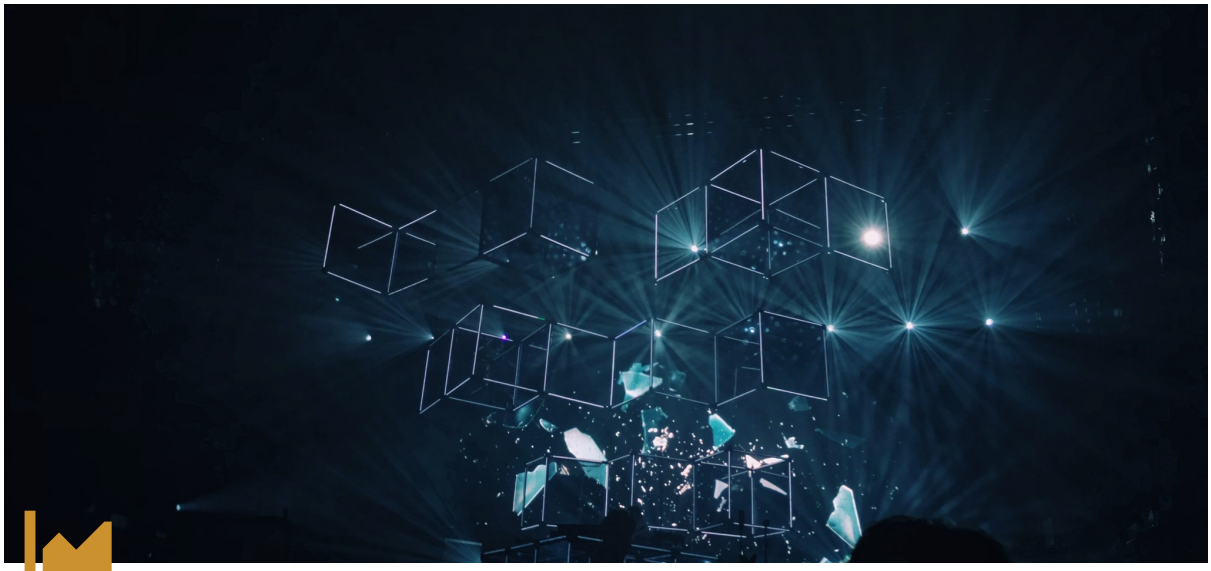


DATA and DATA EVERYWHERE

What is Data?

It is the collection of facts, such as numbers, words, measurements, observations or even just description of things.

What is Metadata?

Information about data(data about data) this is called metainformation or meta data. - this is semi structured data.

What is String?

Sequence of character

Structured & UnStructured Data

What is Structured data?

Structured data - typically categorized as quantitative data

It is highly organized and easily interpretable by machine learning algorithm.

Examples of structured data include dates, names, addresses, credit card numbers, etc. Their benefits are tied to ease of use and access, while liabilities revolve around

data inflexibility:

PROS

- Easily used by machine learning (ML) algorithms: The specific and organized architecture of structured data eases manipulation and querying of ML data.
- Easily used by business users: Structured data does not require an in-depth understanding of different types of data and how they function. With a basic understanding of the topic relative to the data, users can easily access and interpret the data.
- Accessible by more tools: Since structured data predates unstructured data, there are more tools available for using and analyzing structured data.

CONS

- Limited usage: Data with a predefined structure can only be used for its intended purpose, which limits its flexibility and usability.
- Limited storage options.

Structured data tools

- SQLite, MySQL, PostgreSQL

Use Cases for structured data

- Online Booking
- Accounting.

What is Unstructured data?

Unstructured data - typically categorized as qualitative data, cannot be processed and analyzed via conventional data tools and methods.

It is managed in non-relational databases.

Another way to manage unstructured data is to use data lakes to preserve it in raw form.

95% of businesses prioritize unstructured data management.

Examples of unstructured data include text, mobile activity, social media posts, Internet of Things (IoT) sensor data, etc. Their benefits involve advantages in format, speed and storage, while liabilities revolve around expertise and available resources:

PROS

- Fast Accumulation rates: Since there is no need to predefine the data, it can be collected quickly and easily
- Native format: It is stored in its native format, remains undefined until needed. Its adaptability increased file formats in the database, which widens the data pool and enables data scientists to prepare and analyze only the data they need.

