CHAPTER – 4

File

File is nothing but an Electronic document. The contents can be ordinary Text or it can be an executable program. Each file is given a file name to identify it. The File name is in the form

File Name . Extension

Filename can consist of Alphabets or combinations of alphabets, numerals and special characters. Extension indicates the type of file.

Example : XY.Doc

Here File name is XY

Extension name is .DOC which indicates Document file.

Folder

Folder contains a group of files. Folder is otherwise called as directory. Folder may have a set of files under it. It may have other folders under it also. This files and folders can be arranged in hierarchical manner or a tree like structure.

File organization

The arrangement of records in a file is known as File Organisation. File Organisation deals with the arrangement of data items in the secondary storage devices like magnetic disk. That is, the file organisation deals with how the logical tuples (rows) of tables (relations) are organised on the physical storage medium.

For organising records efficiently in the form of a computer file, following three things are important:

(a) A logical method should be selected to organise records in a file.

(b) File structure should be so designed that it would allow quick access to needed data items.

(c) Means of adding or deleting data items or records from files must be present.

Depending on the above considerations, a file may be organised as:

(a) Sequential file

(b) Direct or random access file

(c) Indexed-sequential file

Sequential file

A sequential file is a file in which the records are stored in some order, say the student file contains records of students in the ascending order of roll number of students. It is not necessary that all the records of a sequential file should be in physical adjacent positions. On a magnetic tape, the records are written one after the other along the length of the tape. In case of magnetic disks, the records of a sequential file may not be in contiguous locations. The sequential order may be given with the help of pointers on each record.

Sequential files are preferable when they are to be stored on sequential access devices such as Magnetic tapes.

1001	1002	1003	1004	1005	1006
Rec.1	Rec.2	Rec.3	Rec.4	Rec.5	Rec.6

A sequential file on tape

1001	1002		1003	1004
Rec.1	Rec.2		Rec.3	Rec.4

1015	1017
Rec.1	Rec.2

A sequential file on disk

The main advantages of sequential file organisation are:

(a) File design is simple.

(b) Location or records requires only the record key.

(c) When the activity rate is high, simplicity of the accessing method makes processing efficient.

(d) Low-cost file media such as magnetic tapes can be used for storing data.

The main drawbacks of sequential file organisation are.

(a) Updating requires that all transaction records are sorted in the record key sequence.

(b) A new master file, physically separate and exclusive, is always created as a result of sequential updating.

(c) Addition and deletion of records is not simple.

Direct Access File.

A sequential file is not suitable for on-line enquiry. Suppose a customer at a bank wishes to know the balance amount in his savings account. If the customer file Ls organised sequentially, the record of this customer has to be obtained by searching sequentially from the beginning. There is no way of picking out the particular record wit'hout traversing the file from the beginning and this mpy take a long time. Hence, in such situations, random access or direct access file organisation provides a means of accessing records speedily.

In random access or direct access method of file organisation, each record has its own address on the file. With the help of this physical address, the record can be directly accessed for reading or writing. The records need not be in any sequence within the file and also need not be in adjacent locations on the storage medium. Such a file cannot be created on a magnetic tape medium. Random (or direct) files

are created only on magnetic disks. Since every record can be independently accessed, every transaction can be manipulated individually.

Random access file organisation is best suited for on-line processing systems where current information is the one that is always required.

The advantages of Direct Access file organisation are:

(a) Immediate access to records is possible.

(b) Up-to-date information will always be available on the file.

(c) Several files can be simultaneously updated.

(d) Addition and deletion or records is not very complex.

(e) No new master file is created for updating a random access file.

The disadvantages of Direct Access file organisation are:

(a) Less efficient in the use of storage space.

(b) Uses a relatively expensive medium.

(c) Not well suited for batch processing.

(d) Data security is less due to direct access facility.

Indexed Sequential File

Some files may be required to support both batch processing and on-line processing. For example, an inventory or stock file may be updated periodically by batch processing and at the same time may have to provide current information about stock availability on-line. They can be thus, organised as indexed sequential files. Indexed Sequential file combines the advantages of sequential and direct access file organisations.

An indexed sequential file is basically a sequential file organised serially on key fields. In addition, an index is maintained which speedsúp the access of isolated records. Just as you may se indexes to locate information in book, similarly an index

is provided for the file. The file is divided into a number of blocks and the highest key in each block is indexed.

