DATA SCIENCE COURSE TUTORIAL # 05

Chapter # 2: Understanding Data Types and Formats

2.1 Structured vs Unstructured Data

In data science, data can come in different formats. Understanding the difference between **structured**, **unstructured**, and **semi-structured** data is important for selecting the right storage, processing, and analysis methods.

1: Structured Data

Structured data is **organized in a fixed format** (such as rows and columns in a table). It can be stored in **databases, spreadsheets, or tabular formats**.
Because it follows a clear structure, it is **easy to store, search, and analyze** using tools like **SQL, Excel, or Pandas**.

Examples of structured data:

- A student database with name, roll number, and grades.
- Sales records with date, product name, and quantity.

Key characteristics of structured data:

- Organized in a predefined schema.
- Easy to query with structured query languages.
- Usually stored in relational databases.

2: Unstructured Data

Unstructured data does **not follow a predefined format or structure**. It cannot be easily arranged into rows and columns like structured data.
It includes **text, images, audio, video**, and other formats that are more complex to store and analyze.

Examples of unstructured data:

- Social media posts.
- Email content.
- Customer reviews.
- Photos and videos.

Key characteristics of unstructured data:

- No fixed schema or structure.
- Difficult to store and query in traditional databases.
- Requires advanced processing techniques such as natural language processing (NLP) or image recognition.

3: Semi-Structured Data

Semi-structured data is **not fully organized like structured data** but still contains tags or markers to separate elements.
It lies between structured and unstructured data, making it easier to analyze compared to completely unstructured formats.

Examples of semi-structured data:

- JSON files.
- XML files.

Key characteristics of semi-structured data:

- Flexible structure with defined markers or tags.
- Easier to store and parse than unstructured data.
- Often used for web data, APIs, and configuration files.