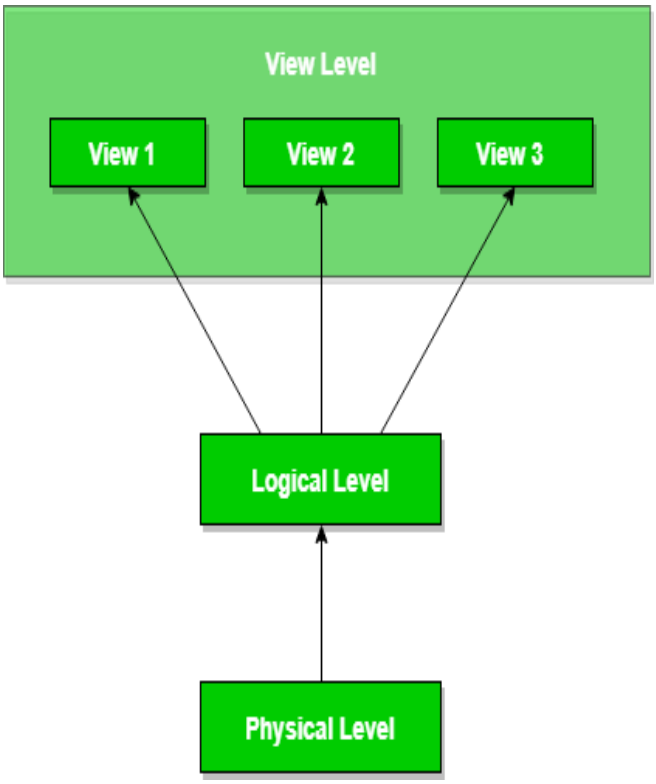
Database Concepts (File System and DBMS)

A database is a shared collection of related data used to support the activities of a particular organization. A database can be viewed as a repository of data that is defined once and then accessed by various users.

A collection of data, commonly called a database, contains information about a particular enterprise. It maintains any information that may be necessary to the decision-making process involved in the management of that organization.

The different levels of the database are implemented through three layers:

1. **Internal Level (Physical Level: T**he 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.
1. **External Level (View Level):** It is the level closest to users and is related to the way the data is viewed by individual users.



A. <u>Key Concept of Database</u>

To store and manage data efficiently in the database there are some key terms:

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.

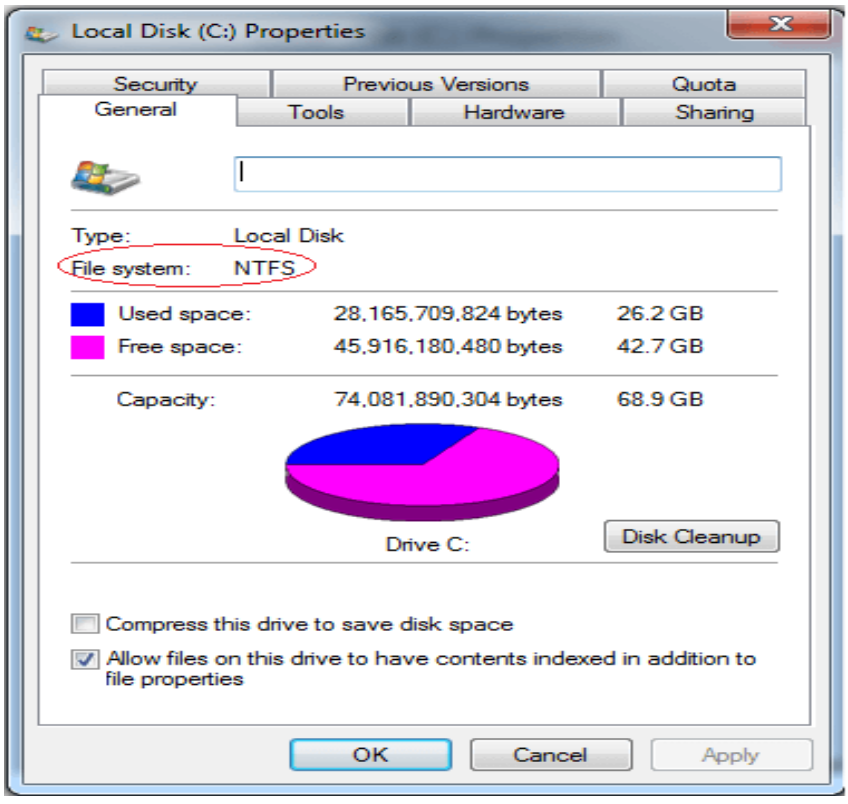
5. Query: In a database, a query is used to access data from the database. So users have to write queries to retrieve or manipulate data from the database.

