Data

Data:

It refers to the symbols that represent people, events, things, and ideas. Data can be a name, a number, the colors in a photograph, or the notes in a musical composition.

The term "data" refers to factual information used for analysis or reasoning.

Data Representation:

It refers to the form in which data is stored, processed, and transmitted.

Devices such as smartphones, iPods, and computers store data in digital formats that can be handled by electronic circuitry (motherboard).

Digitization is the process of converting information, such as text, numbers, photo, or music, into digital data that can be manipulated by electronic devices.

The Digital Revolution has evolved through four phases, beginning with big, expensive, standalone computers, and progressing to today's digital world in which small, inexpensive digital devices are everywhere.

> **Digital data**, in information theory and information systems, is information represented as a string of discrete symbols each of which can take on one of only a finite number of values from some alphabet, such as letters or digits. An example is a text document, which consists of a string of alphanumeric characters. The most common form of digital data in modern information systems is binary data, which is represented by a string of binary digits (bits) each of which can have one of two values, either 0 or 1.

Digital data can be contrasted with analog data, which is represented by a value from a continuous range of real numbers. Analog data is transmitted by an analog signal, which not only takes on continuous values, but can vary continuously with time, a continuous real-valued function of time. An example is the air pressure variation in a sound wave.

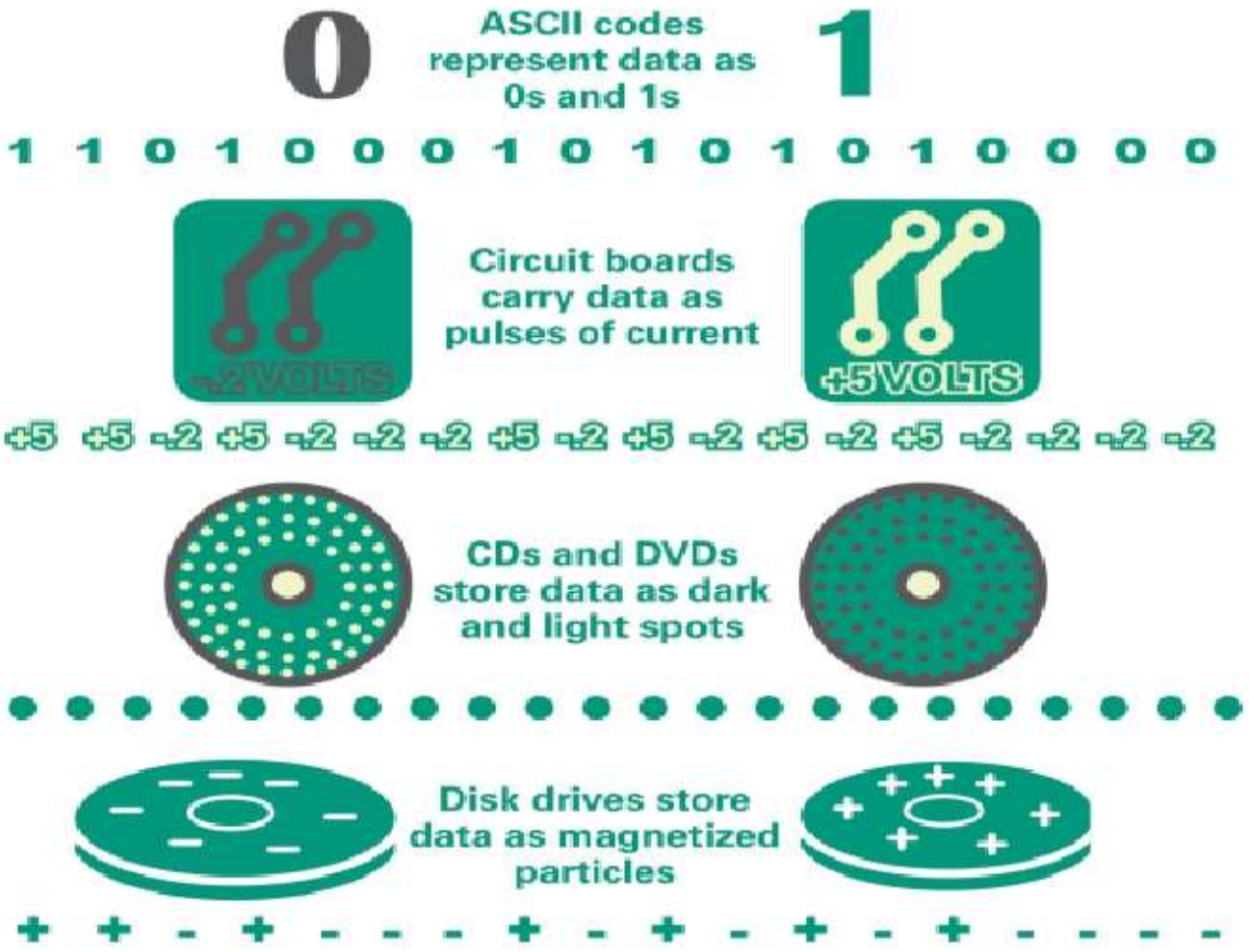
Data Types:

The data types found in the registers of digital computers may be classified as being one of the following categories:

1. Numbers used in arithmetic computation,
2. Letters of the alphabet used in data processing,
3. Other discrete symbols used for specific purposes.

Data are numbers and other binary-coded information that are operated on to achieve required computational results.

Data Representation



Internal Storage Coding of Characters

We know that a computer understands only binary language (0 and 1). Moreover, it is not able to directly understand or store any alphabets, other numbers, pictures, symbols, etc. Therefore, we use certain coding schemes so that it can understand each of them correctly. Let us understand this in the following chapter called Internal Storage Encoding of Characters.

ASCII

ASCII stands for American Standard Code for Information Interchange. The American Standards Association (ASA) introduced it in 1963.

Types of ASCII Codes

We can broadly divide it into two categories as follows:

Standard ASCII

The standard form ranges from 0-127. Hence, it represents a total of 128 characters. Moreover, it uses 7 bits to represent each character with the first bit as 0. We can also call it ASCII-7. This again consists of two parts as follows:

1. Non-Printable Characters: These are certain special characters that are not printed directly. Besides, they perform a certain type of function. For example delete, backspace, escape sequence, end of line, etc. They range from 0-32.

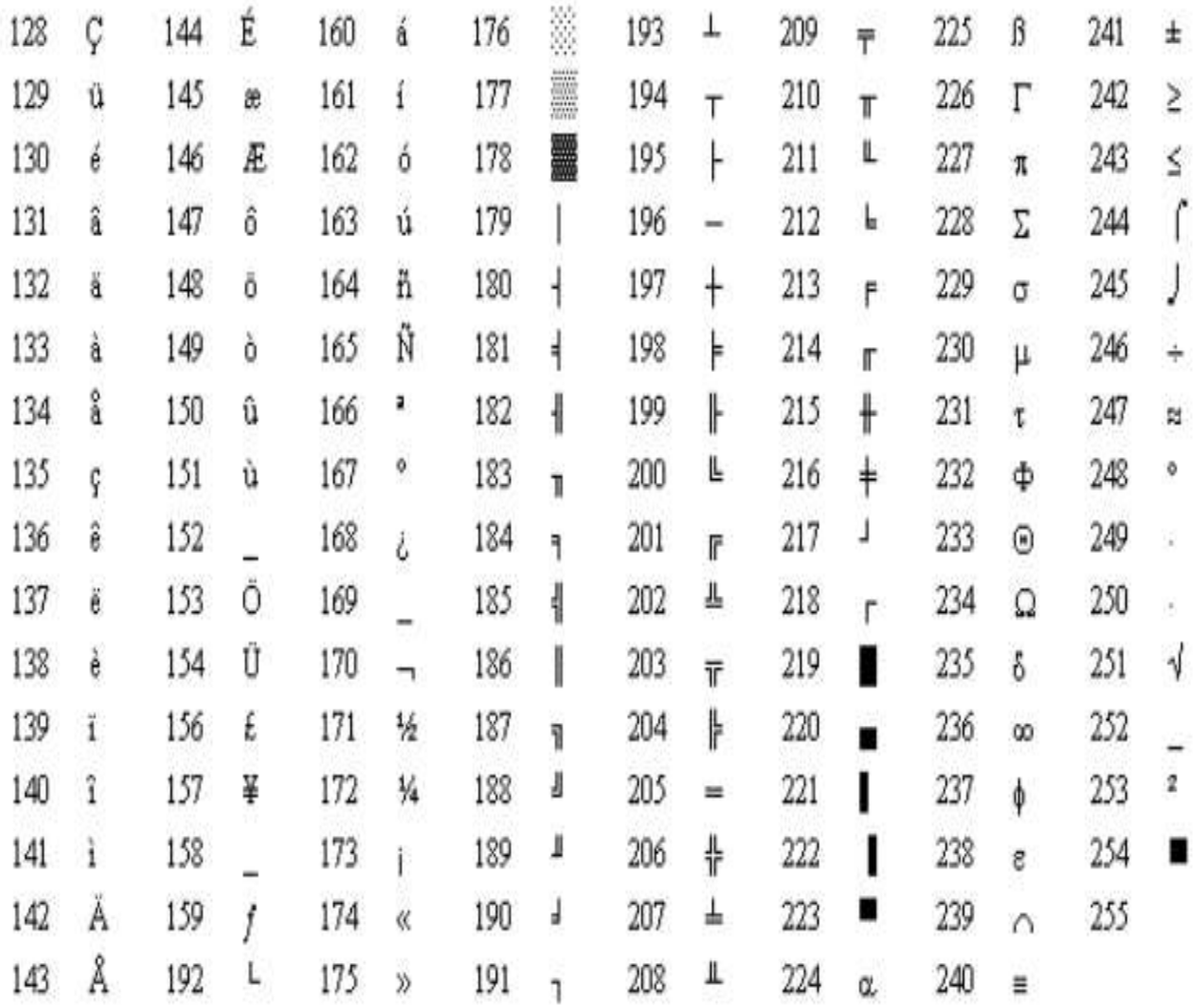
2. Printable Characters: They consist of basic printable characters such as numbers, small alphabets, capital alphabets, special symbols, etc. Moreover, they range from 33 to 127. Furthermore, the code for numbers is between the range of 48-57, for small alphabets, it is between 97-122 and for capital alphabets, it is from 65-90. And the rest is the special symbols like !, @, #, $, %, ^, &, etc.

ASCII TABLE

Character	Binary code	Character	Binary code
A	100 0001	0	011 0000
B	100 0010	1	011 0001
C	100 0011	2	011 0010
D	100 0100	3	011 0011
E	100 0101	4	011 0100
F	100 0110	5	011 0101
G	100 0111	6	011 0110
H	100 1000	7	011 0111
I	100 1001	8	011 1000
J	100 1010	9	011 1001
K	100 1011		
L	100 1100		
M	100 1101	space	010 0000
N	100 1110	.	010 1110
O	100 1111	(	010 1000
P	101 0000	+	010 1011
Q	101 0001	$	010 0100
R	101 0010	*	010 1010
S	101 0011	)	010 1001
T	101 0100	–	010 1101
U	101 0101	/	010 1111
V	101 0110	,	010 1100
W	101 0111	=	011 1101
X	101 1000		
Y	101 1001		
Z	101 1010		

Extended ASCII

This is an extension of the standard form. Though the digits and letters can be easily represented by this code. But, it is not sufficient to represent some special characters or characters of some other language. Therefore, an extension of 128 characters was introduced. Hence, the total number of code now becomes 256. Moreover, these characters have their starting bit as 1 instead of 0. Moreover, it uses 7 bits to represent each character. Hence we can also call it **ASCII-8**.



