Introduction to Data

By Younus Chilakala

Data vs Information

Data:

Data is raw, unanalyzed, unorganized, unrelated, uninterrupted facts that are used to derive information, after analysis.

Information:

Information is acquired when data is analyzed, structured, and given composure or context to make it useful.

How is data useful to us?

Data is essential because it is what provides us with knowledge and insights about almost every single thing in this world. We are living in a world of data and data is all around us.

Good data provides indisputable evidence, while anecdotal evidence, assumptions, or abstract observation might lead to wasted resources due to taking action based on an incorrect conclusion.

Data empowers us to take informed and insightful decisions that drive our lives and organizations forward.

General classification of Data

Structured data:

Structured data is data that has been predefined and formatted to a set structure before being placed in data storage.

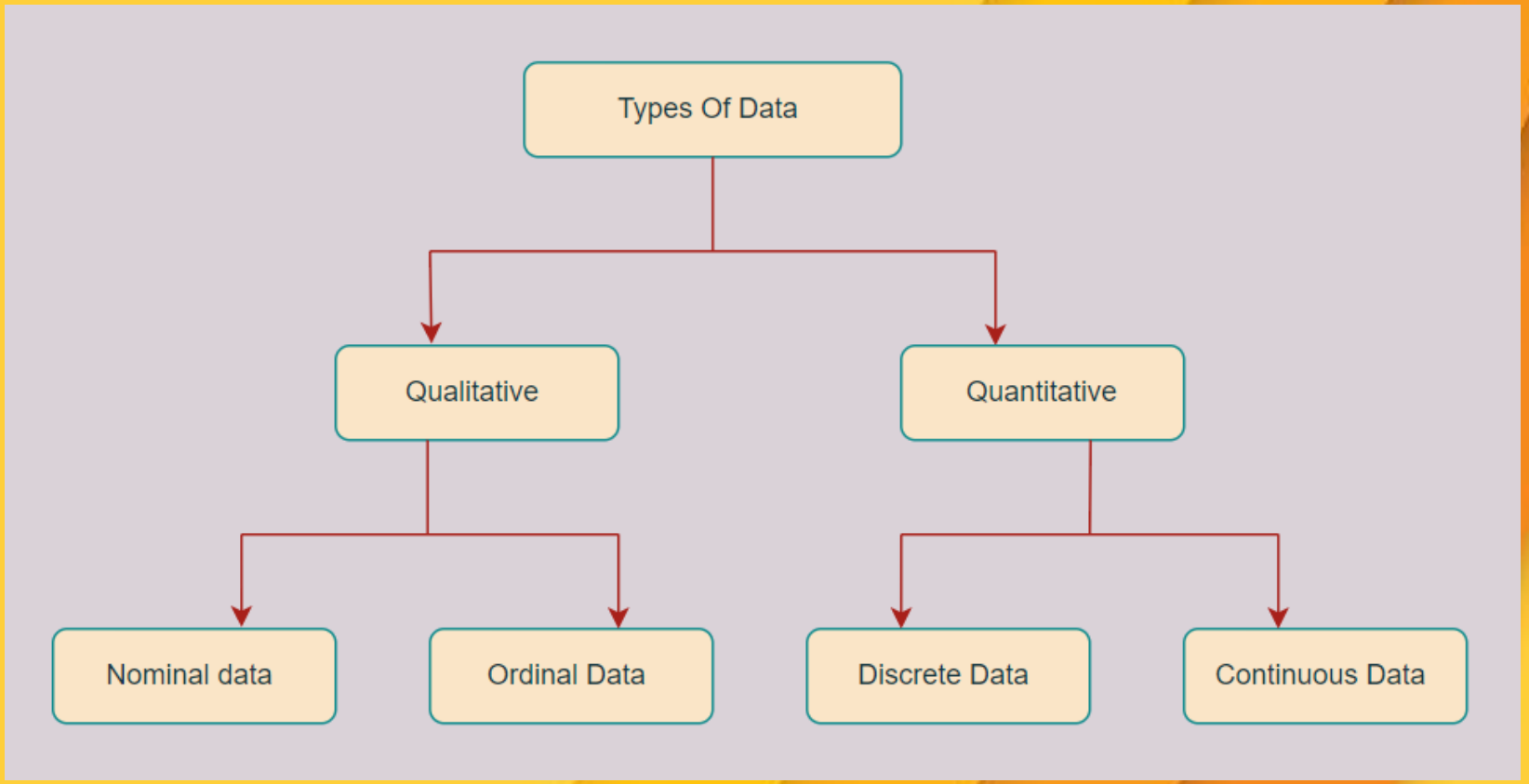
Unstructured data:

Unstructured data is data stored in its native format and not processed until it is used. It cannot be presented in a data model or schema.

Semi-structured data:

Semi-structured data refers to data that does not follow the format of a tabular data model or relational databases because it does not have a fixed schema. However, the data is not completely raw or unstructured, and does contain some structural elements

Types of Data



Qualitative vs Quantitative data

Qualitative data:

Qualitative data is non-statistical and is typically unstructured or semi-structured. This data isn't necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers.

Quantitative data: Quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Types of Data(continued)

Nominal Data:

Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can’t be compared with another color.

Ordinal Data:

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can’t do any arithmetical tasks on them.

Types of Data(continued)

Discrete data:

Data that can only take on certain values are discrete data. These values do not have to be complete numbers, but they are values that are fixed. It only contains finite values, the subdivision of which is not possible. The data can not be split further into decimals or fractions.

Continuous data:

Continuous data is the data that can be of any value. Over time, some continuous data can change. It may take any numeric value, within a potential value range of finite or infinite. The data can be split further into decimals or fractions.

Big Data

According to Oracle, Big data can be described as follows-

The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them. But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.

Currently, there is so much big data that International Data Corporation (IDC) predicts the “Global Datasphere” will grow from 33 Zettabytes (ZB) in 2018 to 175 ZB in 2025.

The three V's of Big Data

Volume:

Big data is about volume. Volumes of data that can reach unprecedented heights. It is not uncommon for large companies to have Terabytes – and even Petabytes – of data in storage devices and on servers.

Velocity:

Velocity is the rate at which data is received and (perhaps) acted on. It essentially measures how fast the data is coming in. Some data will come in in real-time, whereas other will come in fits and starts, sent to us in batches.

Variety:

Variety refers to the many types of data that are available. With the rise of big data, data comes in new unstructured data types. Unstructured and semi-structured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

Tools used in Big Data

Below are some of the Big Data tools and the technologies that they are used for-

Data storage:

Apache Hadoop, MongoDB

Data Analytics:

Apache Spark, Splunk

Data mining:

Presto, RapidMiner

Data visualization:

Tableau, Looker

THANK YOU!