:

- The definition of the data is embedded in the application programs, rather than being stored separately and independently;
- There is no control over the access and manipulation of data beyond that imposed by the application programs.



- Unlike the file system, with its many **separate and unrelated files**, the database system consists of **logically** related data stored in a **single logical data repository**.
- Efficient data management typically requires the use of a computer database.
- A **database** is a shared, integrated computer structure that stores a collection of:
  - **End-user data**, that is, raw facts of interest to the end user.
  - **Metadata**, or data about data, through which the end-user data are integrated and managed.
- To become more effective, a new approach was required. What emerged were the **database** and the **Database Management System** (DBMS).

FIGURE 1.2 The DBMS manages the interaction between the end user and the database



- The **metadata** provide a description of the data characteristics and the set of relationships that links the data found within the database.
- For example, the metadata component stores information such as the name of each data element, the type of values (numeric, dates, or text) stored on each data element, whether or not the data element can be left empty, and so on.

The Database

- A **shared** collection of **logically related data**, and a **description of this data**, designed to meet the information needs of an organization.
- The database is a **single**, possibly large repository of data that can be used simultaneously by many departments and users.
- Instead of disconnected files with redundant data, all data items are integrated with a minimum amount of duplication.
- The database is no longer owned by one department but is a shared corporate resource.
- The database holds not only the organization's operational data but also a description of this data.

Types of Databases

- A DBMS can support many different types of databases.
- Databases can be classified according to the **number of users, the database location(s), the expected type and extent of use** and **degree to which the data are structured.**

a) Number of Users

- The number of users determines whether the database is classified as **single-user** or **multiuser**.
- A **single-user database** supports only one user at a time. In other words, if user A is using the database, users B and C must wait until user A is done. A single-user database that runs on a personal computer is called a **desktop database**.
- In contrast, a **multiuser database** supports multiple users at the same time. When the multiuser database supports a relatively small number of users (usually fewer than 50) or a specific department within an organization, it is called a **workgroup database**. When the database is used by the entire organization and supports many users (more than 50, usually hundreds) across many departments, the database is known as an **enterprise database**.

b) According to database location

- Location might also be used to classify the database.
- For example, a database that supports data located at a single site is called a **centralized database**.
- A database that supports data distributed across several different sites is called a **distributed database**.

c) **The expected type and extent of use**

- The most popular way of classifying databases today, however, is based on **how they will be used** and on the **time sensitivity of the information gathered from them**.
- For example, transactions such as product or service sales, payments, and supply purchases reflect critical day-to-day operations. Such transactions must be recorded accurately and immediately.
- A database that is designed primarily to support a company's day-to-day operations is classified as an **operational database** (sometimes referred to as a **transactional** or **production database**).
- In contrast, a **data warehouse** focuses primarily on storing data used to generate information required to make tactical or strategic decisions.
- Such decisions typically require extensive "data massaging" (data manipulation) to extract information to formulate pricing decisions, sales forecasts, market positioning, and so on.
- Most decision support data are based on data obtained from operational databases over time and stored in data warehouses.
- Additionally, the data warehouse can store data derived from many sources.
- To make it easier to retrieve such data, the data warehouse structure is quite different from that of an operational or transactional database.

b) According to degree to which the data are structured.

- Databases can also be classified to reflect the degree to which the data are structured.
- **Unstructured data** are data that exist in their original (raw) state, that is, in the format in which they were collected.
- Therefore, unstructured data exist in a format that does not lend itself to the processing that yields information.

- **Structured data** are the result of taking unstructured data and formatting (structuring) such data to facilitate storage, use, and the generation of information. You apply structure (format) based on the type of processing that you intend to perform on the data.
- Some data might not be ready (unstructured) for some types of processing, but they might be ready (structured) for other types of processing.
- **Semistructured data** are data that have already been processed to some extent. For example, if you look at a typical Web page, the data are presented to you in a prearranged format to convey some information.

The Database Management System (DBMS)

- A **database management system** (**DBMS**) is a collection of programs that manages the database structure and controls access to the data stored in the database.
- DBMS software stores not only the data structures, but also the relationships between those structures and the access paths to those structures—all in a central location.
- **Note**: In a sense, a database resembles a very well-organized electronic filing cabinet in which powerful software, known as a *database management system*, helps manage the cabinet's contents.
- The DBMS is the software that interacts with the users' application programs and the database. Typically, a DBMS provides the following facilities:
    - It allows users to define and modify the database
    - It provides controlled access to the database. For example, it may provide
    - A security system, which prevents unauthorized users accessing the database;
    - An integrity system, which maintains the consistency of stored data;
    - A concurrency control system, which allows shared access of the database;
    - A recovery control system, which restores the database to a previous consistent state following a hardware or software failure;
    - A user-accessible catalog, which contains descriptions of the data in the database.