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.

B.What is file system, its need?

A file system is a structure of directories that is used to organize and store files. A file system is a structure used by an operating system to organize and manage files on a storage device such as a hard drive, solid state drive (SSD), or USB flash drive. It defines how data is stored, accessed, and organized on the storage device. The most important purpose of a file system is to manage user data. This includes storing, retrieving and updating data.



Examples of file systems

1. **FAT:** FAT is a type of file system, which is developed for hard drives. It stands for file allocation table and was first introduced in 1977, which is used for 12 or 16 bits for each and every cluster access into the file allocation table (FAT).

   On hard drives and other computer systems, it helps to manage files on Microsoft operating systems. In devices like digital cameras, flash memory, and other portable devices, it is also often found that is used to store file information.

   Today, FAT is not used by later versions of Microsoft Windows like Windows XP, Vista, 7, and 10 as they use NTFS.

2. **GFS:** A GFS is a file system, which stands for Global File System. It has the ability to make enable multiple computers to act as an integrated machine, which is first developed at the University of Minnesota. But now it is maintained by Red Hat. When the physical distance of two or more computers is high, and they are unable to send files directly with each other, a GFS file system makes them capable of sharing a group of files directly. A computer can organize its I/O to preserve file systems with the help of a global file system.
3. **HFS**: HFS (Hierarchical file system) is the file system that is used on a Macintosh computer for creating a directory at the time a hard disk is formatted. Generally, its basic function is to organize or hold the files on a Macintosh hard disk. Apple is not capable of supporting to write to or format HFS disks since when OS X came on the market. Also, HFS-formatted drives are not recognized by Windows computers as HFS is a Macintosh format. With the help of WIN32 or NTFS file systems, Windows hard drives are formatted.
4. **NTFS**: NTFS is the file system, which stands for NT file system and stores and retrieves files on Windows NT operating system and other versions of Windows like Windows 2000, Windows XP, Windows 7, and Windows 10. Sometimes, it is known as the New Technology File System.
5. **UDF**: A UDF is a file system, stands for Universal Disk Format and used first developed by OSTA (Optical Storage Technology Association) in 1995 for ensuring consistency among data written to several optical media. It is used with CD-ROMs and DVD-ROMs and is supported on all operating systems. Now, it is used in the process of CD-R's and CD-RW's, called packet writing.

C.What is DBMS and its need?

Database Management System (DBMS) is software for storing and retrieving users' data while considering appropriate security measures. It consists of a group of programs that manipulate the database. The DBMS accepts the request for data from an application and instructs the operating system to provide the specific data. In large systems, a DBMS helps users and other third-party software store and retrieve data.

History of DBMS

- 1960 – Charles Bachman designed the first DBMS system
- 1970 – Codd introduced IBM'S Information Management System (IMS)
- 1976- Peter Chen coined and defined the Entity-relationship model, also known as the ER model
- 1980 – Relational Model becomes a widely accepted database component
- 1985- Object-oriented DBMS develops.
- 1990s- Incorporation of object-orientation in relational DBMS.
- 1991- Microsoft ships MS access, a personal DBMS, and that displaces all other personal DBMS products.
- 1995: First Internet database applications
- 1997: XML applied to database processing. Many vendors begin to integrate XML into DBMS products.
- Characteristics of DBMS

Characteristics and properties of a Database Management System:

- Provides security and removes redundancy
- Self-describing nature of a database system
- Insulation between programs and data abstraction
- Support of multiple views of the data
- Sharing of data and multiuser transaction processing
- Database Management Software allows entities and relations among them to form tables.
- It follows the ACID concept (Atomicity, Consistency, Isolation, and Durability).
- DBMS supports a multi-user environment that allows users to access and manipulate data in parallel.

