Chapter 14: Data Processing

PRELUDE

- ✓ Data processing is any computer process that converts data into information or knowledge.
- ✓ The processing is usually assumed to be automated and running on a computer.
- ✓ Because data are most useful when well-presented and actually informative, data-processing systems are often referred to as information systems to emphasize their practicality.
- ✓ Data-processing systems typically manipulate raw data into information, and likewise information systems typically take raw data as input to produce information as output.

DATA VERSUS INFORMATION

- ✓ Data are plain facts. When data are processed, organized, structured or presented in a given context so as to make them useful, they are called Information.
- ✓ Data themselves are fairly useless. But when these data are interpreted and processed to determine its true meaning, they become useful and can be called Information.
- ✓ Data are assembled in a set order or format for further processing to produce a meaningful result that may support an existing theory or hypothesis or even help to formulate a new one.

Computer data is processed in two fundamental ways:

i. File Processing
ii. Database processing

FILE PROCESSING

- ✓ method of processing data in separate files
- ✓ consists of creating, storing, and/or retrieving the contents of a file to or from a recognizable medium, reading from and writing to files.
- ✓ has two types
  - o Sequential File Processing
  - o Direct Access File Processing

Sequential File Processing

- ✓ stores and accesses files in sequence using a tape or a disk storage.
- ✓ Records are sorted before they are processed
- ✓ Used where data can be processed in batches and where a substantial portion of a master file is changed with the processing of each batch.

Direct Access File Processing

- ✓ Data is accessed directly from any location on the disk
- ✓ Indexes are used to track the location of the file randomly.

PROBLEMS WITH FILE PROCESSING

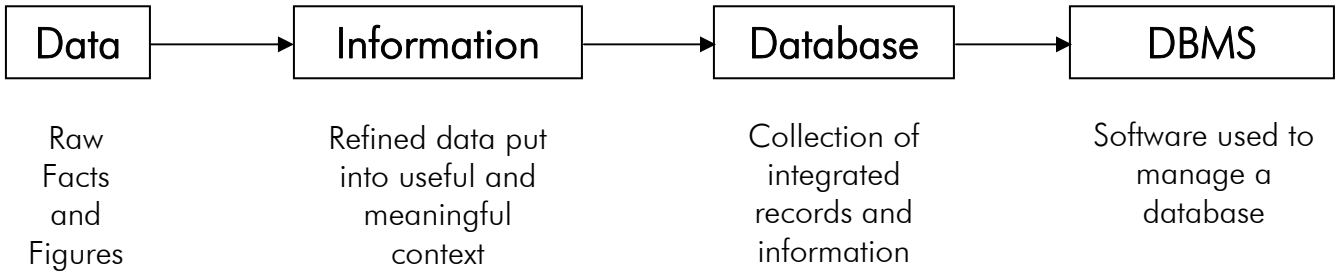
- ✓ Duplication of data leads to wastage of disk space.
- ✓ If the same data is stored in two different places, an update is required in both the places. Otherwise, there would arise a data integrity problem.
- ✓ Difficult to relate records in one file to records in another. [Either a duplicate copy is to be created or a special program (query) is to be prepared to extract the records.]

DATABASE PROCESSING

- ✓ A database is a collection of integrated records which may contain multiple files specifically called tables.
- ✓ Records in the tables are processed by their relationship to one another
- ✓ The DBMS stores and processes the data so that data can be accessed via their relationships to other data.

PRELUDE

- ✓ Information is the refined data, data that have been put into a meaningful and useful context and communicated to a recipient who uses it to make decisions.
- ✓ It involves the communication and reception of intelligence or knowledge.
- ✓ It appraises, notifies, stimulates, reduces uncertainty, reveals additional alternatives or helps in eliminating irrelevant or poor ones and influence individuals and stimulates them to action.
- ✓ Information consists of data, images, text, documents and voice often intertwined but always organized in a meaningful context.



QUALITY OF INFORMATION

1. Accuracy

- ✓ Information should be error free and should clearly reflect the meaning of data on which it is based.
- ✓ It should depict a clear picture to the recipient and may require a graphical presentation rather than tabular structure.
- ✓ Information should be free from biasness without manipulation and distortion.

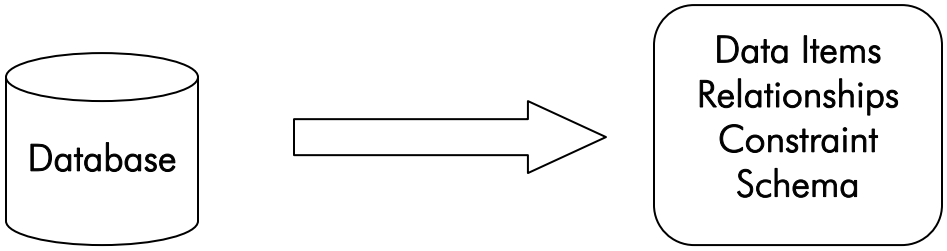
2. Timeliness

- ✓ Timeliness means delivering the information to the recipients with the needed time frame.

