Chapter 1

Introduction to Types of Digital Data

Classification of Digital Data:-

- Irrespective of the size of the enterprise whether it is big or small, data continues to be a precious and irreplaceable asset.
- Data is present in homogeneous sources as well as in heterogeneous sources. The need of the hour is to understand, manage, process, and take the data for analysis to draw valuableinsights. Digital data can be structured, semi-structured or unstructured data.
- Data generates information and from information we can draw valuable insight. As represented in below Figure, digital data can be broadly classified into structured, semi-structured, and unstructured data.

Classification of Digital Data

1) **Unstructured data:-** This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters. researches, white papers, body of an email, etc.
2) **Semi-structured data:-** Semi-structured data is also referred to as self-describing structure. This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer

program. About 10% data of an organization is in this format; for example, HTML, XML, JSON, email data etc.

3) **Structured data:-** When data follows a pre-defined schema/structure we say it is structured data. This is the data which is in an organized form (e.g., in rows and columns) and be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. About 10% data of an organization is in this format. Data stored in databases is an example of structured data.

Structured Data :-

- This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program.
- Relationships exist between entities of data, such as classes and their objects.
- Data stored in databases is an example of structured data.
- If our data is highly structured, one can look at leveraging any of the available RDBMS such as [Oracle Corp. — Oracle, IBM — DB2, Microsoft — Microsoft SQL Server, EMC — Greenplum, Teradata — Teradata, MySQL (open source), PostgreSQL (advanced open source) etc.] to house it.
- These databases are typically used to hold transaction/operational data generated and collected by day-to-day business activities. In other words, the data of the On-Line Transaction Processing (OLTP) systems are generally quite structured.

Sources of Structured Data :-

- Below figure shows the different sources of structured data .

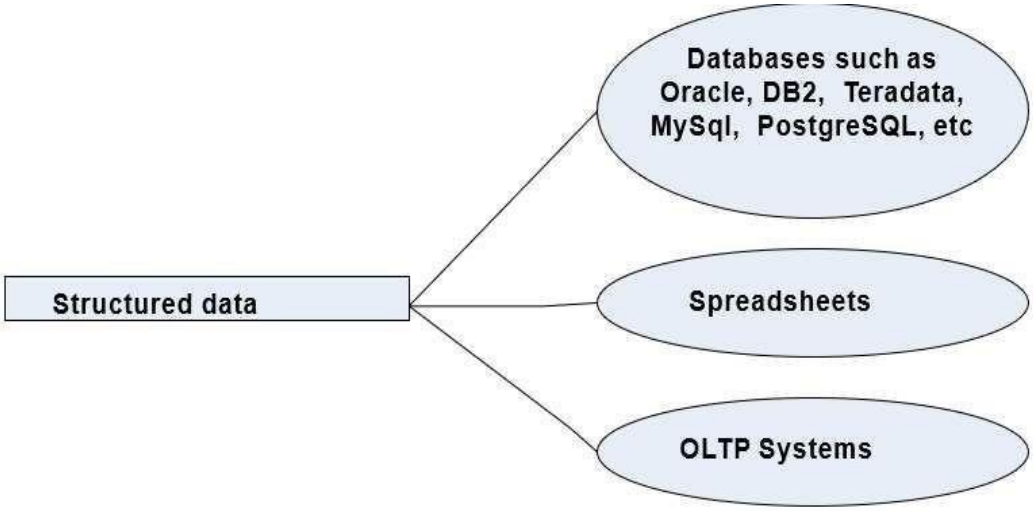


Figure:Sources of Structureddata

Ease with Structured Data:-

- Structure data provide the ease of working .The ease is with respect to the following:
  1) **Insert/update/delete:** The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.
  2) **Security:** How does one ensure the security of information? There are available check encryption and tokenization solutions to warrant the security of information throughout its lifecycle. Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.
  3) **Indexing:-** Indexing is a way to optimize the performance of a database by minimizing the number of disk accesses required when a query is processed. It is a data structure technique which is used to quickly locate and access the data in a database. It speeds up the data retrieval operations (primarily the SELECT DML statement) at the cost of additional writes and storage space, but the benefits that ensue in search operation are worth the additional writes and storage space.

