6

DATA PROCESSING CONCEPTS

'Data' means collection of unorganized facts, which can be organized in to useful information. For example, collection of purchase or sales orders, registration cards, etc. can form data. 'Processing' refers to a group of actions which can convert inputs into outputs. The series of operations performed to convert unorganized data into organized information is called data processing, and includes resources like people, procedures and devices to convert the data into information.

Data Hierarchy

Data and information are stored in computer files for processing and retrieval. If such files are not carefully organized and managed, the decision makers will not be able to find the data when they need it. Thus the data in the computer system is organized in a hierarchy, which is called data hierarchy chain, and the ascending order of the chain is bits, bytes, fields, records, files and databases.

A 'bit' is a value, which represents the presence or absence of an electronic signal and is represented as '0' or '1'. A combination of eight bits makes a 'byte'. The basic building block of data, called 'character' (number, alphabet etc.), is represented with the help of a byte. A meaningful group of characters or bytes is referred to as a 'field' (also called fact, data item or data element). Field is the smallest logical data entity, which is a single unit in data processing.

A group of interrelated fields is called a 'record'. A collection of records of the same type is called a 'file'. Files may be of two types, master files containing permanent data, and transaction files recording temporary data. A collection of integrated and related master files is called a 'database'. A data hierarchy can be represented as follows:

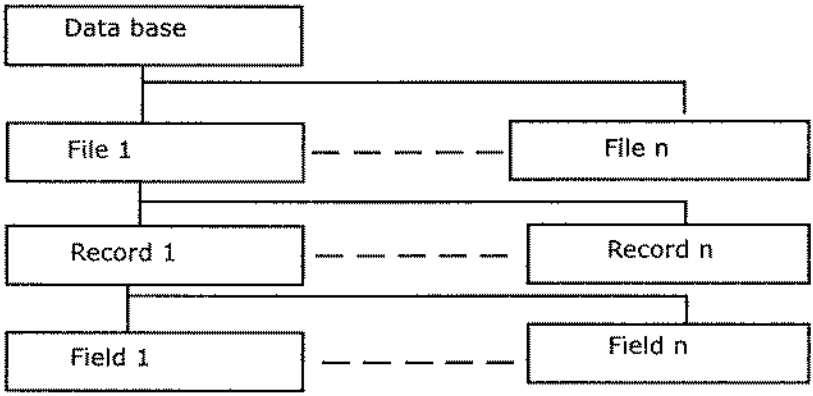

Fig. 6.1: Data Hierarchy

Method of Organizing Data in Files

There are a number of methods to organize data in files. The selection of the method depends upon factors like storage media, accessing methods, processing techniques, etc. The commonly used file organizations are:

1. Sequential file organization: In sequential file organization, records are written and stored on a secondary storage device in the same sequence or order in which they are collected. The records are arranged in order, using a unique key and are physically adjacent to one another. For instance, employee records are arranged in the order of their Date Of Joining. In this case, 'DOJ' is the unique key. In order to retrieve a record or process it, a file is sequentially processed from the beginning, in the order in which it was created, until the desired record is located. That is, the ordering of data cannot be changed when the file is processed. If any record is modified, then the entire file needs to be rearranged.

Merits of sequential files

1. Sequential file organization is very efficient and economical when a large number of file records need to be updated at regular intervals.

2. This method of file organization is inexpensive when compared to other methods.
3. Sequential files are very easy to arrange, store and understand.
4. Locating a specific record is very easy since it requires the specification of the key field.

Demerits

1. In order to retrieve a record, an entire file may need to be read.
2. In order to make the files sequential, transactions must be arranged in sequence before processing.
3. There is scope for high data redundancy.
4. Sequential files are inefficient and uneconomical, when the number of files to be arranged is very small.

In spite of its limitations, sequential file organization is well suited for batch processing applications like pay-roll preparation. The method requires very little storage space and is useful for applications where records have to be retrieved in the same order, every time the file is processed.

2. Direct file organizations/Hashed file organizations: This type of file organization is otherwise called Random access file organization, which allow data to be retrieved quickly in a random manner, regardless of the way in which the data was originally stored. The data are arranged in such a way that the computer can directly locate the key of the desired record without searching the series of records in the file. The problem is that it is very difficult to find out the location of the records. It requires a direct access storage device like drum, disk strip file, etc.