Key	Starting address of the block
125	12
860	19
1420	24
1600	42
1829	49
2225	159
2890	165
3200	807

Indexed sequential files are also known as Indexed Sequential Access Method (ISAM) files.

Within each block, the record is searched sequentially. This method is much faster than searching the entire file sequentially. It is also possible to have more than one level of indexing to make the search process faster.

The main advantage of indexed sequential file organisation is that it is suitable for both sequential and on-line or direct access processing.

The main disadvantages of this organisation are:

(a) Less efficient in the use of storage space.

(b) Additions and deletions of records are more complex as they effect both the index and the record number in the file.

Deciding on a File Organisation

The major factors to be considered while deciding which file organisation should be chosen are the following:

(a) Percentage of actual records processed in a day. If large number of records are accessed at a time, direct access file organisation should be used. If very few records are accessed, sequential file organisation will be more suitable and cheaper.

(b) Files that are frequently updated must be stored on a direct access storage device, such as disk.

(c) Selection of file organisation also depend on the mode of processing i.e. where the system requires an online processing or batch processing.

DATA PROCESSING

Data Processing means, processing the input data to produce some meaningful and purposeful information. Computer has the capability of processing high-volume of data in less time with higher accuracy. Hence the data processing performed by computer is sometimes called Electronic Data Processing (EDP). Data processing involves 5 distinct steps.

- Data capturing
- Data validation
- Processing / Execution
- Data storage
- Data Retrieval/Out generation

Data capturing encompasses the activities of inputing data to the computer. Before giving input to the system the required data are to be first identified and put in the defined format called source data layout. The aim of this layout is to have faster data entry. To reduce the volume of data and also have better organisation and easy

access to data those can suitably be coded. After data are ready they can be entered to the computer through keyboard. This is sometimes called data capture through intelligent terminal. The other form of data capture is through scanners or optical devices. In this type of data capturing data are not entered rather data are captured from the source document or paper as it is Photographs, fingerprints, signatures, objective multiple type answers in answer papers etc. are captured through this method. Another form of data capturing is through some interfacing devices from where data can be transferred directly to the computer. Example of this is Electronic cash Registers used in shops and cash counters.

After data capturing, the data is validated. Data validation involves checking of input data to fit to requirements or specifications. For example price of a book can be numeric only. If by mistake Alphabetic data are entered then it is checked and error is shown to revalidate the data. This prevents unwanted and unspecified data to enter into the system and causing errors.

The valid input data are stored in file or database and processed as per the instructions. The instructions are put in programs or software. This software or program, when executed does the processing of input data and produces the output. Again outputs are stored in files in memory.

After processing of data, the outputs are produced. The outputs may be formed in different ways depending on the requirements and specifications. The same set of data can be printed in tabular form or in form of graphs. There are variety of ways for presenting data. The printed output is sometimes called hardcopy. There can be provision of answering to queries of user where the answer is displayed on the monitor screen itself. So it depends on the requirements of user.

An important step of Data Processing is maintaining Database. Database is nothing but collection of data which is controlled centrally with many provisions of data security. This is where, normally data are stored for reference. The input data and output data are stored in database. Even after processing of data and producing of data is over, database is maintained properly with safety for future needs and reference. The important tool of data processing is file. File is nothing but an electronic document where data can be stored. Depending on the type of file structure and organisation the data access speed varies.

10 Effective Ways to Data Capture

Depending upon the procedure of collecting information, the data capture process can be divided into two segments:

1. **Manual Data Capture:**

   In manual data capture process, the data is entered manually by an operator using input devices like keyboard, touch screens, mouse etc. for keying in data in the form of figures or text into particular software such as Excel or any other data or word processing program. This method of data collection is labor intensive, time consuming and so businesses find it efficient to migrate to automated methods of data capture. However, the manual method is not totally extinct and still finds application in many business processes. Briefly, the methods of manual data capture include using:

- Mouse
- Graphics tablet
- Keyboard
- Touch-screen – e.g. PDA
- Tracker ball

2. Automated Data Capture:

Automated data capture involves the use of computerized technology to capture data. This method has a high initial cost on account of the initial investment required as for instance, the purchase of technology but as the project proceeds, is found to lower the operating costs significantly on account of low manpower requirement. Further, with the majority of data today existing in electronic forms, the cost of using such automated technology has also reduced. Hence, there has been proliferation of techniques and technology of automated methods of data capturing, each suitable for a particular type data or source of data. Automated data capture includes the use of different technologies such as OCR, ICR, OMR and others, which are individually described here.