128	Ç	144	É	160	á	176	▒	193	┴	209	╤	225	ß	241	±
129	ü	145	æ	161	í	177	▓	194	┬	210	╥	226	Γ	242	≥
130	é	146	Æ	162	ó	178	▓	195	├	211	╙	227	π	243	≤
131	â	147	ô	163	ú	179	|	196	−	212	╘	228	Σ	244	⌠
132	ä	148	ö	164	ñ	180	┤	197	┼	213	╒	229	σ	245	⌡
133	à	149	ò	165	Ñ	181	╡	198	╞	214	╓	230	µ	246	÷
134	å	150	û	166	ª	182	╢	199	╟	215	╫	231	τ	247	≈
135	ç	151	ù	167	º	183	╖	200	╚	216	╪	232	Φ	248	°
136	ê	152	_	168	¿	184	╕	201	╔	217	┘	233	Θ	249	·
137	ë	153	Ö	169	_	185	╣	202	╩	218	┌	234	Ω	250	·
138	è	154	Ü	170	¬	186	║	203	╦	219	▉	235	δ	251	√
139	ï	156	£	171	½	187	╗	204	╠	220	▄	236	∞	252	_
140	î	157	¥	172	¼	188	╝	205	=	221	▌	237	ϕ	253	²
141	ì	158	_	173	¡	189	╜	206	╬	222	▐	238	ε	254	■
142	Ä	159	ƒ	174	«	190	╛	207	╧	223	▀	239	∩	255	
143	Å	192	└	175	»	191	┐	208	╨	224	α	240	≡		

Unicode

This coding scheme was introduced in 1991. Since the ASCII characters were not enough to represent all the characters there was a need for another coding standard. As, the old scheme was maybe sufficient for the English characters but, it was not able to represent other languages. Today, the Unicode covers more than 128,000 characters.

Moreover, it has different types according to the bits used for encoding such as UTF-8, UTF-16, UTF-32.

Difference Between Unicode and ASCII
The differences between them are as follows:

Unicode Coding Scheme	ASCII Coding Scheme
• It uses variable bit encoding according to the requirement. For example, UTF-8, UTF-16, UTF-32	• It uses 7-bit encoding. As of now, the extended form uses 8-bit encoding.
• It is a standard form.	• It is not a standard all over the world.
• People use this scheme all over the world.	• It has only limited characters hence, it cannot be used all over the world.
• The Unicode characters themselves involve all the characters of the ASCII encoding. Therefore we can say that it is a superset for it.	• It has its equivalent coding characters in the Unicode.
• It has more than 128,000 characters.	• In contrast, it has only 256 characters.

What is the main difference between Unicode and ASCII coding?

- The main difference between the two coding schemes is that Unicode has over 128000 characters whereas, the other scheme contains only 256 characters. Therefore, Unicode nearly covers all the languages' characters hence, people use it worldwide.

Input Devices and Media
In computing, an input device is a piece of equipment used to provide data and control signals to an information processing system, such as a computer or information appliance. Examples of input devices include keyboards, mouse, scanners, cameras, joysticks, and microphones.

(details in chapter iv)

Computer peripheral device
A peripheral device is an internal or external device that connects directly to a computer or other digital device but does not contribute to the computer's primary function, such as computing. It helps end users access and use the functionalities of a computer.

Since it's not a core device for the system, the computer can still function without the peripheral, which simply provides extra functions. However, some peripherals such as a mouse, keyboard, or monitor tend to be pretty much fundamental to the interaction between the user and the computer itself.

A peripheral device is also called a peripheral, computer peripheral, input-output device, or I/O device.

Peripherals are commonly divided into three kinds: input devices, output devices, and storage devices. (all these are described in details in chapter iv)

Off-line peripherals
A device may be physically connected but offline if the system has been instructed not to use it.

What is data validation?
In a nutshell, data validation is the process of determining whether a particular piece of information falls within the acceptable range of values for a given field.

What is data verification?

Data verification, on the other hand, is actually quite different from data validation. Verification performs a check of the current data to ensure that it is accurate, consistent, and reflects its intended purpose.

Verification may also happen at any time. In other words, verification may take place as part of a recurring data quality process, whereas validation typically occurs when a record is initially created or updated.