DBMS vs. Flat File System

S.No	DBMS	Flat File Management System
1	Multi-user access	It does not support multi-user access
2	Design to fulfil the need of small and large businesses	It is only limited to smaller DBMS systems.
3	Remove redundancy and Integrity.	Redundancy and Integrity issues
4	Expensive. But in the long term Total Cost of Ownership is cheap	It's cheaper
5	Easy to implement complicated transactions	No support for complicated transactions

Popular DBMS Software systems:

MySQL, Microsoft Access, Oracle, PostgreSQL, FoxPro, SQLite, IBM DB2, LibreOffice Base MariaDB, Microsoft SQL Server

E.Codd's 12 rules for RDBMS

Codd's rule in DBMS also known as Codd's 12 rules/commandments is a set of thirteen rules (numbered 0 to 12) that define a database to be a correct Relational Database Management System (RDBMS). If a database follows Codd's 12 rules, it is called a True relational database management system. Cod's 12 rules are used to determine whether a relational database management system is a true relational database management system or not.

Rule 0: The Foundation Rule

For a system to be qualified as a relational database management system, it must be able to manage databases entirely through its relational capabilities.

Rule 1: The Information Rule

The information in a relational database must be stored in columns or rows of a table, i.e., a cell.

Rule 2: The Guaranteed Access Rule

Each and every datum in a relational database must be logically accessible using the combination of the table name, primary key value, and column name.

A datum is an atomic value, i.e., a piece of information that cannot be broken down further.

Rule 3: Systematic Treatment Of NULL Values

NULL values are fully supported in a relational database and represent missing information or inapplicable information in a systematic way, independent of the data type. NULL values are different from empty strings, blank spaces, and 00.

Rule 4: Active/Dynamic Online Catalog Based On The Relational Model

Database Description (Catalog) of a complete database must be stored online. The rules of the rest of the database must also apply to the Catalog. The query language used to query the database should be used for the catalog also.

Rule 5: The Comprehensive Data Sublanguage Rule

Relational systems can support multiple languages and different modes of using terminals, such as fill-in-the-blanks mode. However, there must be at least one language whose statements are expressible according to a well-defined syntax.

Rule 6: The View Updating Rule

Theoretically, updatable *views* are also practically updatable by the database system.

Rule 7: High-Level Insert, Update & Delete Rule

The database system must follow high-level relational operations such as insertion, updation, and deletion at each level or row by row. It also supports the union, intersection, and subtraction operations in database systems.

Rule 8: Physical Data Independence

The working of a database system should be independent of the physical storage of its data. If a file is modified (renamed or moved to another location), it should not interfere with the working of the system.

Rule 9: Logical Data Independence

If there is a change in the logical structure (table structure) of the database, the user view of the data must not change.
Say a table is partitioned into two tables, the new view should give the result as the join of the two tables.

Rule 10: Integrity Independence

Integrity constraints specific to a particular relational database must be defined in the relational data sub-language and stored in the catalogue and not in the application.

Rule 11: Distribution Independence

A database should work properly regardless of its distribution across a network. The end-user should not be able to see that the data is distributed over many locations, they should always get the impression that the data is located at a single site only.

Rule 12: The Non-subversion Rule

If a relational system allows low-level access, that low cannot be used to subvert or bypass the integrity rules to modify the data. This can be achieved with some sort of looking or encryption.

Session 2:

1. **Database Storage Structure**
   - **A. Table Space**

     A table space is a storage structure containing tables, indexes, large objects, and long data. They are used to organize data in a database into logical storage groupings that relate to where data is stored on a system. Table spaces are stored in database partition groups.