4) **Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server (increasing the primary and secondary or peripheral storage capacity, processing capacity of the processor, etc.).

5) **Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction.

   a) **Atomicity:-** It mean that either the entire transaction takes place at once or doesn't happen at all. There is no midway i.e. transactions do not occur partially. Each transaction is considered as one unit and either runs to completion or is not executed at all. It involves the following two operations.

      - ✓ Abort: If a transaction aborts, changes made to database are not visible.
      - ✓ Commit: If a transaction commits, changes made are visible.

      Atomicity is also known as the 'All or nothing rule'.

   b) **Consistency :-** This means that integrity constraints must be maintained so that the database is consistent before and after the transaction. It refers to the correctness of a database.

      The database moves from one consistent state to another consistent state . In other words , if the same piece of information is stored at two or more places , they are in complete agreement

   c) **Isolation :-** This property ensures that multiple transactions can occur concurrently without leading to the inconsistency of database state.

      - Transactions occur independently without interference. Changes occurring in a particular transaction will not be visible to any other transaction until that particular change in that transaction is written to memory or has been committed.
      - This property ensures that the execution of transactions concurrently will result in a state that is equivalent to a state achieved these were executed serially in some order'.

   d) **Durability:-** Al changes made to database during a transaction are permanent and that accounts for the durability of the transaction .

- This property ensures that once the transaction has completed execution, the updates and modifications to the database are stored in and written to disk and they persist even if a system failure occurs.
- These updates now become permanent and are stored in non-volatile memory. The effects of the transaction, thus, are never lost.

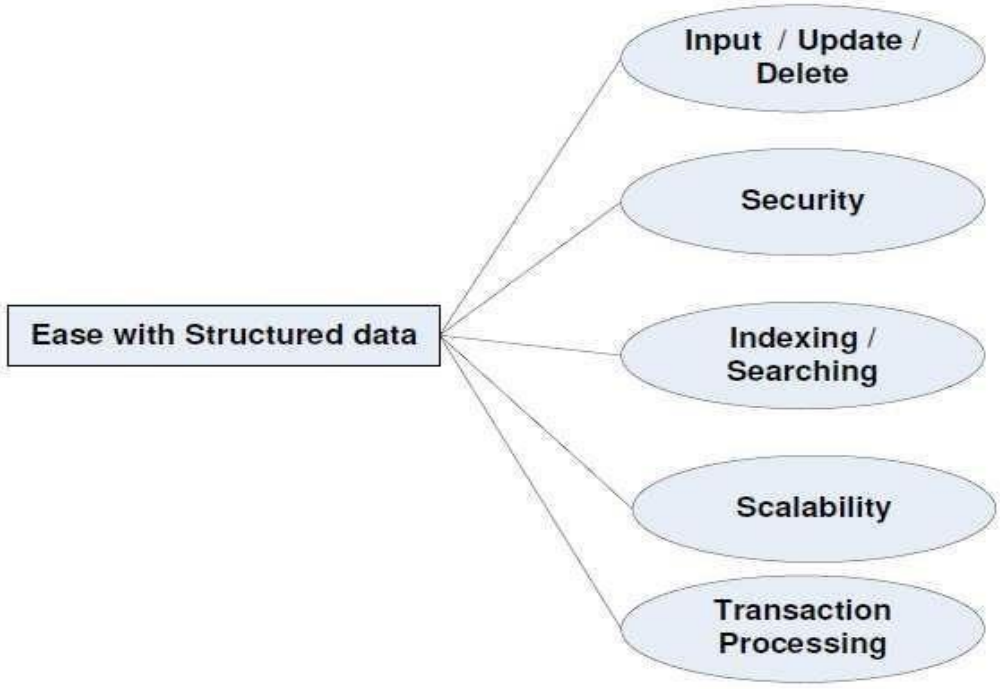


Figure : Ease of Working with structure Data

Semi-Structured data:-

- This is the data which does not conform to a data model but has some structure.
- However, it is not in a form which can be used easily by a computer program.
- Example, emails, XML, markup languages like HTML, etc. Metadata for this data is

It has the following features:

1) It does not conform to the data models that one typically associates with relational databases or any other form of data tables.
2) It uses tags to segregate semantic elements.