3. Relevancy

- ✓ Relevancy means the use of any specific information for a specific person.
- ✓ Information relevant for one person may not be relevant for another.

ELEMENTS OF A DATABASE



TERMS USED IN DATABASE SYSTEM

1. Table
   A table is a part of database with rows and columns where rows are records and columns are fields.
2. File
   A file is a collection of data on disk accessed by a unique name.
3. Record
   A record is a group of related fields of information treated as a unit, an entity. Each row in a table is a record.
4. Fields
   The fields of a record contain the data items, attribute of the entity. The field has its length and data type.
5. Primary Key
   A primary key is a unique field that identifies the records in a table. It is the key that prevents redundant data as no duplicate value possible.
6. Entity
   An entity is a "thing" or "object" in the real world that is distinguishable from all other objects. e.g. each person in an enterprise is an entity.
7. Attribute (Fields)
   The individual properties of the entity, about which data is recorded are its attribute. e.g. the attributes of Report (Entity) will include "Roll No.", "SName", "Class", "Subject1", "Subject2", "Total"
8. Relationship
   A relationship is an association among several entities.

DRAWBACKS OF OLDER DATABASE SYSTEMS

- Encoded Data (data hard-coded in the application)
- Interdependence between programs and data files
- Data repetition or redundancy
- Data inconsistency (Irregularity)
- Ad hoc (Unplanned/Informal) representation of relationships
- Ad hoc data management techniques
- Lack of data security mechanisms
- Non uniform back-up and recovery methods

DATABASE AND ITS IMPORTANCE

A database is a structured collection of records or data that is stored in a computer system. In order for a database to be truly functional, it must not only store large amounts of records well, but be accessed easily. In addition, new information and changes should also be fairly easy to input. In order to have a highly efficient database system, you need to incorporate a program that manages the queries and information stored on the system called the Database Management System. Besides these features, all databases that are created should be built with high data integrity and the ability to recover data if hardware fails.

ADVANTAGES OF DATABASE

Data redundancy can be reduced
Data redundancy refers to the repetition of data in multiple places. This wastes the storage spaces. By using a computerized database system, redundancy can be reduced to remarkable amount. Redundancy can be removed using a primary key.

Inconsistency can be avoided
Data becomes inconsistent when there is a redundancy error. When there is same data on two sites and changes are made at one site only without propagating to the next site, data remain inconsistent as the entries regarding the same data do not agree.

Data Sharing
Data in a database can be shared with any existing application.

Standards can be enforced
With the central control of the database, the database administrator can enforce standards.

Security restrictions can be applied
The DBA can ensure the security of data by the use of a proper channel to access data. Authorization checks can be carried out while accessing sensitive data.

Integrity (Accuracy) can be maintained
Integrity of a database refers to it accuracy and precision. A centralized control of the data helps in permitting the administrator to define integrity constraints to the data in database.

Conflicting requirements can be balanced
Knowing the overall requirement as opposed to the individual requirements, the database can be structured to provide an overall service that is best for the organization.

CHARACTERISTICS OF DATA IN A DATABASE

The data in a database should have the following features.

- ✓ **Shared:** Shared among different users and applications
- ✓ **Persistence:** Data in a database exist permanently and has thoroughness.
- ✓ **Validity/Integrity/Correctness:** Data should be correct with respect to the real world entity that they represent.
- ✓ **Security:** Data should be protected from unauthorized users and access.
- ✓ **Consistency:** Whenever more than one data element in a database represents related real-world values, the values should be consistent (regular) with respect to the relationship.
- ✓ **Non-redundant:** No two data items in a database should represent the same real-world entity.

DATABASE MANAGEMENT SYSTEM

- ✓ A DBMS is a software that provides services for accessing a database, while maintaining all the required features of the data.
- ✓ A DBMS is a set of software programs that controls the organization, storage, management, and retrieval of data in a database.
- ✓ It is a set of prewritten programs that are used to store, update and retrieve a Database.
- ✓ The DBMS accepts requests for data from the application program and instructs the operating system to transfer the appropriate data.
- ✓ When a DBMS is used, information systems can be changed much more easily as the organization's information requirements change. New categories of data can be added to the database without disruption to the existing system.

