# Chapter 12:
# Data Files and Databases

## Data Files & Databases

- Before databases were introduced, all data was stored in files.
- Files are still used today, but are much less common.
- A file stores only 1 type of record.
- Databases can store different types of records.

## Data Files

- Manual office systems used files to structure the information and computer systems to group common data together in files.
- The way in which data files for computerized systems are structured follows a set pattern.
  - The file is divided into records.
  - Each record is divided into fields.
  - The field is a particular data item which relates to a record.
  - Each record in a file would have the same set of fields.

## Storing Data in Files

- Traditional Method
  - Keep paper documents in 2-ring binder or file.
  - Each file contains information about a particular subject
  - E.g. students data, there could be many registration forms. Each form may include name, student number, address, etc…
- Earlier application system each had their own file

## File Organization

- Data files must be organized logically into files and be accessible and updated as needed.
- There are three major methods of file organization
  - Sequential Organization
  - Direct Organization
  - Indexed-sequential Organization

## Sequential File Organization

- This is the simplest form of file organization. Records are stored in sequential order by a particular record key.
  - For example, if the file contains employee data, then the likely data field chosen to be the record is the NRIC number or the Employee Number.
- Sequential Files can be stored on both magnetic tape and magnetic disks. However, magnetic tapes can only store sequential files, and not the others.
- Data which is stored sequentially can only be accessed sequentially. For this reason, it is called **Sequential Access Method (SAM)** file.

## Direct File Organization

- In a direct file, data records are not read (accessed) sequentially. Instead, the data management systems (the software which controls these files) is able to determine the exact location on the disk (address) where the data record is stored.
- This method of access is called **Direct Access** and the file is called **Direct Access Method (DAM) File**. This is only found in disks and not on tapes.
- How does that data management system determine the address on the disk to store and retrieve a particular data record?
  - The most common method is to use a program specially created for this purpose.
- The input will be the record key field, and the output from this program is the address. This program maps the record key into a storage area with the objective of spreading the data records as uniformly as possible in the disk.

## Indexed-Sequential File Organization

- The indexed-sequential file organization is a combination of the earlier two methods. Like the direct access method, it is found only on disks.
- The physical space on a disk where records are stored is called a track. A track is the storage area of the disk consisting of a circular band.
- In an indexed-sequential file, data records are stored sequentially by record key. More specifically, a few data records are stored on any one track. An index is created automatically by the data management system.
- The index will carry the highest record key of those records in a track. When a particular record is to be accessed, the index is looked up to find the track where the record can be found. Then the records in this track are read sequentially until the particular record desired is found.
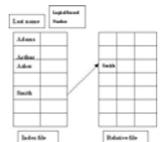
## Redundancy in Data Files

- In the traditional data files, the same data may be stored in several functional files.
- This situation duplicates data unnecessarily, creating what is referred to as **Redundancy**.
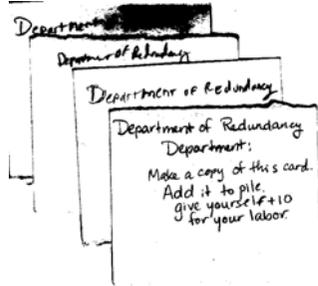
## Databases

- A **Database** is a data store for accepting, storing and providing on demand data for multiple independent users.
- It is essentially a means of organizing data in related file structure to allow quick, user specified access and manipulation of the data.
- They are controlled by systems software called **Database Management Systems** (DBMS).

## Database Management System (DBMS)

- The **Database Management System** (DBMS) is the software that manage (builds and maintains) the database.
  - This includes **Adding**, **Deleting** and **Changing** records in the database.
- The DBMS will also control access to the data. It provides the environment through which an end-user can access the data base.
  - A financier user, for example, may not be able to view personnel data, even though both finance and personnel data are kept in the same database.

## DBMS

- The DBMS makes it easier for programmers to develop application systems.
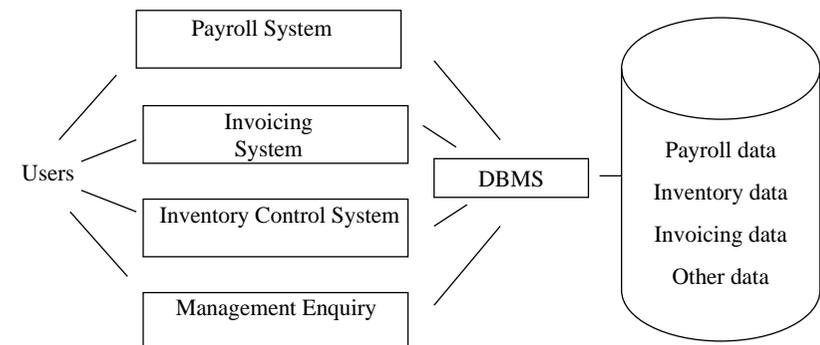
# Functions of a DBMS

- What are the functions performed by a DBMS?
  - Record Creation and Updates
  - Record Selection by Search Criteria
  - Minimize Data Redundancy
  - Security
  - User Interface

# Record Creation and Updates

- Creating /deleting/ updating records
  - The database is maintained by the user creating, deleting or amending records.
- Selecting required records
  - The Data Base Management System will also allow the user to process selected groups of records, the whole file or individual records.
  - This is accomplished by selecting the required option from a menu displayed on the video screen or by keying in the relevant command from the keyboard.