     A table space is a unit of database storage that is roughly equivalent to a file group in Microsoft SQL Server. Table spaces allow storage and management of database objects within individual groups

   - **B. Control File**

     Every Oracle database has a control file. A control file is a small binary file that records the physical structure of the database and includes:

     - The database name
     - Names and locations of associated data files and online redo log files
     - The timestamp of the database creation
     - The current log sequence number
     - Checkpoint information

     A control file is an automatically generated file that records process specifications and the success or failure of processing. Control files use the **.cf** file extension by default.

   - **C. Data file**

     Data files contain data and objects such as tables, indexes, stored procedures, and views. Log files contain the information that is required to recover all transactions in the database. Data files can be grouped together in file groups for allocation and administration purposes.

     A data file is a computer file which stores data to be used by a computer application or system, including input and output data.

2. Structured and Unstructured Data

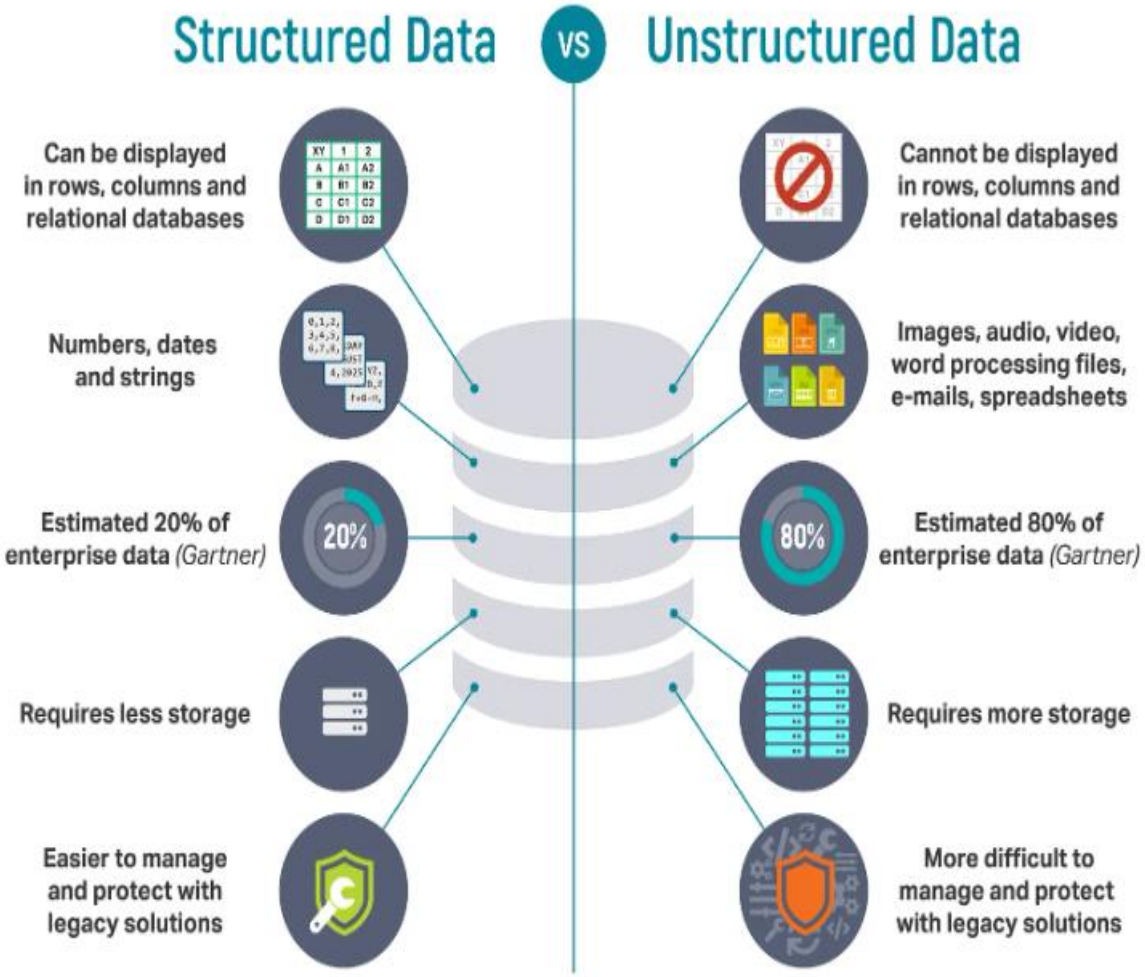
Structured data is highly specific and is stored in a predefined format, where unstructured data is a compilation of many varied types of data that are stored in their native formats.