Advantages of Direct files

1. They can provide minute up-to-date information.
2. They are capable of easy access and quick retrieval of data.
3. No sequential arrangement of data is required before processing.
4. It is possible to arrange them sequentially if required.
5. This type of file organization is highly suitable for interactive online applications like airline or railway reservation systems.

6. The transactions can be processed as and when are generated.

Disadvantages

1. They require larger storage capacity and space than sequential files.
2. They require special security measures.
3. Updating of files is very difficult
4. They are highly expensive, since the direct access devices like hardware, software etc. are required.

3. Indexed sequential files: In indexed sequential files also, data is stored in a special sequence. The peculiarity is that a special index is created to show the memory address of each piece of data. The index table is used to speed up access to the records without searching the entire file. Such files provide the user with sequential access. For example, a telephone index or book index shows the location of each topic. Similarly, a file index shows the physical location of each piece of data. The method of keeping indexed files is called Indexed Sequential Acess Method (ISAM) and files of this type are referred to as ISAM files.

Merits of indexed sequential files:

1. Direct access to records is possible when the number of files is very small.
2. They allow efficient and economical use of sequential processing techniques when the number of files is very large.
3. They help to identify the location of a record without searching the entire file.

Demerits of indexed sequential files:

1. They are inefficient in using storage space.
2. They are relatively expensive, since they require direct access storage devices like hardware and software.
3. Their slow access to records problem.

The Concept of Database

In earlier days, data was stored and processed, using file-processing system. Today, database is preferred to traditional files due to the several problems associated with

the file processing system. Important among such problems are:

1. Data Independence: In file processing system, each file is independent of other files, and in order to integrate data in different files customized programmes have to be written.

2. Data and application: that use data are so tightly interwoven that any change to the data requires changing all the programmes that use the data.

3. Data redundancy: This is the situation where same data resides in several files, which causes wastage of storage capacity, difficulty in updating and maintaining files, and inconsistency of data values.

4. Data Duplication: It may compromise the integrity of data. When one file is updated, file containing the same record may not be updated. The data base overcomes many of these problems.

Unlike file processing systems, in which files are independent, a database is a collection of interrelated files. Files are dependent on the application programmes that access and process the files. The important features of database are:

1. Databases are organized and structured in different ways.
2. Duplication of data is minimized.
3. A database is an integrated and centralized data files consisting of all the data required by the organization.
4. It provides for easy and equal access to all data stored.
5. Programs and data in a database are independent.
6. It can be stored on a direct – access device.
7. Logical relationships among various records are defined in database.

Database Management System

Database Management System is a support programme that works in conjunction with the operating systems, to create, process control and manage data. It is a collection of programs required to store and retrieve data from a database.

DBMS is defined as a set of programs that act as an interface between the application programs and the data in the database.

One of the primary advantage of a DBMS is data independence, which means that the user can access a piece of data without being burdened with the question of physical location of the data. The accessing of data in a DBMS is simply based on its contents and associations with other data. Further, in a DBMS, the physical organization of data is independent of the program that uses the data. Hence , if data is physically relocated, the programs that use the data do not have to be modified, making it considerably easy to maintain the database. Moreover, data redundancy is considerably reduced, since there is no need to store the same data in different locations. Accessing and processing of data is comparatively easy in DBMS than in traditional files.

The primary disadvantage of a database is that it requires a considerable outlay of resources. Mainframe hardware is expensive and the cost of software development for DBMS is high. Moreover, integration or co-ordination of the activities of different departments is not an easy task.

Components of DBMS

The principal components of a DBMS are a data description module, which analyses the data requirements of application programs and transfers the control, and data manipulation module which retrieves the needed data elements from the database. These two components can again be into the following three categories:

1. Data Definition Language (DDL): The contents of a database is created, using the DDL. It defines the relationships between different data elements and serves as an interface for the application programs that use the data.

2. Data Manipulation Language (DML): Data is processed, and updated, using a language called DML, whose commands process, update, and retrieve data. It includes special user-friendly structured query languages. They deal exclusively with data integrity, data manipulation, data access, data retrieval , data query, and data security. Most DBMS products use some version of SQL, whose primary purpose is to allow users to query. There are four basic operations in an SQL, such as select update, insert, and delete. Users can ask

two kinds of questions in SQL. They are: static questions which are routine, standardized questions that once defined can repeatedly be used for and are appropriate for generating weekly, monthly and quarterly reports, and dynamic questions that are adhoc and specific to the decision-maker.