SERVICES OF DBMS

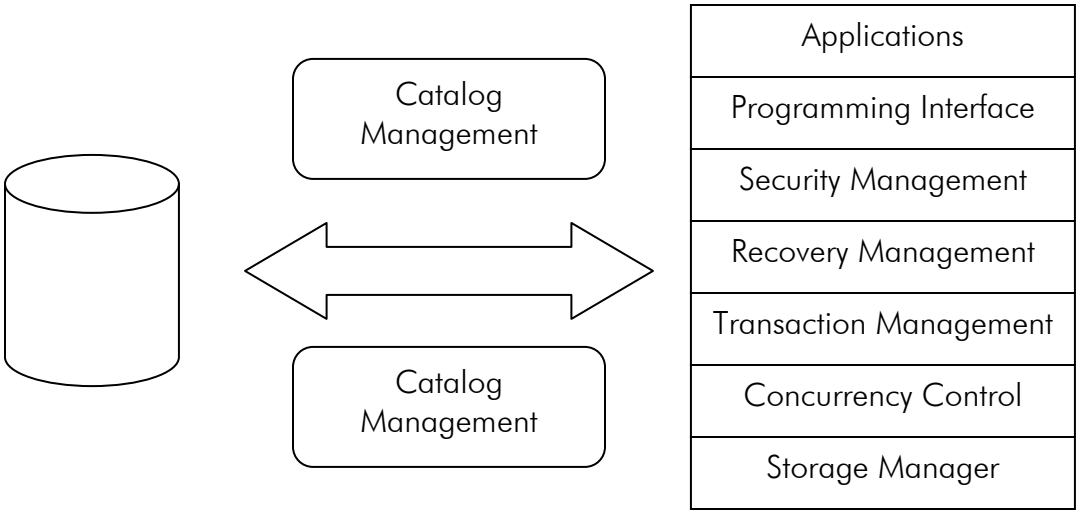


Fig: Services provided by a DBMS

Transaction Processing: A transaction is a sequence of database operations that represents a logical unit of work. It assesses a database and transforms it from one state to another. A transaction can update a record, delete one or modify a set of records. When a DBMS does a "commit", the changes made by the transaction are made permanent. If a user does not want to make the change permanent, he can rollback the transaction and the database will remain in its original state.

Concurrency Management: It is the database management activity of coordinating the actions of database manipulation processes that operates concurrently, access shared data and can potentially interfere with each other. The goal of an ideal concurrency management mechanism is to allow concurrency while maintaining the consistency of the shared data.

Recovery: The objective of recovery in a database is to ensure that the aborted of failed transactions do not create any adverse effect on the database or other transactions. It makes sure that the database is returned to a consistent state after a transaction fails or aborts. Recovery is a very much related to concurrency in the sense that, the more the concurrency, the more is the chance of an aborted transaction affecting many other transactions.

Security: Security refers to the protection of data against unauthorized access. Security mechanisms make sure that only authorized users are given access to the data in the database. The level of access is for each user and the operations that each user can perform on the data will be monitored and controlled by the DBMS depending on the access privileges of the users.

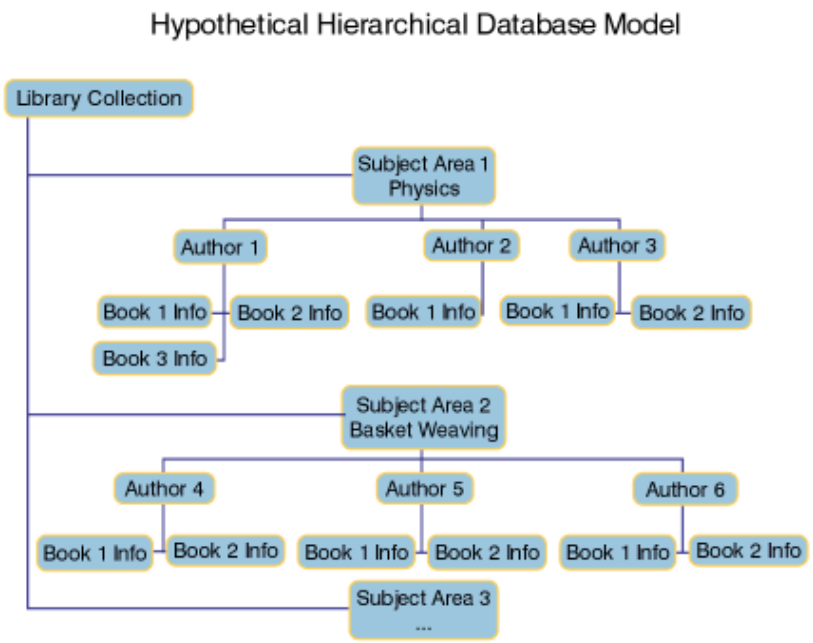
Language Interface: The DBMS provides support languages used for the definition and manipulation of the data in the database. By providing language support for data definition and manipulation the DBMS create an environment where the users can do their jobs without worrying about the physical implementation.

Data Catalog: Data catalog or Data Dictionary is a system database that contains the descriptions of data in the database (metadata). It contains information about data, relationship, constraints and the entire schema that organize these features into unified database. It also gives the information about the structure of the database.

Storage Management: The DBMS provides a mechanism for the management of permanent storage of the data.