Role and Advantages of the DBMS

- The DBMS serves as the intermediary between the user and the database.
- The database structure itself is stored as a collection of files, and the only way to access the data in those files is through the DBMS.
- The DBMS presents the end user (or application program) with a single, integrated view of the data in the database.
- The DBMS receives all application requests and translates them into the complex operations required to fulfill those requests.
- The DBMS hides much of the database's internal complexity from the application programs and users.

DBMS provides advantages such as:

- *Improved data sharing*. The DBMS helps create an environment in which end users have better access to more and better-managed data.
- *Improved data security*. The more users access the data, the greater the risks of data security breaches. Corporations invest considerable amounts of time, effort, and money to ensure that corporate data are used
- properly.

- *Better data integration.* Wider access to well-managed data promotes an integrated view of the organization's operations and a clearer view of the big picture.
- *Minimized data inconsistency.* **Data inconsistency** exists when different versions of the same data appear in different places.
- *Improved data access.* The DBMS makes it possible to produce quick answers to ad hoc queries. From a database perspective, a **query** is a specific request issued to the DBMS for data manipulation—for example, to read or update the data. Simply put, a query is a question, and an **ad hoc query** is a spur-of-the-moment question. The DBMS sends back an answer (called the **query result set**) to the application.
- *Improved decision making.* Better-managed data and improved data access make it possible to generate better-quality information, on which better decisions are based. The quality of the information generated depends on the quality of the underlying data. **Data quality** is a comprehensive approach to promoting the accuracy, validity, and timeliness of the data.
- *Increased end-user productivity.* The availability of data, combined with the tools that transform data into usable information, empowers end users to make quick, informed decisions that can make the difference between success and failure in the global economy.

(Database) Application Programs

- **Application program:** A computer program that interacts with the database by issuing an appropriate request (typically an SQL statement) to the DBMS.
- Users interact with the database through a number of **application programs** that are used to create and maintain the database and to generate information.
- These programs can be conventional batch applications or, more typically nowadays, they will be online applications.
- The application programs may be written in some programming language or in some higher-level fourth-generation language.

The Database System Environment

- *The term database system refers to an organization of components that define and regulate the collection, storage, management, and use of data within a database environment.*
- From a general management point of view, the database system is composed of the five major parts: **hardware, software, people, procedures, and data.**

a) Hardware

Hardware refers to all of the system's physical devices; for example, computers (PCs, workstations, servers, and supercomputers), storage devices, printers, network devices (hubs, switches, routers, fiber optics), and other devices (automated teller machines, ID readers, and so on).

The DBMS and the applications require hardware to run. The hardware can range from a single personal computer, to a single mainframe, to a network of computers. The particular hardware depends on the organization's requirements and the DBMS used. Some DBMSs run only on particular hardware or operating systems, while others run on a wide variety of hardware and

operating systems. A DBMS requires a minimum amount of main memory and disk space to run, but this minimum configuration may not necessarily give acceptable performance.

b)Software

Although the most readily identified software is the DBMS itself, to make the database system function fully, three types of software are needed: operating system software, DBMS software, and application programs and utilities.

- ***Operating system software*** manages all hardware components and makes it possible for all other software to run on the computers. Examples of operating system software include Microsoft Windows, Linux, MacOS, UNIX, and MVS.
- ***DBMS software*** manages the database within the database system. Some examples of DBMS software include Microsoft's SQL Server, Oracle Corporation's Oracle, Sun's MySQL, and IBM's DB2.
- ***Application programs and utility software*** are used to access and manipulate data in the DBMS and to manage the computer environment in which data access and manipulation take place. Application programs are most commonly used to access data found within the database to generate reports, tabulations, and other information to facilitate decision making.
- **Utilities** are the software tools used to help manage the database system's computer components. For example, all of the major DBMS vendors now provide graphical user interfaces (GUIs) to help create database structures, control database access, and monitor database operations.

c)People.

This component includes all users of the database system. On the basis of primary job functions, five types of users can be identified in a database system: **system administrators, database administrators, database designers, system analysts and programmers, and end users**. Each user type, described below, performs both unique and complementary functions.