3. Data Dictionary (DD): It means data about data and can be defined as a component of a DBMS, which describes the data and its characteristics like location, size and type of data. Every organization, whether small or large, needs a tool to describe, identify, locate, control and manage each piece of data in the organization and to ensure consistency and standardization in the use of data through out the organization. A data dictionary identifies the origin, use, ownership, and methods of accessing and searching data. The DBMS uses DD to address all questions pertaining to data, such as definitions, storage locations, use and access privileges.

Database Models

Database model is the way of organizing data and its interrelationships. There are three such models. They are developed on the basis of three types of basic relationships, like one-to-one, one-to-many, and Many-to-many. These relation ships can be represented as:

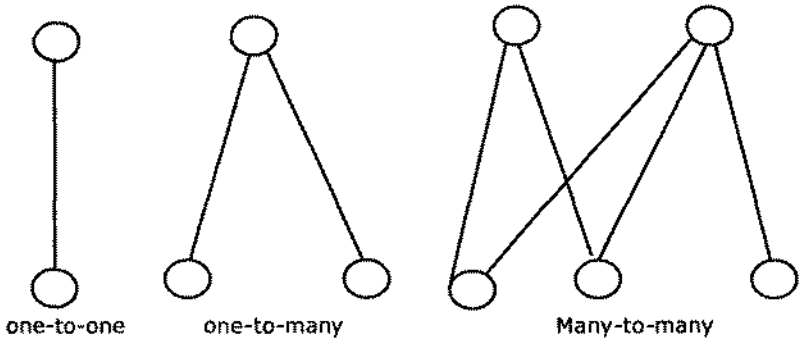

Fig. 6.2: Database relationships

The relationship between husband and wife is an example of one-one-one relationship, where as the relationship between a mother and her children is an example of one-to-many relationship, many-to-many relationship exists among teachers and students. The three logical models of database are:

1. Hierarchical data model: In a hierarchical data model, logical relationships among various data elements are represented as a hierarchy which is similar to an organization chart. A given data element can be accessed only by going through the proper hierarchy. Each box in the hierarchical model is a record, sometimes referred to as a 'node'. The topmost node is called the 'root node'. The relationships between different nodes is referred to as a parent-child relationship. In a hierarchical structure, each node, except the root node, has exactly one parent. In other words, this has a one-to-many relationship. It can be represented as follows:

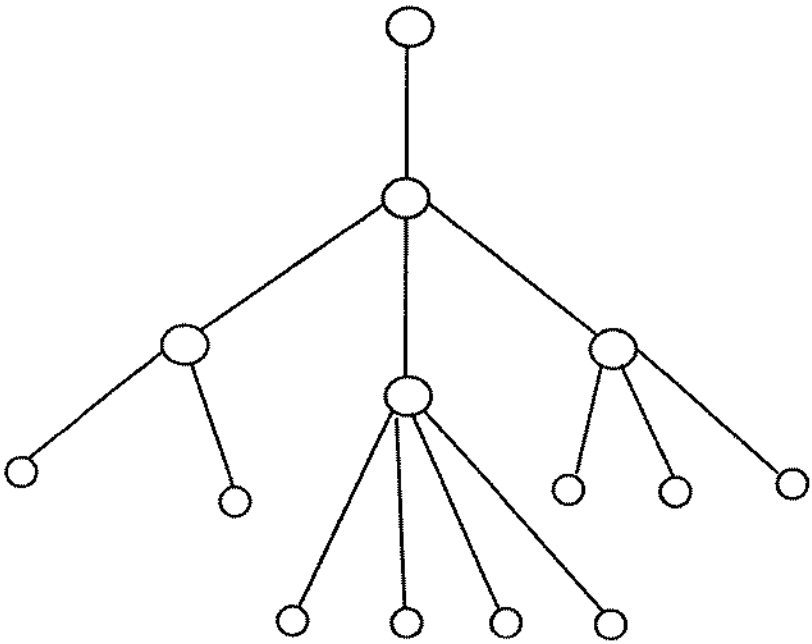

Fig. 6.3: Hierachical data model

The hierarchical model is ideally suited to problems in which data elements have a natural hierarchical structure. However, this model has some disadvantages. The data values stored at a lower level cannot be accessed without accessing data values above them. So, retrieval from a database, especially from a large one is time-consuming. Moreover, this model has a rigid structure in which the relationships between different elements must be clearly identified before development begins, and any change to the model require

major programming effort. So, the model is not always be flexible to accommodate the dynamic data needs of an organization.