DATABASE MODELS

Hierarchical Model

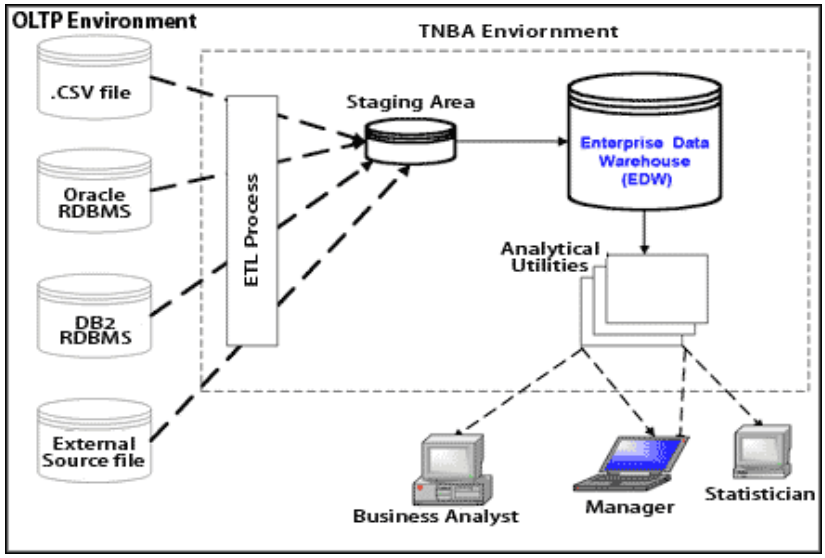


The hierarchical data model organizes data in a tree structure. There is a hierarchy of parent and child data segments. This structure implies that a record can have repeating information, generally in the child data segments. Data is a series of records, which have a set of field values attached to it. It collects all the instances of a specific record together as a record type. These record types are the equivalent of tables in the relational model, and with the individual records being the equivalent of rows. To create links between these record types, the hierarchical model uses Parent Child Relationships. These are a 1:N mapping between record types. This is done by using trees, like set theory used in the relational model, "borrowed" from maths.

For example, an organization might store information about an employee, such as name, employee number, department, salary. The organization might also store information about an employee's children, such as name and date of birth. The employee and children data forms a hierarchy, where the employee data represents the parent segment and the children data represents the child segment. If an employee has three children, then there would be three child segments associated

with one employee segment. In a hierarchical database the parent-child relationship is one to many. This restricts a child segment to having only one parent segment.

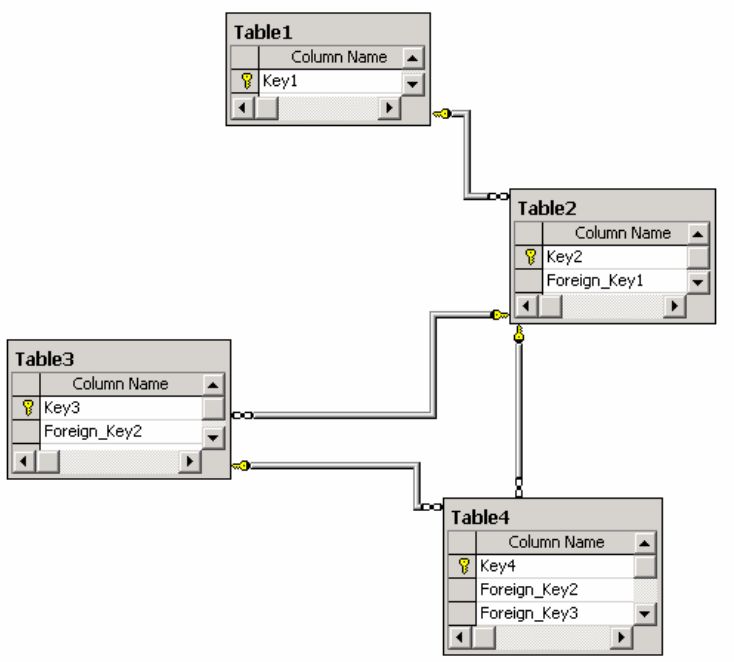
Network Model



The popularity of the network data model coincided with the popularity of the hierarchical data model. Some data were more naturally modeled with more than one parent per child. So, the network model permitted the modeling of many-to-many relationships in data. In 1971, the Conference on Data Systems Languages (CODASYL) formally defined the network model. The basic data modeling construct in the network model is the set construct. A set consists of an owner record type, a set name, and a member record type. A member record type can have that role in more than one set, hence the multiparent concept is supported. An owner record type can also be a member or owner in another set. The data model is a simple network, and link and intersection record types (called junction records by IDMS) may exist, as well as sets between them. Thus, the complete network of relationships is represented by several pairwise sets; in each set some (one) record type is owner (at the tail of the network arrow) and one or more record types are members (at the head of the relationship arrow). Usually, a set defines a 1:M relationship, although 1:1 is permitted. The CODASYL network model is based on mathematical set theory.

Relational Model

Developed by E.F. Codd, a relational database allows the definition of data structures, storage and retrieval operations and integrity constraints. In such a database the data and relations between them are organized in tables. A table is a collection of records and each record in a table contains the same fields.