	Data validation	Data verification
Purpose	Check whether data falls within the acceptable range of values	Check data to ensure it's accurate and consistent
Usually performed	When data is created or updated	When data is migrated or merged
Example	Checking whether user-entered ZIP code can be found	Checking that all ZIP codes in dataset are in ZIP+4 format

	Data verification	Data validation
1	When user enters email to reset password and presses submit then that email address is matched with the system repository and this process is called verification.	When user enters email to login or in a form, check the syntax to determine if it is correct or incorrect is called validation.
2	Check data against the source to verify its correctness e.g. checking the age against a valid document.	Check if a data value is within a range e.g. age should not be negative and should not be a very large number greater than say 110.
3	Match member id in the membership table, match member benefits in the plan benefits, etc is verification.	Insurance companies validate a claim (logged by provider) to ascertain that the claim is valid using parameters such as eligibility, membership, COB (coordination of benefits) etc. (Note: a claim can be invalid even if it passes all the pre-determined verification steps because of many reasons like claimant hiding some facts that leads to claim rejection on grounds of unclear information.)

UNIT (IX): Data files

Character

In computer science, a character is a display unit of information equivalent to one alphabetic letter or symbol.

It is important to distinguish that a character in computer science is not equal to one bit of machine language. Instead, individual characters are represented by segments of compiled machine language. A universal system for characters has been developed called ASCII. Individual ASCII characters require one byte, or eight bits, of data storage.

The character also plays a critical role in computer programming, where it may be represented in code languages as "chr" or "char." A character is one single unit of a text or character string, where the individual characters and the entire string are manipulated in various ways by code functions.

Field

A combination of one or more characters is called field. It is the smallest unit of data that can be accessed by the user. The name of each field in a record is unique. The data type of a field indicates the type of data that can be stored in the field. Each field contains one specific piece of information. A field size defines the maximum number of characters that can be stored in a field.

For example, Employee Number, Employee Name, Grade and Designation are fields.

Fields	
Employee Number	0007
Employee Name	Umar akmal
Grade	1/4
Designation	Senior Manager

Record

A collection of related fields treated as a single as a single unit is called a record. For example a employee's record includes a set of fields that contains Employer Number, Employee Name, Grade and designation etc.

File

A collection of related records treated as a single unit is called file. File is also known as data set. Files are stored in disk like hard disk, CD-ROM or DVD-ROM etc. A Student file may contain the records of hundreds of students. Each student's record consists of same fields but each field contains different data.

What is File Organization?

File Organization refers to the logical relationships among various records that constitute the file, particularly with respect to the means of identification and access to any specific record. In simple terms, Storing the files in certain order is called file Organization. File Structure refers to the format of the label and data blocks and of any logical control record.

File organization

A file consists of a number of records. Each record is made up of a number of fields and each field consists of a number of characters. In order to produce useful information by means of computerized data processing, it is very necessary to organize data in systematic way. Methods of organizing data are referred to as data structures.

COMPUTER FILE ORGANIZATION METHODS

File organization refers to the way data is stored in a file. File organization is very important because it determines the method of access, efficiency, flexibility and storage devices to be used. There are four methods of organizing files on a storage media namely:

- serial
- Sequential
- indexed- sequential
- random

Serial file Organization

• Serial file organization is the simplest file organization method. This is type of file design where records are stored in the storage media chronologically. i.e. in the order they occur. In serial files, records are entered in the order of their creation. As such, the file is unordered, and is at best in chronological order.

• Serial files are primarily used as transaction files in which the transactions are recorded in the order that they occur.

• The records are accessed are accessed from the storage media serially from head to tail.

• This type of access is normally used magnetic tapes.

• The hit rate for serial files is high- Hit rate refers to the number of records accessed at a given period of time.

• Serial files are suited for high activity processing e.g. batch processing where a group of records are collected and processed at the same time.

• Serial file can only be accessed serially, that is search through the file starting from the head of the file to tail of the file.

Sequential file Organization

• A sequentially organized file consists of records arranged in the sequence in which they are written to the file e.g. Alphabetically, numerically etc. (the first record written is the first record in the file, the second record written is the second record in the file, and so on). As a result, records can be added only at the end of the file. Attempting to add records at some place other than the end of the file will result in the file begin truncated at the end of the record just written.

• Sequential files are usually read sequentially, starting with the first record in the file. Sequential files with a fixed-length record type that are stored on disk can also be accessed relative record number (direct access).