Structured data examples

- Excel files, SQL databases, Point-of-sale data, Web form results.
- Search engine optimization (SEO) tags, Product directories.
- Inventory control, Reservation systems.

Examples of unstructured data:

Rich media. Media and entertainment data, surveillance data, geo-spatial data, audio, weather data. Document collections,Pdf file etc. Retailers, manufacturers and other companies analyse unstructured data to improve customer experience and enable targeted marketing. They also do sentiment analysis to better understand customers and identify attitudes about products, customer service and corporate brands.



Structured vs unstructured data

Properties	Structured data	Unstructured data
Formats	Several formats	A huge variety of formats
Data model	Pre-defined/not flexible	Not pre-defined/flexible
Storages	Data warehouses	Data lakes
Databases	SQL Relational databases	NoSQL Non-relational databases
Ease of search	Easy to search	Difficult to search
Data nature	Quantitative	Qualitative
Analysis methods	• Classification • Regression • Data clustering	• Data stacking • Data mining
Tools and technologies	• RDBMS • CRM • OLAP • OLTP	• NoSQL DBMS • AI-driven tools • Data storage architectures • Data visualization tools
Specialists to handle data	Business analysts Software engineers Marketing analysts	Data scientists, engineers, and analysts with deep expertise

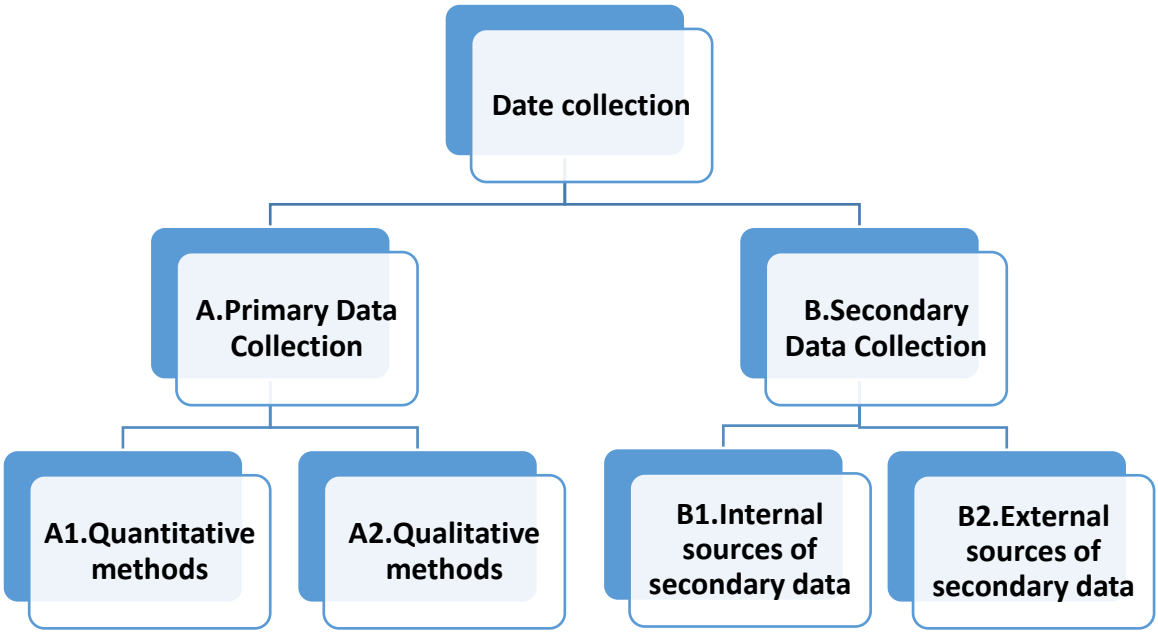
Key difference between Structure and unstructured data