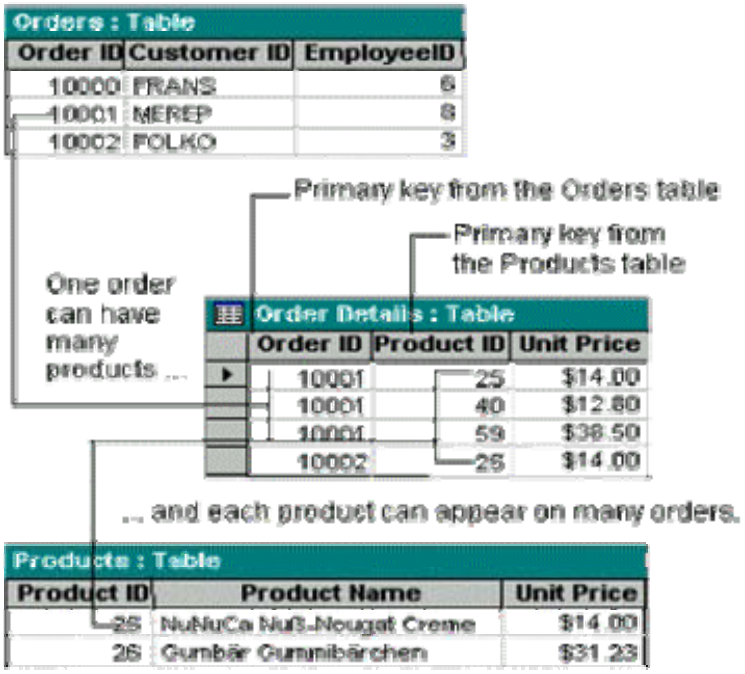
Properties of Relational Tables:

- Values Are Atomic
- Each Row is Unique
- Column Values Are of the Same Kind
- The Sequence of Columns is Insignificant
- The Sequence of Rows is Insignificant
- Each Column Has a Unique Name

Certain fields may be designated as keys, which mean that searches for specific values of that field will use indexing to speed them up. Where fields in two different tables take values from the same set, a join operation can be performed to select related records in the two tables by matching values in those fields. Often, but not always, the fields will have the same name in both tables. For example, an "orders" table might contain (customer-ID, product-code) pairs and a "products" table might contain (product-code, price) pairs so to calculate a given customer's bill you would sum the prices of all products ordered by that customer by joining on the product-code fields of the two tables. This can be extended to joining multiple tables on multiple fields. Because these relationships are only specified at retrieval time, relational databases are classed as dynamic database management system. The RELATIONAL database model is based on the Relational Algebra.

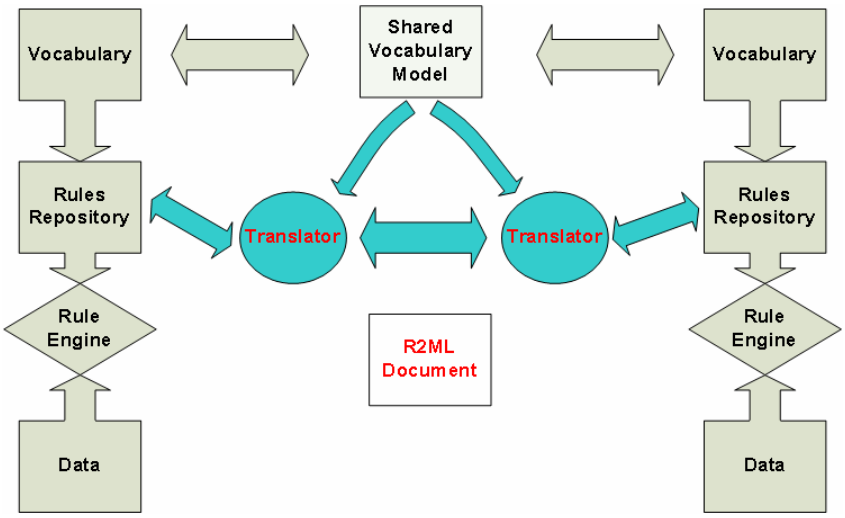
Object-Oriented Model

Object oriented DBMS add database functionality to object programming languages. They bring much more than persistent storage of programming language objects. Object oriented DBMS extends the semantics of the C++, Smalltalk and Java object programming languages to provide full-featured database programming capability, while retaining native language compatibility. A major benefit of this approach is the unification of the application and database development into a seamless data model and language environment. As a result, applications require less code, use more natural data modeling, and code bases are easier to maintain. Object developers can write complete database applications with a modest amount of additional effort.



In contrast to a relational DBMS where a complex data structure must be flattened out to fit into tables or joined together from those tables to form the in-memory structure, object DBMSs have no performance overhead to store or retrieve a web or hierarchy of interrelated objects. This one-to-one mapping of object programming language objects to database objects has two benefits over other storage approaches: it provides higher performance management of objects, and it enables better management of the complex interrelationships between objects. This makes object DBMSs better suited to support applications such as financial portfolio risk analysis systems, telecommunications service applications, World Wide Web document structures, design and manufacturing systems, and hospital patient record systems, which have complex relationships between data.

Deductive/Inference Model



A deductive model stores as little data as possible but compensates by maintaining rules that allow new data combinations to be created when needed. A deductive database system is a database system which can make deductions based on rules and facts stored in the database. Datalog is the language typically used to specify facts, rules and queries in deductive databases. Deductive databases have grown out of the desire to combine logic programming with relational databases to construct systems that support a powerful formalism and are still fast and able to deal with very large datasets. Deductive databases are more expressive than relational databases but less expressive than logic programming systems. Deductive databases have not found widespread adoptions outside academia, but some of their concepts are used in today's relational databases to support the advanced features of more recent SQL standards.

