There are three types of data : structured , unstructured and semi-structured data.
But in this article I'll just shine a light on semi-structured data.

What is Semi-structured data?

Semi-structured data is a type of data that is not purely structured, but also not completely unstructured. It contains some level of organization or structure, but does not conform to a rigid schema or data model, and may contain elements that are not easily categorized or classified.

1. Semi-structured data is typically characterized by the use of metadata or tags that provide additional information about the data elements. For example, an XML document might contain tags that indicate the structure of the document, but may also contain additional tags that provide metadata about the content, such as author, date, or keywords.

2. Other examples of semi-structured data include JSON, which is commonly used for exchanging data between web applications, and log files, which often contain a mix of structured and unstructured data.

Semi-structured data is becoming increasingly common as organizations collect and process more data from a variety of sources, including social media, IoT devices, and other unstructured sources. While semi-structured data can be more challenging to work with than strictly structured data, it offers greater flexibility and adaptability, making it a valuable tool for data analysis and management.

Semi-structured data is data that does not conform to a data model but has some structure. It lacks a fixed or rigid schema. It is the data that does not reside in a relational database but that has some organizational properties that make it easier to analyze. With some processes, we can store them in the relational database.

Characteristics of semi-structured Data:

- Data does not conform to a data model but has some structure.
- Data can not be stored in the form of rows and columns as in Databases
- Semi-structured data contains tags and elements (Metadata) which is used to group data and describe how the data is stored
- Similar entities are grouped together and organized in a hierarchy
- Entities in the same group may or may not have the same attributes or properties
- Does not contain sufficient metadata which makes automation and management of data difficult
- Size and type of the same attributes in a group may differ
- Due to lack of a well-defined structure, it can not used by computer programs easily

Sources of semi-structured Data:

- E-mails
- XML and other markup languages
- Binary executables
- TCP/IP packets
- Zipped files
- Integration of data from different sources
- Web pages

Advantages of Semi-structured Data:

- The data is not constrained by a fixed schema
- Flexible i.e Schema can be easily changed.
- Data is portable
- It is possible to view structured data as semi-structured data
- Its supports users who can not express their need in SQL
- It can deal easily with the heterogeneity of sources.

- Flexibility: Semi-structured data provides more flexibility in terms of data storage and management, as it can accommodate data that does not fit into a strict, predefined schema. This makes it easier to incorporate new types of data into an existing database or data processing pipeline.
- Scalability: Semi-structured data is particularly well-suited for managing large volumes of data, as it can be stored and processed using distributed computing systems, such as Hadoop or Spark, which can scale to handle massive amounts of data.
- Faster data processing: Semi-structured data can be processed more quickly than traditional structured data, as it can be indexed and queried in a more flexible way. This makes it easier to retrieve specific subsets of data for analysis and reporting.
- Improved data integration: Semi-structured data can be more easily integrated with other types of data, such as unstructured data, making it easier to combine and analyze data from multiple sources.
- Richer data analysis: Semi-structured data often contains more contextual information than traditional structured data, such as metadata or tags. This can provide additional insights and context that can improve the accuracy and relevance of data analysis.

Overall, semi-structured data provides a number of advantages over traditional structured data, particularly when it comes to managing and analyzing large volumes of data that do not fit neatly into predefined data models.

Disadvantages of Semi-structured data

- Lack of fixed, rigid schema make it difficult in storage of the data
- Interpreting the relationship between data is difficult as there is no separation of the schema and the data.
- Queries are less efficient as compared to structured data.
- Complexity: Semi-structured data can be more complex to manage and process than structured data, as it may contain a wide variety of formats, tags, and metadata. This can make it more difficult to develop and maintain data models and processing pipelines.
- Lack of standardization: Semi-structured data often lacks the standardization and consistency of structured data, which can make it more difficult to ensure data quality and accuracy. This can also make it harder to compare and analyze data across different sources.
- Reduced performance: Processing semi-structured data can be more resource-intensive than processing structured data, as it often requires more complex parsing and indexing operations. This can lead to reduced performance and longer processing times.
- Limited tooling: While there are many tools and technologies available for working with structured data, there are fewer options for working with semi-structured data. This can make it more challenging to find the right tools and technologies for a particular use case.
- Data security: Semi-structured data can be more difficult to secure than structured data, as it may contain sensitive information in unstructured or less-visible parts of the data. This can make it more challenging to identify and protect sensitive information from unauthorized access.

Overall, while semi-structured data offers many advantages in terms of flexibility and scalability, it also presents some challenges and limitations that need to be carefully considered when designing and implementing data processing and analysis pipelines.

Problems faced in storing semi-structured data

- Data usually has an irregular and partial structure. Some sources have implicit structure of data, which makes it difficult to interpret the relationship between data.
- Schema and data are usually tightly coupled i.e they are not only linked together but are also dependent on each other. Same query may update both schema and data with the schema being updated frequently.
- Distinction between schema and data is very uncertain or unclear. This complicates the designing of structure of data
- Storage cost is high as compared to structured data

Possible solution for storing semi-structured data

- Data can be stored in DBMS specially designed to store semi-structured data
- XML is widely used to store and exchange semi-structured data. It allows its user to define tags and attributes to store the data in hierarchical form.
  Schema and Data are not tightly coupled in XML.
- Object Exchange Model (OEM) can be used to store and exchange semi-structured data. OEM structures data in the form of graphs.
- RDBMS can be used to store the data by mapping the data to relational schema and then mapping it to a table

Extracting information from semi-structured Data

Semi-structured data have different structure because of heterogeneity of the sources. Sometimes they do not contain any structure at all. This makes it difficult to tag and index. So extracting information from them is a tough job. Here are possible solutions :

- Graph based models (e.g OEM) can be used to index semi-structured data
- Data modeling technique in OEM allows the data to be stored in graph based models. The data in a graph based model is easier to search and index.
- XML allows data to be arranged in hierarchical order which enables the data to be indexed and searched
- Use of various data mining tools