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

3. **Introduction to Data Collection like what is data collection.**

Data collection is a process of gathering information from all the relevant sources. Data is a collection of facts, figures, objects, symbols, and events gathered from different sources. Organizations collect data with various data collection methods to make better decisions. Without data, it would be difficult for organizations to make appropriate decisions, so data is collected from different audiences at various points in time.

<u>Date collection may be categorised in to two ways:</u>



A. Primary Data Collection Methods

Primary data is collected from first-hand experience and is not used in the past. The data gathered by primary data collection methods are specific to the research's motive and highly accurate.

Primary data collection methods can be divided into two categories:

A1. Quantitative methods

Quantitative techniques for market research and demand forecasting usually use statistical tools. In these techniques, demand is forecasted based on historical data. These methods of primary data collection are generally used to make long-term forecasts. Statistical analysis methods are highly reliable as subjectivity is minimal in these methods.

A2. Qualitative methods.

Qualitative data collection methods are especially useful in situations when historical data is not available. Or there is no need of numbers or mathematical calculations.

Qualitative research is closely associated with words, sounds, feeling, emotions, colours, and other elements that are non-quantifiable. These techniques are based on experience, judgment, intuition, conjecture, emotion, etc. There are following ways to deal this type of data:

Surveys: Surveys are used to collect data from the target audience and gather insights into their preferences, opinions, choices, and feedback related to their products and services. Most survey software often has a wide range of question types to select.

Polls: Polls comprise one single or multiple-choice question. You can go for polls when it is required to have a quick pulse of the audience's sentiments. Because they are short in length, it is easier to get responses from people.

Interviews: In this method, the interviewer asks the respondents face-to-face or by telephone. In face-to-face interviews, the interviewer asks a series of questions to the interviewee in person and notes down responses. If it is not feasible to meet the person, the interviewer can go for a telephone interview. This form of data collection is suitable for only a few respondents. It is too time-consuming and tedious to repeat the same process if there are many participants.

Delphi Technique: In Delphi method, market experts are provided with the estimates and assumptions of forecasts made by other experts in the industry. Experts may reconsider and revise their estimates and assumptions based on the information provided by other experts. The consensus of all experts on demand forecasts constitutes the final demand forecast.

Focus Groups: A focus group is one of the examples of qualitative data in education. In a focus group, a small group of people, around 8-10 members, discuss the common areas of the research problem. Each individual provides his or her insights on the issue concerned. A moderator regulates the discussion among the group members. At the end of the discussion, the group reaches a consensus.

Questionnaire: A questionnaire is a printed set of questions, either open-ended or closed-ended. The respondents must answer based on their knowledge and experience with the issue. The questionnaire is a part of the survey, whereas the questionnaire's end goal may or may not be a survey.

B. Secondary Data Collection Methods

Secondary data is the data that has been used in the past. The researcher can obtain data from the data sources, both internal and external, to the organizational data.

B1. Internal sources of secondary data:

- Organization's health and safety records
- Mission and vision statements
- Financial Statements
- Magazines
- Sales Report
- CRM Software
- Executive summaries

B2. External sources of secondary data:

- Government reports
- Press releases
- Business journals
- Libraries
- Internet

4. **<u>The tools and how data can be gathered in a systematic fashion</u>**

Here are some of the data collection techniques used by the Data Collection Tools-

- **Interviews**
- **Questionnaires**
- **Case Studies**
- **Usage Data**
- **Checklists**
- **Surveys**
- **Observations**
- **Documents and records**
- **Focus groups**
- **Oral histories**

<u>Data can be gathered in a systematic fashion as following cycle:</u>