PRELUDE

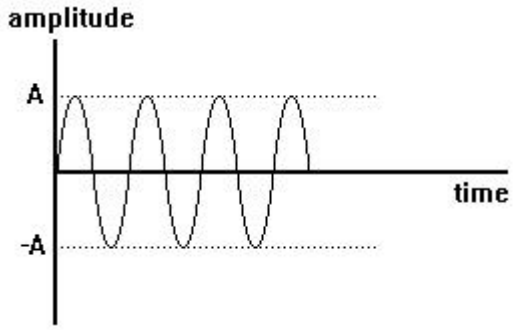
- ✓ refers to all types of data transmission from voice to video.
- ✓ is the transmission of signals over a distance for the purpose of communication.
- ✓ A telecommunication system consists of three basic elements:
  - o a sender that sends information
  - o a transmission medium that carries the information and,
  - o a receiver that receives the information and converts it back into usable information.
- ✓ Telecommunication over a telephone line is called point-to-point communication because it is between one transmitter and one receiver.
- ✓ Telecommunication through radio broadcasts is called broadcast communication because it is between one powerful transmitter and numerous receivers.

The information carrying signals are divided into two broad classes;

- ✓ Analog
- ✓ Digital

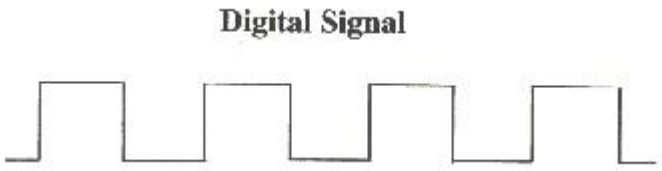
ANALOG SIGNALS

- ✓ Analog signals are continuous electrical signals that vary in time.
- ✓ Analog systems are very tolerant to noise, make good use of bandwidth, and are easy to manipulate mathematically.
- ✓ analog signals require hardware receivers and transmitters that are designed to perfectly fit the particular transmission. While working on a new system, one needs to completely change your transmitters and receivers if there is a change in analog signal.



DIGITAL SIGNALS

- ✓ A digital signal is a discrete signal. It is depicted as discontinuous values of voltage.
- ✓ A digital signal has the following characteristics:
  - o Holds a fixed value for a specific length of time
  - o Has sharp, abrupt changes
  - o A preset number of values allowed



- ✓ Digital signals are intolerant to noise, and digital signals can be completely corrupted in the presence of excess noise. In digital signals, noise could cause a 1 to be interpreted as a 0 and vice versa, which makes the received data different than the original data.
- ✓ Imagine if the army transmitted a position coordinate to a missile digitally, and a single bit was received in error? This single bit error could cause a missile to miss its target by miles. Luckily, there are systems in place to prevent this sort of scenario, such as checksums and CRCs, which tell the receiver when a bit has been corrupted and ask the transmitter to resend the data. The primary benefit of digital signals is that they can be handled by simple,

standardized receivers and transmitters, and the signal can be then dealt with in software (which is comparatively cheap to change).

MODULATION

- ✓ is the process, or results of the process, whereby some characteristic of one signal is varied in accordance with another signal. The modulated signal is called the carrier.
- ✓ Is the process of changing some characteristics of the carrier wave (high frequency wave used to carry the signals) is modulation.
- ✓ The carrier may be modulated in three fundamental ways: by varying the amplitude, called amplitude modulation; by varying the frequency, called frequency modulation; by varying the phase, called phase modulation.
- ✓ A device that performs modulation is known as a modulator and a device that performs the inverse operation of modulation is known as a demodulator (sometimes detector or demod). A device that can do both operations is a modem (short for "Modulator-Demodulator").

WHY MODULATION?

Modulation is necessary in communication due to the following reasons.

- ✓ **Antenna Length**: In order to transmit a wave efficiently, the length of the transmitting antenna should be nearly equal to the wavelength of the wave. So, for transmitting audio wave of very high wavelength, very long antennas are required which is practically impossible.
- ✓ **Operation Range**: The energy of a wave depends upon its frequency. The greater the frequency of the wave, the greater the energy possessed by it. As the audio signal frequencies are small, they cannot be transmitted over long distance.
- ✓ **Wireless Communiation:** At audio frequencies, radiation is not practical because the efficiency of radiation is poor at low frequencies. Hence, wireless communcation would be impractical in such case.