3) Tags are also used to enforce hierarchies of records and fields within data.
4) There is no separation between the data and the schema. The amount of structure used is dictated by the purpose at hand.
5) In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes. And if at all, they have the same set of attributes, the order of attributes may not be similar and for all practical purposes it is not important as well.

Sources of Semi-structured Data :-

- Amongst the sources for semi-structured data, the front runners are XML and JSON
- **XML:** eXtensible Markup Language (XML) is hugely popularized by web services developed utilizing the Simple Object Access Protocol (SOAP) principles.
- **JSON:** Java Script Object Notation (JSON) is used to transmit data between a server and a web application.
- JSON is popularized by web services developed utilizing the Representational State Transfer (REST) - an architecture style for creating scalable web services.
- MongoDB (open-source, distributed, NoSQL, documented oriented database) and Couchbase (originally known as Membase, open-source, distributed, NoSQL, document-oriented database) store data natively in JSON format.

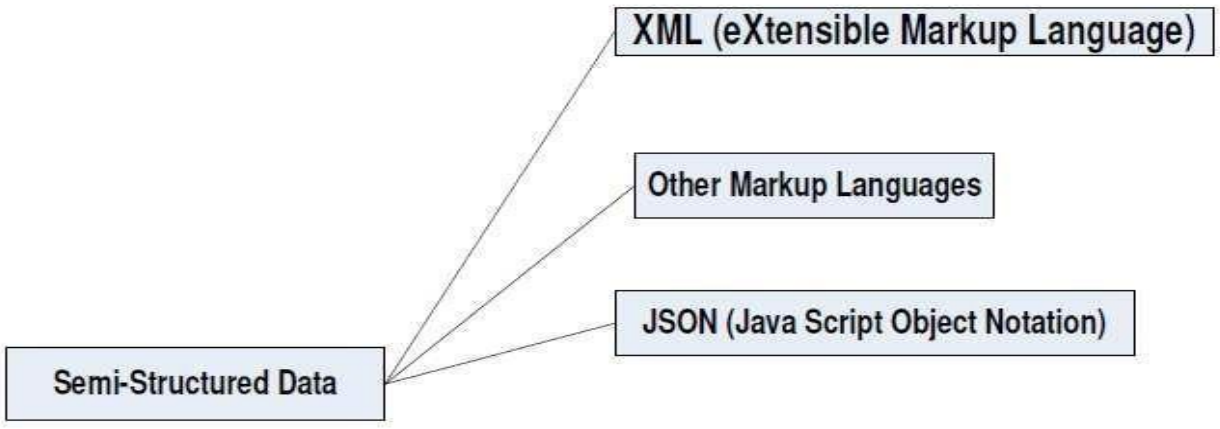


Figure: Sources of Semi-structure Data

Example of HTML is as below

An example of HTML is as follows:

```
<HTML>
    <HEAD>
    <TITLE>Place your title here</TITLE>
    </HEAD>
<BODY BGCOLOR="FFFFFF">
    <CENTER><IMG SRC="clouds.jpg" ALIGN="BOTTOM"x/CENTER>
    <HR>          <a href="http://bigdatauniversity.com">Link Name</a>
    <HI>this is a Header</HI>
    <H2>this is a sub Header</H2>
    Send me mail at <a href="mailto:support@yourcompany.com"> support@yourcompany.com</a>.
    <P>a new paragraph!
    <PxB>a new paragraph!</B>
    <BRxBxI>this is a new sentence without a paragraph break, in bold italics.</Ix/B>
    <HR>
</BODY>
</HTML>
```