3. Optical Character Recognition (OCR):

OCR technology is used to convert different types of machine-printed documents including image files, PDF files or scanned paper documents, into searchable and editable data.

4. Intelligent Character Recognition (ICR):

ICR technology helps to recognize and capture handwritten printed characters from image files. As handwritten text caries significantly, so ICR is less accurate and complicated as compared to other technologies. However, the technology evolves continuously by itself and as the number of samples processed increases, the accuracy increases. This self-learning process differentiates it from others and gives it the name of 'intelligent'.

5. **Optical Mark Reading (OMR):**

OMR technology is used to capture human marked data from documents such as forms and surveys. The technology has the capacity to differentiate between marked and unmarked boxes and so, is used for capturing data through boxes that are checked manually on documents.

6. **Magnetic Ink character Recognition (MICR):**

It is a data capture technology capable of recognizing characters. It involves the recognition of specially formatted characters that are printed in magnetic ink, by a machine. The technology is mainly used by banking industry to speed up the processing of cheques and other documents. The added advantage of this method is that humans can also read the data as well.

7. Magnetic Stripe Cards:

Magnetic stripe cards store data using magnetic properties of certain materials. They possess stripes of iron-based magnetic materials on the card. They are used for electronically storing particular numbers related to credit cards, identity cards and they enable automated data transfer when they reswiped in magnetic readers.

8. Smart-Cards:

Smart cards are pocket-sized cards with embedded integrated circuits. They can function on contact or can be contactless. They contain more memory than magnetic cards and can be used for data related to personal identification, authentication, biometrics etc. Upon interaction with suitable reading devices they enable automated information transfer and data access.

9. **Web-Data Capture:**

Data capture from web involves the capture of data from electronic forms through internet or intranet.

10. **Voice- Recognition:**

Voice recognition is the process of converting speech into text. The text can be simple text or can be a set of commands. It finds application in dictation systems, small controlling systems and certain processes of data entry and word-processing environment.

methods of data storage

Data storage is the holding of data in an **electromagnetic form** for **access** by a **computer processor**. There are two main kinds of storage:

Primary storage is data that is held in in random access memory (RAM) and other memory devices that are built into computers.

Secondary storage is data that is stored on external storage devices such as hard disks, tapes, CD's.

The table below summarises the main methods of data storage

Method	Commentary
Hard disks	Often called a disk drive, hard drive or hard disk drive, this method of data storage stores and provides relatively quick access to large amounts of data. The information is stored on electromagnetically charged surfaces called 'platters'.
Floppy disks	A floppy disk is a type of magnetic disk memory which consists of a flexible disk with a magnetic coating. Almost all floppy disks for personal computers now have a capacity of 1.44 megabytes. Floppy disks are readily portable, and are very popular for transferring software from one PC to another. They are, however, very slow compared to hard disks and lack storage capacity. Increasingly, therefore, computer manufacturers are not including floppy disk drives in the products as a built-in storage option.
Tape storage	Tape is used as an **external storage medium**. It consists of a loop of flexible celluloid-like material that can store data in the form of electromagnetic charges. A tape drive is the device that positions, writes from, and reads to the tape. A tape cartridge is a protectively-encased tape that is portable.
Optical disks	An optical disc is a storage medium that can be written to and read using a low-powered laser beam. A laser reads these dots, and the data is converted to an electrical signal, finally converted into the original data.
CD-R	**Compact Disc-Recordable ("CD-R") discs** have become a universal data storage medium worldwide. CD-Rs are becoming increasingly popular for music recording and for file storage or transfer between personal conmputers. CDR discs are **write-once media**. This means that - once used -they cannot be erased or re-recorded upon. CD-R discs can be played back in any audio CD player or CD-ROM drive, as well as many DVD players and drives.
CD-RW	Compact Disc-Rewritable (CD-RW) disks are rewritable and can be erased and re-recorded upon over and over again. CD-RW discs can only be used on CD players, CD-ROM drives, and DVD players and drives that are CD-RW playback-compatible.
DVD	A DVD (Digital Versatile Disc or Digital Video Disc) is a high density optical disc with large capacity for storage of data, pictures and sound. The capacity capacity is 4.7 GB for single sided, singe layer DVD disc - which is approximately 7 times larger than that of a compact disc.