• Records in sequential files can be read or written only sequentially.

• After you have placed a record into a sequential file, you cannot shorten, lengthen, or delete the record. However, you can update (REWRITE) a record if the length does not change. New records are added at the end of the file.

• If the order in which you keep records in a file is not important, sequential organization is a good choice whether there are many records or only a few. Sequential output is also useful for printing reports.

• The most suitable storage media for sequential files is magnetic tapes.

• Sequential files are ideal for high activity processing e.g. batch processing.

Direct access and serial access

"Sequential access must begin at the beginning and access each element in order, one after the other. Direct access allows the access of any element directly by locating it by its index number or address. Arrays allow direct access. Magnetic tape has only sequential access, but CDs had direct access. If you are on a railroad train, to go from one car to another you must use sequential access. But when you board the train initially you have direct access. Direct access is faster than sequential access, but it requires some external mechanism (array index, file byte number, railroad platform)."

Indexed- Sequential file Organization

• Indexed file contains records ordered a record key. Each record contains a field that contains the record key. The record key uniquely identifies the record and determines the sequence in which it is accessed with respect to other records. A record key for a record might be, for example, an employee number or an invoice number.

• An indexed file can also use alternate indexes, that is, record keys that let you access the file using a different logical arrangement of the records. For example, you could access the file through employee department rather than through employee number.

• The records are arranged sequentially as in sequential files but the difference is that there is an index that allows for selective access. The indexes are used to point particular portion where the records are stored in groups, this allows the by-passing of a group of records that are not
required in a particular processing run.

• The best storage media for index sequential files is a magnetic disc (hard disk) the records can be accessed using the following methods: –

i. Sequential Access – this is where the user will use the specific sequence e.g. alphabetic or numeric to retrieve a record of interest.

ii. Use of Indices – this is where the user will use the unique index number to retrieve a record of interest.

iii. Random access – this is where user moves up and down in a none orderly manner in order to retrieve a record of interest.

The hit rate is both high and low – hit rate refers to the number of records that can be accessed at a given period of time. It is high because of the sequential access and low because of index access. Index sequential files are ideal for online processing and batch processing.

Database Concepts

Database system is an excellent computer-based record-keeping system. A collection of data, commonly called a database, contains information about a particular enterprise. It maintains any information that may necessary to the decision-making process involved in the management of that organization. It can also be defined as a collection of interrelated data stored together to serve multiple applications, the data is stored so that it is independent of programs that use the data. A generic and controlled approach is used to add new data and modify and retrieve existing data within the database. The data is structured so as to provide the basis for future application development.