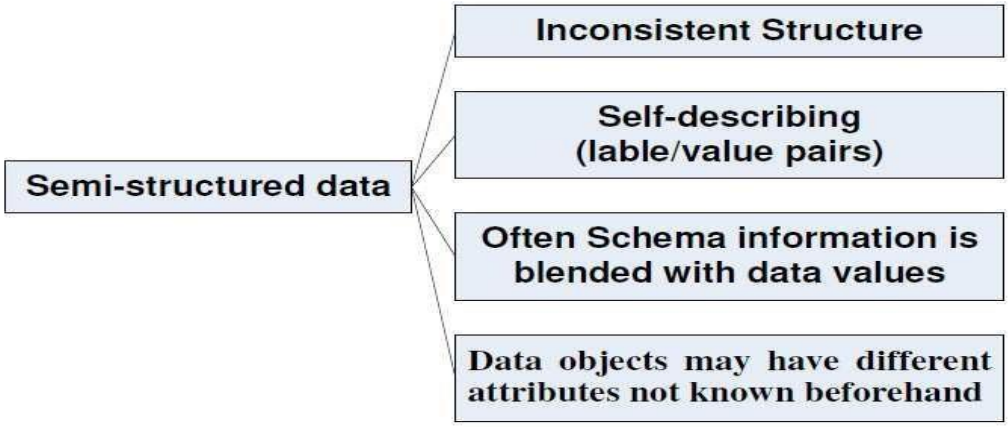
Example of JSON document

Sample JSON document

```
{
_id:9,
BookTitle: "Fundamentals of Business Analytics",
AuthorName: "Seema Acharya",
Publisher: "Wiley India",
YearofPublication: "2011"
}
```

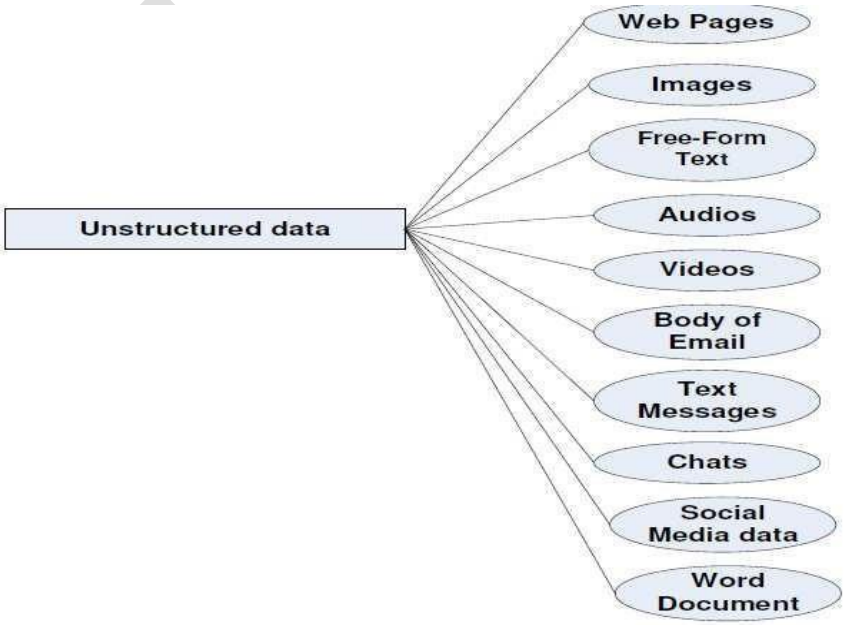
Characteristics of semi structure Data :

- Following figure will illustrate what are the characteristics of Semi Structure data .