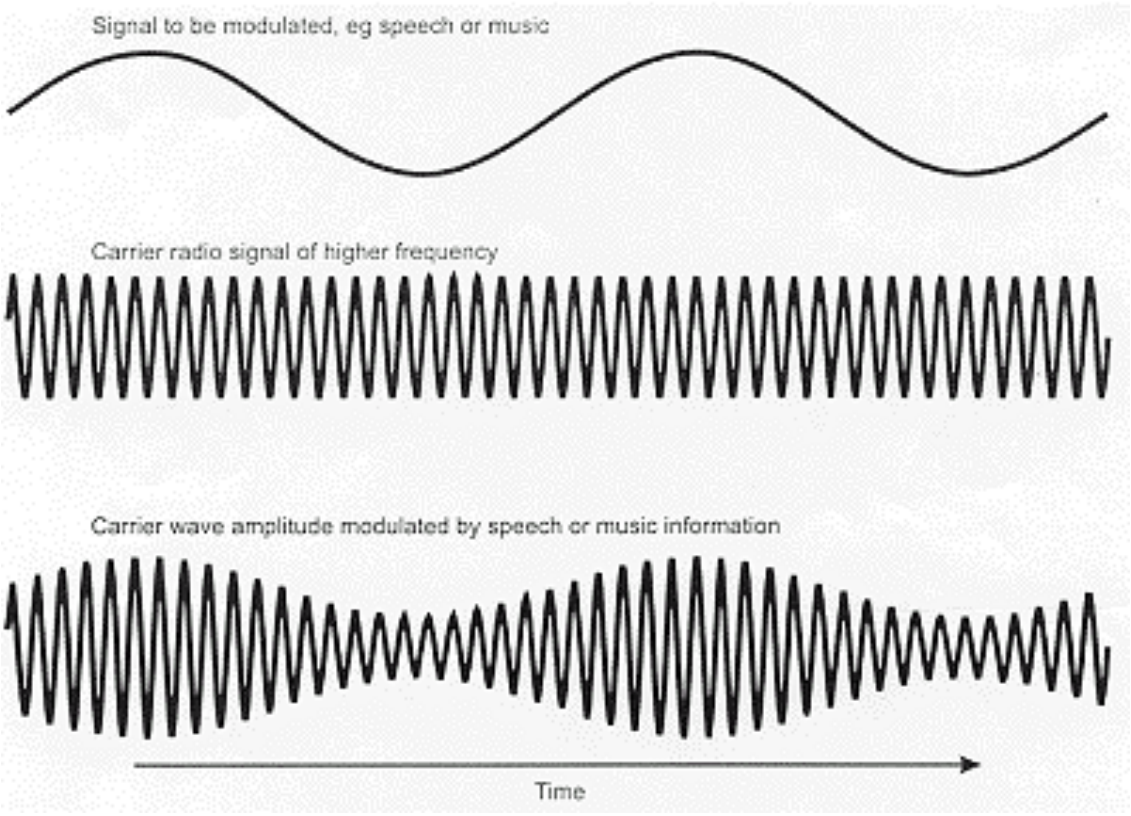
The only practical solution for the above problems is to modulate a high frequency carrier wave with the audio signal and permit the transmission to occur at this high frequency.

TYPES OF MODULATION

1. Amplitude Modulation (AM)
2. Frequency Modulation (FM)
3. Phase Modualtion (PM)

1. Amplitude Modulation (AM)

- ✓ When the amplitude of the high frequency carrier wave is changed in accordance with the intensity of the signal it is called amplitude modulation.
- ✓ The amplitude of the carrier wave changes according to the intensity of the signal.
- ✓ The amplitude variations of the carrier wave is at the signal frequency
- ✓ The frequency of the amplitude moduated wave remains the same - carrier frequency.
- ✓ The depth of modulation is defined by a modulation factor. It determines the extent to which the amplitude of the carrier wave is changed.
- ✓ AM radio ranges from 535 to 1705kHz



Limitations of Amplitude Modulation

- ✓ Noise Reception: In an AM wave, the signal is in the amplitude variations of the carrier. Practically, all natural and human voices consist of electrical amplitude disturbances. A radio receiver can't distinguish between amplitude vibrations that represent noise and those that contain the desired signal, reception is generally noisy.
- ✓ Small Operting Range: Due to low frequency of amplitude modulation, transmitters employing this method have a small operating range.
- ✓ Lack of audio quality: In order to attain high-fidelity reception all audio frequencies upto 15 KHz must be reproduced. This necessitates bandwidth of 30 KHz, but AM broadcasting stations are assigned bandwidth of only 10KHz to minimize interference from adjacent broadcating stations. This means that the highest modulating frequency can be 5 KHz, which is hardly sufficient to reproduce the signal properly.

2. Frequency Modulation (FM)

- ✓ Frequency modulation conveys information over a carrier wave by varying its frequency and the amplitude of the carrier remains constant.
- ✓ The FM radio band goes from 88 to 108 MHz

Advantages

- It gives noiseless reception. Noise is a form of amplitude vibration and FM receivers reject such signals.
- Operating range is quite large.
- It gives high-fidelity (quality) reception
- The efficiency of transmission is very high.