2. Network model: In a network model, each record in a database can have multiple parents. The relationship among data elements is many-to-many. For example, each student in a class can attends the classes of different teachers. Similarly, each teacher can teach different students. The main difference between hierarchical model and network model is that, in the latter, a child can have a number of parents where as in the former; a child can have only one parent. It can be represented as:

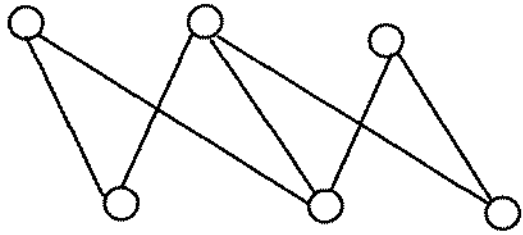

Fig. 6.4: Network model

The network model has many advantages. Its structure promotes flexibility and data accessibility, since data elements at a lower level can be accessed without accessing the data elements above them. The model is efficient, easy to understand, and can be applied to many real world problems that require routine transactions. The disadvantages of this model are that it is complex to design and develop, and it has to be developed and fine-tuned frequently, so that relationships among different pieces of data are true representatives of the real world. Like the hierarchical model, the network model requires that the relationships among all data elements be defined before development begins, and changes often demand a major programming effort. Further, for large databases, operation and maintenance of network model are time-consuming and expensive.

3. Relational model: Here, data is represented using two-dimensional tables, which are made up of columns and rows. For example, the faculty table can be represented as follows:

Faculty Table

Teacher	*Department*	*Qualification*
P	Computer	MCA
Q	Management	MBA
R	Chemistry	MSc.
S	Commerce	M.com.

The biggest advantage of the relational model over others is that it can relate data in a table to data in any other table as long as two tables or files share at least one common attribute. Its most appealing quality is its simplicity. Users can easily relate the data in tables and thus find the data structure in a relational model easy to understand and implement. Secondly, uses are not burdened with issues like storage structure and access strategy, because the relational database automatically address these issues. Thirdly, the relational model is flexible and can integrate data and information from multiple files. The fourth advantage is that relational models support ad hoc queries. Finally, new data can be easily added and old ones can be deleted or updated without significant design changes to the database.

The relational model suffers from certain drawbacks also. It is very slow when compared with other data models, since it has to access data from different files which is time consuming. Secondly there is the problem of data redundancy. However, the relational model continues to gain popularity.

Conclusion

Databases are vital parts of information systems, which provide a number of advantages over the traditional file processing system. Data models can be either logical or physical. In logical models, a user-oriented way is used to describe and understand data, whereas in the physical model, we describe physical storage of data. The traditional file organization systems are of three types like sequential, direct, and indexed sequential files. In order to overcome the shortcomings of the conventional file processing system, the concept of database and DBMS is developed. At present, there are three models of databases which offer the users various user depending upon his time and requirement. They are hierarchical, network, and relational models.

Exercise

Short Answer Questions

1. Explain data hierarchy.
2. What are the different methods of organizing data in files?
3. Compare and contrast the hashed file organization and sequential file organization.
4. Briefly explain the merits and limitations of sequential files.
5. What are the advantages of direct files over sequential files?
6. What are the additional benefits of indexed sequential files when compared with sequential files?
7. Briefly describe the concept of database.
8. What are the important problems of file processing system?
9. What do you mean by data redundancy?
10. What is a DBMS?
11. What are the specific advantages of DBMS over traditional file processing system.
12. What are the different components of a DBMS?
13. Write short note on Data Manipulation Language and Data Definition Language.
14. What is a Data Dictionary?
15. Briefly explain various database models.
16. What is a hierarchical data model ?
17. Explain the network model of database.
18. Discuss the relational model of database with suitable example.
19. Distinguish between data and information.
20. Explain the advantages and disadvantages of indexed sequential files.

Essay Questions

1. Describe the traditional file processing system with its advantages and limitations.
2. What is a database and database management system and how it overcomes the limitations of file processing system?
3. Explain DBMS, its features and various components.
4. Describe database models representing one-to-one, one to many and many-to-many relationships, with examples.
5. Explain the concept of data processing.