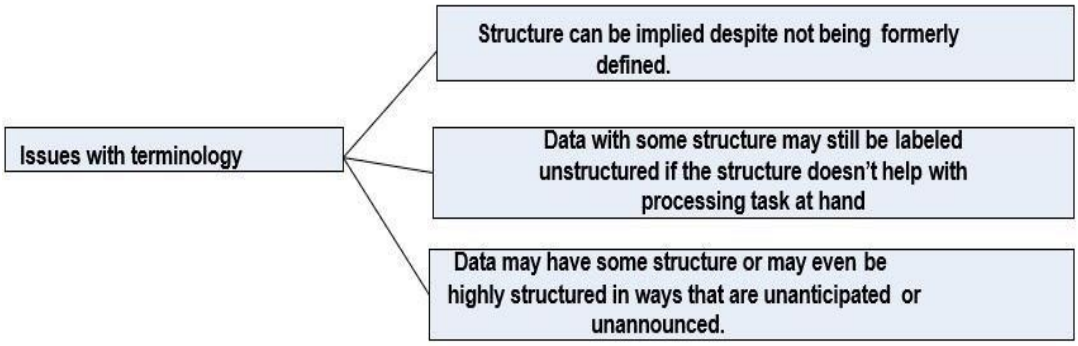


Unstructured Data :-

- This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.
- About 80% data of an organization is in this format.
- Example: memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc. which is illustrated in following figure .



Issues with terminology Unstructured Data :-



- Although unstructured data is known NOT to conform a pre-defined data model or to be organised in a pre-defined manner , there are incidents wherein the structure of the data (placed in the unstructured category) can be implied .
- Above figure mention some reasons behind placing data in the unstructured category despite it having some structure or being highly structured.

How to Deal with Unstructured Data?

- The following techniques are used to find pattern in or interpret unstructured data

1) Data Mining:-

- ✓ First, we deal with large data sets.
- ✓ Second, we use methods at the intersection of artificial intelligence, machine learning, statistics, and database systems to unearth consistent patterns in large data sets and/or systematic relationships between variables.
- ✓ It is the analysis step of the knowledge discovery in databases process**.**
- ✓ Few popular data mining algorithm are as follow .

a) Association rule mining:

- It is also called "market basket analysis" or "affinity analysis".
- It is used to determine ―What goes with what?
- It is about when you buy a product, what is the other product that you are likely to purchase with it.
- For example, if you pick up bread from the grocery, are you likely to pick eggs or cheese to go with it.

b) Regression analysis:

- It helps to predict the relationship between two variables.
- The variable whose value needs to be predicted is called the dependent variable and the variables which are used to predict the value are referred to as the independent variables.

c) Collaborative filtering:

- It is about predicting a users preference or preferences based on the preferences of a group of users. For example, take a look at Table below.

	Learning using Audio	Learning using Video	Textual Learners
User 1	Yes	Yes	No
User 2	Yes	Yes	Yes
User 3	Yes	Yes	No
User 4	Yes	?	?

- We are looking at predicting whether User 4 will prefer to learn using videos or is a textual leaner depending on one or a couple of his or her known preferences.
- We analyze the preferences of similar user profiles and on the basis of it, predict that User 4 will also like to learn using videos and is not a textual learner.

2) Text Analytics or Text Mining:-

- ✓ Compared to the structured data stored in relational databases, text largely unstructured, amorphous, and difficult to deal with algorithmically.
- ✓ Text mining is the process of gleaning high quality and meaningful information (through devising of patterns and trends by means of statistical pattern learning) from text.

- ✓ It includes tasks such as text categorization, text clustering, sentiment analysis, concept/entity extraction, etc.

3) **Natural language processing (NLP):-** It is related to the area of human computer interaction. It about enabling computers to understand human or natural language input.
4) **Noisy text analytics:-** It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message-boards, text messages, etc. The noisy unstructured data usually comprises one or more of the following: Spelling mistakes, abbreviations, acronyms, non- standard words, missing punctuation, missing letter case, filler words such as " Uh " , "Um" etc
5) **Manual tagging with metadata:-** This is about tagging manually with adequate metadata to provide the requisite semantics to understand unstructured data.
6) **Part-of-speech tagging:-** It is also called POS or POST or grammatical tagging. It is the process reading text and tagging each word in the sentence as belonging to a particular part of speech such as "noun", "verb", "adjective" , etc.
7) **Unstructured Information Management Architecture (UIMA):-** It is an open source platform from IBM. It is used for real-time content analytics. It is about processing text and other unstructured to find latent meaning and relevant relationship