Purpose of Database

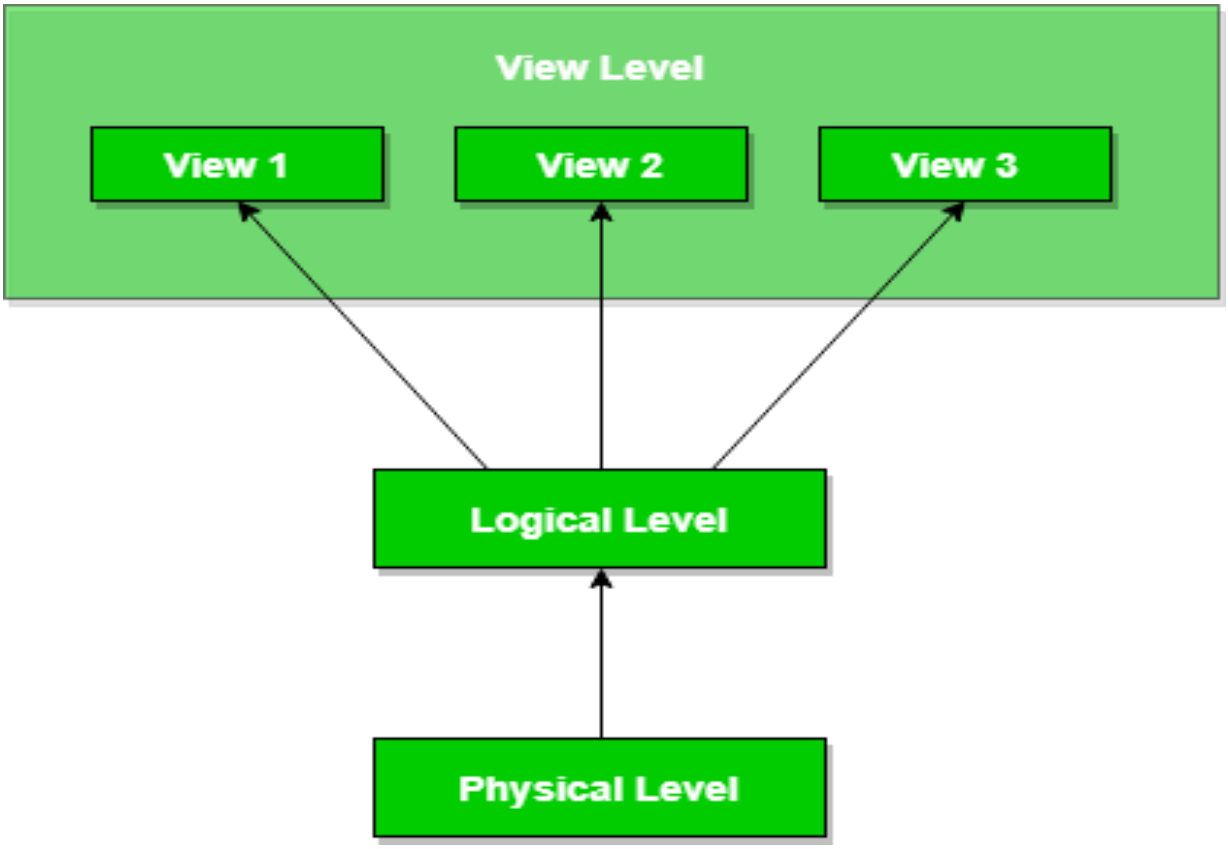
The intent of a database is that a collection of data should serve as many applications as possible. Therefore, a database is often thought of as a repository of information needed to run certain functions in a corporation or organization. It would permit only the retrieval of data but also the continuous modification of data needed for the control of operations. It may be possible to search the database to obtain answers to questions or information for planning purposes.

In a typical file-processing system, permanent records are stored in different files. Many different application programs are written to extract the records and add the records to the appropriate files. But this scheme has several major limitations and disadvantages, such as data redundancy (duplication of data), data inconsistency, maladaptive data, non-standard data, insecure data, incorrect data, etc. A database management system is an answer to all these problems as it provides centralized control of the data.

Database Abstraction

A major purpose of a database is to provide the user with only as much information as is required of them. This means that the system does not disclose all the details of the data, rather it hides some details of how the data is stored and maintained. The complexity of databases is hidden from them which, if necessary, are ordered through multiple levels of abstraction to facilitate their interaction with the system. The different levels of the database are implemented through three layers:

1. Internal Level(Physical Level): The lowest level of abstraction, the internal level, is closest to physical storage. It describes how the data is stored concretely on the storage medium.
2. Conceptual Level: This level of abstraction describes what data is concretely stored in the database. It also describes the relationships that exist between the data. At this level, databases are described logically in terms of simple data structures. Users at this level are not concerned with how these logical data structures will be implemented at the physical level.
3. External Level(View Level): It is the level closest to users and is related to the way the data is viewed by individual users.



Since a database can be viewed through three levels of abstraction, any change at one level can affect plans at other levels. As databases continue to grow, there may be frequent changes to it at times. This should not lead to redesign and re-implementation of the database. In such a context the concept of data independence proves beneficial.

