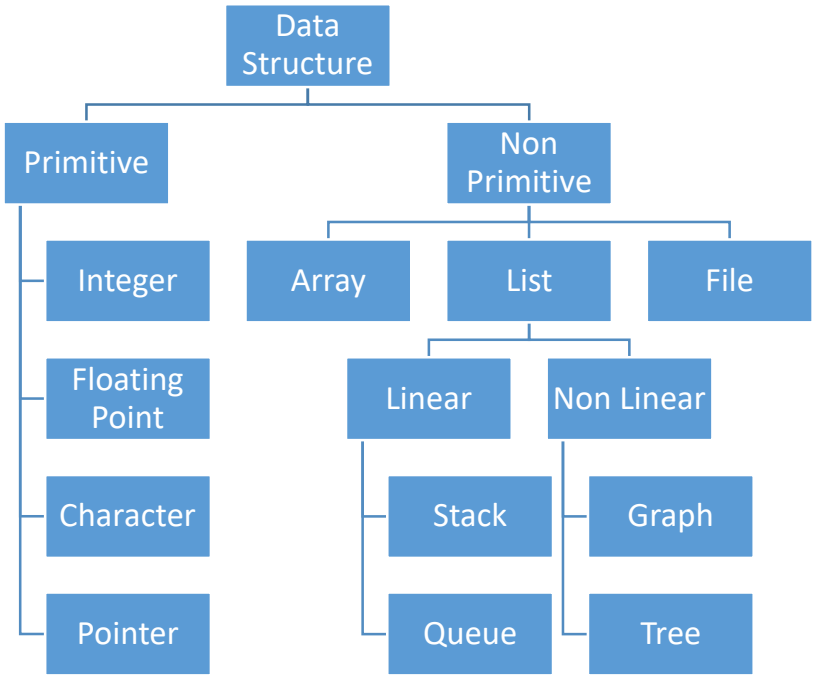
4.3 Data Structure, Database and Data Normalization

4.3.1 Data Structure

Data Structure is a representation of the logical relationship existing between individual elements of data. It is a way of organizing all data items that considers not only the elements stored but also their relationship to each other. Data Structure mainly specifies the following four things:

- Organization of Data
- Accessing methods
- Degree of associativity
- Processing alternatives for information

4.3.1.1 Classification of Data Structure



4.3.1.2 Operation on Data Structures

Design of efficient data structure must take operations to be performed on the data structures into account. The most commonly used operations on data structure are as follows:

- Create
- Delete
- Select
- Update

- Search
- Sort
- Merge
- Split
- Traverse

4.3.2 Database

A database is a collection of information that is organized so that it can easily be accessed, managed, and updated. Databases are designed to offer an organized mechanism for storing, managing and retrieving information. They do so through the use of tables.

Relational DataBase Management System (RDBMS) is a program that lets you create, update, and administer a relational database. Most commercial RDBMS's use the Structured Query Language (SQL) to access the database.

4.3.2.1 Entity-relationship diagrams

An entity relationship diagram (ERD) shows the relationships of entity sets stored in a database. An entity in this context is a component of data. In other words, ER diagrams illustrate the logical structure of databases.

4.3.2.2 Common ER Diagram Symbols

	Entities, which are represented by rectangles. An entity is an object or concept about which you want to store information.
	Actions, which are represented by diamond shapes, show how two entities share information in the database.
	Attributes, which are represented by ovals. A key attribute is the unique, distinguishing characteristic of the entity.

4.3.3 Data Normalization

Normalization is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships. It is a multi-step process that puts data into tabular form, removing duplicated data from the relation tables.

Normalization is used for mainly two purposes,

- Eliminating redundant (useless) data.
- Ensuring data dependencies make sense i.e data is logically stored.

4.3.3.1 List of Database Normal Forms

- 1NF (First Normal Form)
- 2NF (Second Normal Form)
- 3NF (Third Normal Form)
- BCNF (Boyce-Codd Normal Form)
- 4NF (Fourth Normal Form)
- 5NF (Fifth Normal Form)

i. First Normal Form

- Remove horizontal redundancies
    - No two columns hold the same information
    - No single column holds more than a single item
- Each row must be unique
    - Use a primary key
- Benefits
    - Easier to query/sort the data
    - More scalable
    - Each row can be identified for updating

ii. Second Normal Form

- Table must be in First Normal Form
- Remove vertical redundancy
    - The same value should not repeat across rows
- Composite keys
    - All columns in a row must refer to BOTH parts of the key
- Benefits
    - Increased storage efficiency
    - Less data repetition

iii. Third Normal Form

- Table must be in Second Normal Form
    - If your table is 2NF, there is a good chance it is 3NF
- All columns must relate directly to the primary key
- Benefits
    - No extraneous data

iv. BCNF (Boyce-Codd Normal Form)

Even when a database is in 3rd Normal Form, still there would be anomalies resulted if it has more than one Candidate Key.

v. Fourth Normal Form

A table is said to be in the Fourth Normal Form when,

- It is in the Boyce-Codd Normal Form.
- And, it doesn't have Multi-Valued Dependency.

vi. Fifth Normal Form

A table is in 5th Normal Form only if it is in 4NF and it cannot be decomposed into any number of smaller tables without loss of data.