CAPTER – 2

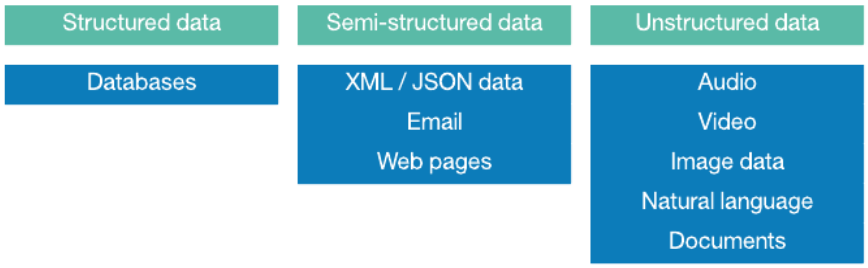
Types Of Data Based on Structure

Learning Topics

- ✓ Structured data
- ✓ Unstructured data
- ✓ Semi structured data

In data science, the data will be divided into three parts based on its structure. Will learn those three types in this chapter.

Structured Data:



- Structured data is usually stored with well-defined schemas such as Databases. It is generally in tabular format with column and rows that clearly define its attributes.
- It has consistent order and can be easily accessed and used by a person or a computer program.
- SQL (Structured Query language) is often used to manage structured data stored in databases.

Characteristics Of Structured Data:

- Data has easily identifiable structure.
- Data is stored in the form of rows and columns.
- Data is well organized so, Definition, Format and Meaning of data is explicitly known
- Easy to access and query, so data can be easily used by other programs
- Easy to analyze and process

Sources Of Structured Data:

- SQL Databases
- Spreadsheets such as Excel

Advantages Of Structured Data:

- It has a well-defined structure that helps in easy storage and access of data.
- Each data record has index number and column name which helps in easy access.
- Data mining is easy i.e., knowledge can be easily extracted from data.
- Operations such as Updating and deleting is easy due to well-structured form of data.
- Business Intelligence operations such as Data warehousing can be easily undertaken.
- Easily scalable in case there is an increment of data.
- Ensuring security to data is easy

Note: Only 20% of structured data is available in world remain data is not in structured. This is the main challenge of every data scientist.

Unstructured Data:

Unstructured data is the data which does not identifiable structure such that it cannot be used by a computer program easily and not easily understandable. Unstructured data is can't organize in a pre-defined manner.

Characteristics:

- Data has any structure.
- Data cannot be stored in the form of rows and columns as in Databases.
- Data does not follow any semantic or rules
- Data don't have any particular format or sequence
- Very hard to identify the its structure.
- Due to lack of identifiable structure, it cannot use by computer programs easily

Sources of Unstructured Data:

- Images (JPEG, GIF, PNG, etc.), Videos
- Reports, Memos
- Word documents and PowerPoint presentations
- Web pages, Surveys

Advantages:

- We can give a required format or sequence to data.
- The data is not constrained by a fixed schema.
- Very Flexible due to absence of schema.
- Data is portable.
- It is very scalable.
- It can deal easily with the heterogeneity of sources.
- These types of data have a variety of business intelligence and analytics applications.

Disadvantages:

- It is difficult to store and manage due to lack of schema and structure.
- Indexing the data is difficult and not having pre-defined attributes.

- Search results are not very accurate.
- Ensuring security to data is difficult task.

Problems Faced In Storing Unstructured Data:

- It requires a lot of storage space.
- It is difficult to store videos, images, audios, etc.
- Due to unclear structure, operations like update, delete and search is very difficult.
- Storage cost is high as compared to structured data.
- Very difficult to indexing.

Possible Solution for Storing Unstructured Data:

- Unstructured data can be converted to easily manageable formats.
- Content addressable storage system (CAS) is used to store unstructured data.
- It stores data based on their metadata and a unique name is assigned to every object stored in it. The object is retrieved based on content not its location.
- Unstructured data can be stored in XML format.
- Unstructured data can be stored in RDBMS which supports BLOBs

Extracting Information from Unstructured Data:

Unstructured data do not have any structure. So, it can not easily be interpreted by conventional algorithms. It is also difficult to tag and index unstructured data. So, extracting information from them is tough job. Here are possible solutions:

- Taxonomies or classification of data helps in organizing data in hierarchical structure. Which will make search process easy.
- Data can be stored in virtual repository and be automatically tagged. For example, Documentum.
- Use of application platforms like XOLAP.
- XOLAP helps in extracting information from e-mails and XML based documents
- Use of various data mining tools

Semi-structured data:

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 rational database but that have some organizational properties that make it easier to analyze. With some processes, we can store them in the relational database.

Characteristics:

- 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
- JSON, XML, YML
- Binary executables
- TCP/IP packets
- Zipped files
- Integration of data from different sources
- Web pages

Advantages:

- 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 cannot express their need in SQL
- It can deal easily with the heterogeneity of sources.

Disadvantages:

- 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.

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 of 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 form of graph.
- 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, while extract information from them is tough job. Here are possible solutions –

- Graph based models (e.g. OEM) can be used to index semi-structured data
- Data modelling technique in OEM allows the data to be stored in graph-based model. The data in 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.