CONS

- Requires expertise
- Specialized tools

Unstructured data tools

- MongoDB, Hadoop, Azure

Use cases for Unstructured data

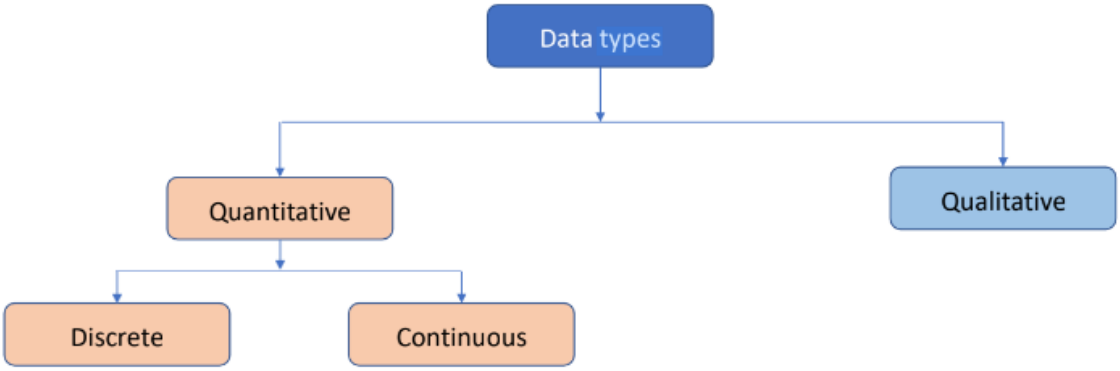
- Data Mining
- Chatbots

What is Semi-structured data?

It is the “bridge” between structured and unstructured data. It does not have a predefined data model and is more complex than structured data, yet easier to store than unstructured data.

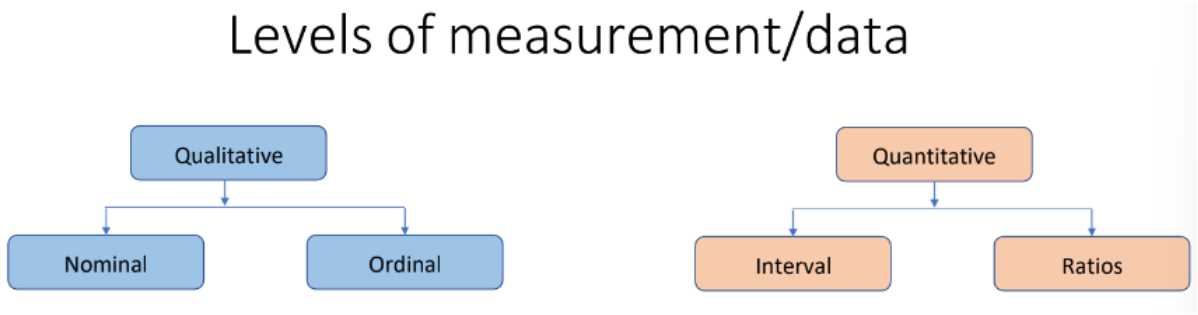
Semi-structured data uses “metadata” (e.g., tags and semantic markers) to identify specific data characteristics and scale data into records and preset fields. Metadata ultimately enables semi-structured data to be better cataloged, searched and analyzed than unstructured data.
• Example of metadata usage: An online article displays a headline, a snippet, a featured image, image alt-text, slug, etc., which helps differentiate one piece of web content from similar pieces.

FOUR TYPES OF DATA



- Qualitative deals with the quality(characteristics)
- Quantitative deals with numbers.
    - Discrete: It is countable and Finite. This type of data can't be measured but it can be counted.
        - Let me give you an example of countable, if you are asked to count number of cars on
          road during certain time of the day, it can take numbers like 100,125 or 1000 but never
          100.5 or 125.72 etc.
        - Likewise, example of finite is, the number which comes on rolling a dice. There are only 6 possible choices like 1,2,3...6 but never more than 6 or 4.5 etc. likewise if you flip a coin it has only heads or tails, so, there are certain number of values you can pick from.
    - Continuous: These are uncountable or infinite. Usually it is a measurement of something and cannot be counted.
        - For example, a person's height or weight. Height can be 5.23, 5.24, 5.76 etc. similarly
          weight can be any value like 75.82 Kgs or 62.35 kgs etc. Height, weight, length, speeds,
          temperatures etc. are examples of continuous data.

Levels of Data



NOMINAL

It means name only.

Nominal scales are used for labelling variables, without any quantitative value

It is a categorical data which has no order. Red car, blue car, black car etc.

For statistical analysis we can assign numbers to these categories for example: Red=1, White=2 and Black=3, these numerical values assigned does not have any mathematical significance.

A sub-type of nominal scale with only two categories (eg. male/female, hot/cold, good/bad) is called `dichotomous`

ORDINAL

It is a categorical data which has an implied order. Like size of clothing as Small, Medium, Large or Xtra Large.

Likert scale questions having a scale from 1 to 5 for example Agree=1, neutral=2 disagree=3 etc. The numbers do not have any mathematical significance, but they are the labels.

Eg. Ranking (ranking in school exams i.e 1st, 2nd and 3rd etc., ranking in the Army i.e Captain, Major, Colonel etc.), survey done on Likertscale, in this example there is an order associated with it.

INTERVAL

These values are categorical and ordered data in addition to that they have scale to them.

Interval values data don't have a `true zero`. **True zero** means absence of the variable

For eg. zero degrees temprature does not mean there is no temprature, it just means, it is too cold.

Here addition and substraction are significant, but division and multiplication are not.

Eg. Time and temprature

RATIOS

These values are categorical, ordered data having scale to them and they have a natural or true zero.

Eg. height, weight, length etc.

Since they have natural zero it allows for a wide range of both descriptive and inferential statistics to be applied.

Eg. Zero dollars means there is no money.

They can be discrete or continuous.