# Record Selection by Search Criteria

- Specify criteria
  - The processing can also involve the extraction of information from the selected records according to specified search criteria.
  - In a student database, if we wanted to list the students who have attended less than 80% of classes, this would be handled by the DBMS provided the required data is recorded in the database.
- Filter out unwanted data
  - A database has searching facilities ranging from limited searches on keywords to full text searching.
- SQL statements
  - A **Query** language is used to specify the search criteria.
  - This language facilitates enquiries using English – style word to define commands, so greatly assisting users to find the information they are seeking.
- Extracting from large amount of data

# Minimize Data Redundancy

- Data duplication
  - If a data element or record (e.g. an employee address) appears twice in an organization's records, this is called **Data Redundancy**, since logically, it should not be necessary to hold the data twice, hence one copy is redundant.
  - In a database, data redundancy should be eliminated, or at least, minimized and this is the task of the DBMS.

# Security

- Access rights
  - The Data Base Management System is also responsible for providing security for the database, ensuring against unauthorized access and data corruption.
  - If many end-users are given access to common data, there must be a way of controlling that access to ensure that no two users are able to change the data simultaneously. The DBMS is responsible for such control.

# User Interface

- User interface
  - Easy to use systems and authorize users only

# Data Dictionary

- A data dictionary contains details of the data in a database.
- Just like an English-language dictionary, each data item is recorded together with its definition and explanation.
- The dictionary assists in the design of a database since the meaning of various types of data are easily located.
- It can be used to reduce the level of redundancy to a necessary minimum.
- It is easier to identify synonyms (data groups having more than one name).

# Example

Date: 12/19/2003                          Project: WEB LANE CAFE                          Page: 11
Time: 10:36:28 AM

Detailed Listing -- Alphabetically
All Entries -- Data Flow Diagrams

Vendor-ID                                                                 Data Element
Vendor File::Vendor-ID
   *Description:*
      A unique identification number assigned to each vendor.
   *Alias:*
      Vendor Code
   *Values & Meanings:*
      Required element
      Cannot be blank
      May not be duplicated
   *Data element attributes*
      *Storage Type:*        Char
      *Length:*              4
      *Display Format:*      AAAA
      *Null Type:*          NotNull
   *Location:*
      File -->                Vendor File
   *Date Last Altered:* 12/19/2003          *Date Created:*   12/19/2003

## Advantages of Databases

- Reduces data redundancy
  - Reduces data duplication which occurs using conventional file structure.
- Avoids duplicating input data
  - Avoids duplicating input data to update multiple functionality independent files holding the same date.
- Offers useful query and report generation languages
  - Offers useful query and report generation languages.
- Provides fast and flexible access to information
  - Provides fast and flexible access to information.

## Disadvantages of Database

- Complex data structures require a longer design and implementation time
  - Data structures in a database are complex and therefore takes a longer time to design and implement.
- High cost compared to the conventional file processing approach
  - Costs more than conventional file systems.
- Requires more technically competent staffing
  - Requires high caliber, experienced and specialized personnel.

## Creating a Database Environment

- Departments have all along maintained their own data and have been jealous of it.
- Inter-departmental rivalry leads to reluctance to divulge data, let alone make it available in a central database.
- There must be a change of attitude in terms of who actually owns data.
- Appointing a Chief Information Officer is one way to give direction and focus, as the CIO will be responsible for the overall use of data and information in the company.
- A data administration function must be established and the CIO must plan the current and future data needs.

## Physical and Logical Records

- The form in which data which is actually stored is not the same as the form presented to the application program.
- The data structure that is actually on the computer media (e.g. disk, tape ) is called a **Physical Structure**.
- The structure that the application program uses is a **Logical Structure**.

## Data Files and Databases

- Database Management Systems (DBMS) do not replace the traditional method of storing data in data files.
- The data is still stored in sequential, direct or indexed sequential file structure even though a database management system is used to provide greater flexibility in accessing data.
- There are a variety of methods that could be used by a DBMS to logically structure the data.
- Two methods that will be discussed are
  - ◆ Multi-lists
  - ◆ Inverted Files

## Multilist

- A list links common items in a database.
- A list is like a chain and each link in that chain is a record with the desired attribute.
- By starting at the first record, the DBMS can retrieve, one by one, the records in the chain which share the desired attribute.

## Inverted Files

- In an inverted file, the addresses of each record with a desired attribute are stored in an index.
- In a fully inverted file, there will be one index for each type of data item.
- In a partially inverted file, only some of the attributes are indexed.



Figure 4.3 An Inverted File