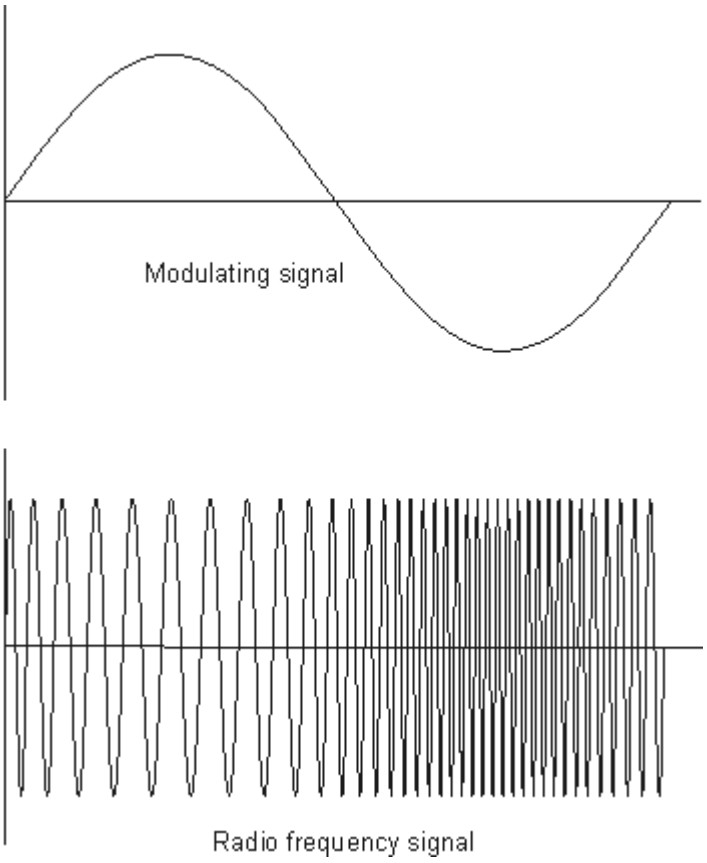


Fig: Frequency Modulation

3. Phase Modulation

- ✓ Phase modulation is the change in the carrier phase angle.
- ✓ The phase angle cannot change without affecting a change in frequency. Hence, it is taken as a second form of frequency modulation.

MODEMS

- ✓ Modem is a device that modulates an analog carrier signal to encode digital information, and also demodulates such a carrier signal to decode the transmitted information.
- ✓ The goal is to produce a signal that can be transmitted easily and decoded to reproduce the original digital data.
- ✓ Modems can be used over any means of transmitting analog signals, from driven diodes to radio.
- ✓ Modems are generally classified by the amount of data they can send in a given time, normally measured in bits per second, or "bps". They can also be classified by Baud, the number of times the modem changes its signal state per second.

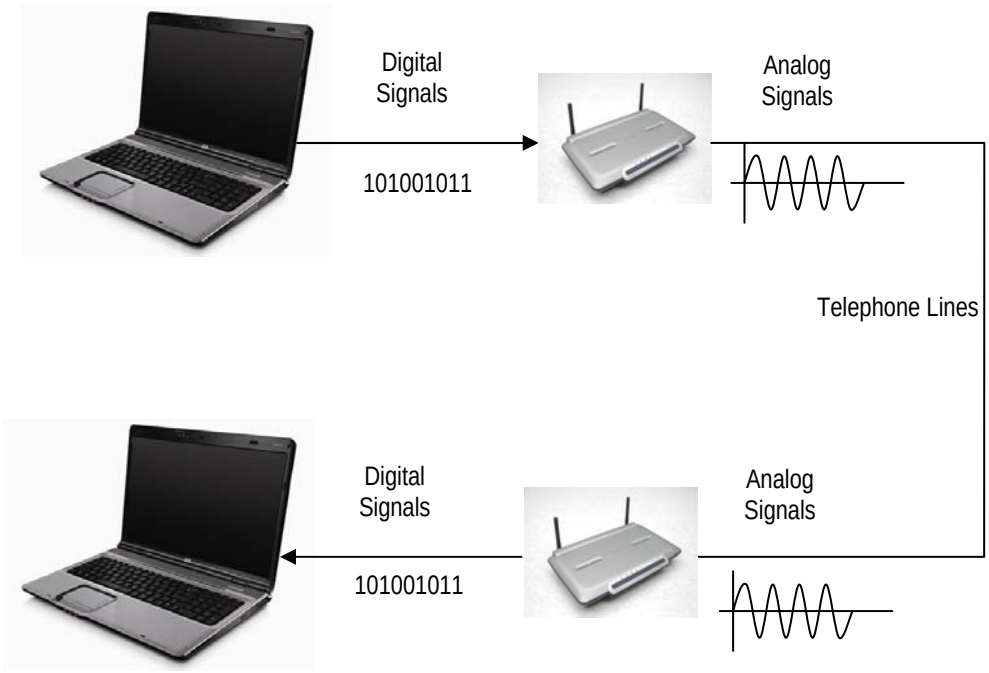


Fig: How MODEM Works

Transfer Speeds

- ✓ Measured in bps (bits per second)
- ✓ General MODEM used for internet connection are 14.4 kbps modems
- ✓ 28.8 kbps, 33.6 kpbs, 56 kbps modems are available in the market.