1. Database Schema: It is a design of the database. Or we can say that it is a skeleton of the database that is used to represent the structure, types of data will be stored in the rows and columns, constraints, relationships between the tables.
2. Data Constraints: In a database, sometimes we put some restrictions on the table that what type of data can be stored in one or more columns of the table, it can be done by using constraints. Constraints are defined while we are creating a table.
3. Data dictionary or Metadata: Metadata is known as the data about the data. Or we can say that the database schema along with different types of constraints on the data is stored by DBMS in the dictionary is known as metadata.
4. Database instance: In a database, a database instance is used to define the complete database environment and its components. Or we can say that it is a set of memory structures and background processes that are used to access the database files.
5. Query: In a database, a query is used to access data from the database. So users have to write queries to retrieve or manipulate data from the database.
6. Data manipulation: In a database, we can easily manipulate data using the three main operations that is Insertion, Deletion, and updation.
7. Data Engine: It is an underlying component that is used to create and manage various database queries.
Advantages of Database

Let us consider some of the benefits provided by a database system and see how a database system overcomes the above-mentioned problems:-

1. Reduces database data redundancy to a great extent
2. The database can control data inconsistency to a great extent
3. The database facilitates sharing of data.
4. Database enforce standards.
5. The database can ensure data security.
6. Integrity can be maintained through databases.

Therefore, for systems with better performance and efficiency, database systems are preferred.

Disadvantages of Database

With the complex tasks to be performed by the database system, some things may come up which can be termed as the disadvantages of using the database system. These are:-

1. Security may be compromised without good controls.
2. Integrity may be compromised without good controls.
3. Extra hardware may be required
4. Performance overhead may be significant.
5. The system is likely to be complex.

Subtraction using 1's and 2's complement methods

In subtraction by 1's complement we subtract two binary numbers using carried by 1's complement.

The steps to be followed in subtraction by 1's complement are:

i) To write down 1's complement of the subtrahend.

ii) To add this with the minuend.

iii) If the result of addition has a carry over then it is dropped and an 1 is added in the last bit.

iv) If there is no carry over, then 1's complement of the result of addition is obtained to get the final result and it is negative.

Evaluate:

(i) 110101 – 100101

Solution:

1's complement of 10011 is 011010. Hence

Minued -		1 1 0 1 0 1
1's complement of subtrahend -		0 1 1 0 1 0
Carry over -	1	0 0 1 1 1 1

1

0 1 0 0 0 0

(ii) 101011 – 111001

Solution: 1's complement of 111001 is 000110. Hence

Minued - 1 0 1 0 1 1

1's complement - 0 0 0 1 1 0

1 1 0 0 0 1

Hence the difference is – 1 1 1 0

2's complement subtraction

steps :

1. At first, find 2's complement of the B(subtrahend).
2. Then add it to the A(minuend).
3. If the final carry over of the sum is 1, then it is dropped and the result is positive.
4. If there is no carry over, then 2's complement of the sum is the final result and it is negative.

1. Find Subtraction of 110 and 101 using 2's complement method

Here A = 110, B = 101.
Find A - B = ? using 2's complement
First find 2's complement of B = 101

Note : 2's complement of a number is 1 added to it's 1's complement number.
1's complement of 101 is

```
    1  1  1
  - 1  0  1
 ------------
    0  1  0
```

Now add 1 : 010 + 1 = 011

Now Add this 2's complement of B to A

```
  1  1
     1  1  0
  +  0  1  1
 ------------
  1  0  0  1
```

+ Hints : (Move mouse over the steps for detail calculation highlight)

The left most bit of the result is called carry and it is ignored.
So answer is 001

2. Find Subtraction of 10110 and 11101 using 2's complement method

Here A = 10110, B = 11101.
Find A - B = ? using 2's complement
First find 2's complement of B = 11101

Note : 2's complement of a number is 1 added to it's 1's complement number.
1's complement of 11101 is

```
    1  1  1  1  1
 -  1  1  1  0  1
-------------------
    0  0  0  1  0
```

Now add 1 : 00010 + 1 = 00011

Now Add this 2's complement of B to A

```
       1  1
    1  0  1  1  0
 +  0  0  0  1  1
-------------------
    1  1  0  0  1
```

✚ Hints : (Move mouse over the steps for detail calculation highlight)

Here there is no carry, answer is - (2's complement of the sum obtained 11001)

Note : 2's complement of a number is 1 added to it's 1's complement number.
1's complement of 11001 is

```
    1  1  1  1  1
 -  1  1  0  0  1
-------------------
    0  0  1  1  0
```

Now add 1 : 00110 + 1 = 00111
So answer is -00111