- *System administrators* oversee the database system's general operations.
- *Database administrators*, also known as DBAs, manage the DBMS and ensure that the database is functioning properly.
- *Database designers* design the database structure. They are, in effect, the database architects. If the database design is poor, even the best application programmers and the most dedicated DBAs cannot produce a useful database environment.
- *System analysts and programmers* design and implement the application programs. They design and create the data entry screens, reports, and procedures through which end users access and manipulate the database's data.
- *Users:* are the people who use the application programs to run the organization's daily operations. For example, salesclerks, supervisors, managers, and directors are all classified as end users. High-level end users employ the information obtained from the database to make tactical and strategic business decisions.
    - **Naïve users** are typically unaware of the DBMS. They access the database through specially written application programs that attempt to make the operations as simple as possible.
    - **Sophisticated users**. At the other end of the spectrum, the sophisticated end-user is familiar with the structure of the database and the facilities offered by the DBMS.

b) Procedures

- Procedures are the instructions and rules that govern the design and use of the database system.
- Procedures are a critical, although occasionally forgotten, component of the system.
- Procedures play an important role in a company because they enforce the standards by which business is conducted within the organization and with customers.
- Procedures are also used to ensure that there is an organized way to monitor and audit both the data that enter the database and the information that is generated through the use of those data.
- The users of the system and the staff that manage the database require documented procedures on how to use or run the system.

d) ***Data.***

- The word *data* covers the collection of facts stored in the database.
- Because data are the raw material from which information is generated, the determination of what data are to be entered into the database and how those data are to be organized is a vital part of the database designer's job.

Advantages of the database approach

- **Control of data redundancy:** Database approach attempts to eliminate the redundancy by integrating the files so that multiple copies of the same data are not stored. However, the database approach does not eliminate redundancy entirely, but controls the amount of redundancy inherent in the database.
- **Data consistency:** By eliminating or controlling redundancy, we reduce the risk of inconsistencies occurring. If a data item is stored only once in the database, any update to its value has to be performed only once and the new value is available immediately to all users.
- **More information from the same amount of data:** With the integration of the operational data, it may be possible for the organization to derive additional information from the same data.
- **Sharing of data:** Typically, files are owned by the people or departments that use them. On the other hand, the database belongs to the entire organization and can be shared by all authorized users. In this way, more users share more of the data.
- **Improved data integrity:** Database integrity refers to the validity and consistency of stored data. Integrity is usually expressed in terms of **constraints**, which are consistency rules that the database is not permitted to violate.
- **Improved security**: Database security is the protection of the database from unauthorized users. Without suitable security measures, integration makes the data more vulnerable than file-based systems.
- **Enforcement of standards:** Again, integration allows the DBA to define and enforce the necessary standards.
- **Balance of conflicting requirements:** Each user or department has needs that may be in conflict with the needs of other users. Since the database is under the control of the DBA, the DBA can make decisions about the design and operational use of the database that provide the best use of resources for them organization as a whole.

- **Improved data accessibility and responsiveness:** Again, as a result of integration, data that crosses departmental boundaries is directly accessible to the end-users.
- **Improved data accessibility and responsiveness:** Again, as a result of integration, data that crosses departmental boundaries is directly accessible to the end-users.
- **Improved maintenance through data independence:** DBMS separates the data descriptions from the applications, thereby making applications immune to changes in the data descriptions.
- **Increased concurrency:** In some file-based systems, if two or more users are allowed to access the same file simultaneously, it is possible that the accesses will interfere with each other, resulting in loss of information or even loss of integrity. Many DBMSs manage concurrent database access and ensure such problems cannot occur.
- **Improved backup and recovery services.**

Disadvantages

- ***Increased costs***. Database systems require sophisticated hardware and software and highly skilled personnel. The cost of maintaining the hardware, software, and personnel required to operate and manage a database system can be substantial. Training, licensing, and regulation compliance costs are often overlooked when database systems are implemented.
- ***Management complexity.*** Database systems interface with many different technologies and have a significant impact on a company's resources and culture. The changes introduced by the adoption of a database system must be properly managed to ensure that they help advance the company's objectives. Given the fact that database systems hold crucial company data that are accessed from multiple sources, security issues must be assessed constantly.
- ***Maintaining currency.*** To maximize the efficiency of the database system, you must keep your system current. Therefore, you must perform frequent updates and apply the latest patches and security measures to all components. Because database technology advances rapidly, personnel training costs tend to be significant.
- ***Vendor dependence.*** Given the heavy investment in technology and personnel training, companies might be reluctant to change database vendors. As a consequence, vendors are less likely to offer pricing point advantages to existing customers, and those customers might be limited in their choice of database system components.
- ***Frequent upgrade/replacement cycles.*** DBMS vendors frequently upgrade their products by adding new functionality. Such new features often come bundled in new upgrade versions of the software. Some of these versions require hardware upgrades. Not only do the upgrades themselves cost money, but it also costs money to train database users and administrators to properly use and manage the new features.