Big Data

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

So It is massive amount of data which cannot be stored, processed and analyzed using traditional tools like RDBMS.

Sources of Big Data

These data come from many sources like

- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

Why is need of Big Data Analytics?

To extract meaningful insights from big data such as hidden patterns, unknown correlations, market trends and customer preferences.

Big Data Characteristics (5 V's):

There are five v's of Big Data that explains the characteristics.

5 V's of Big Data

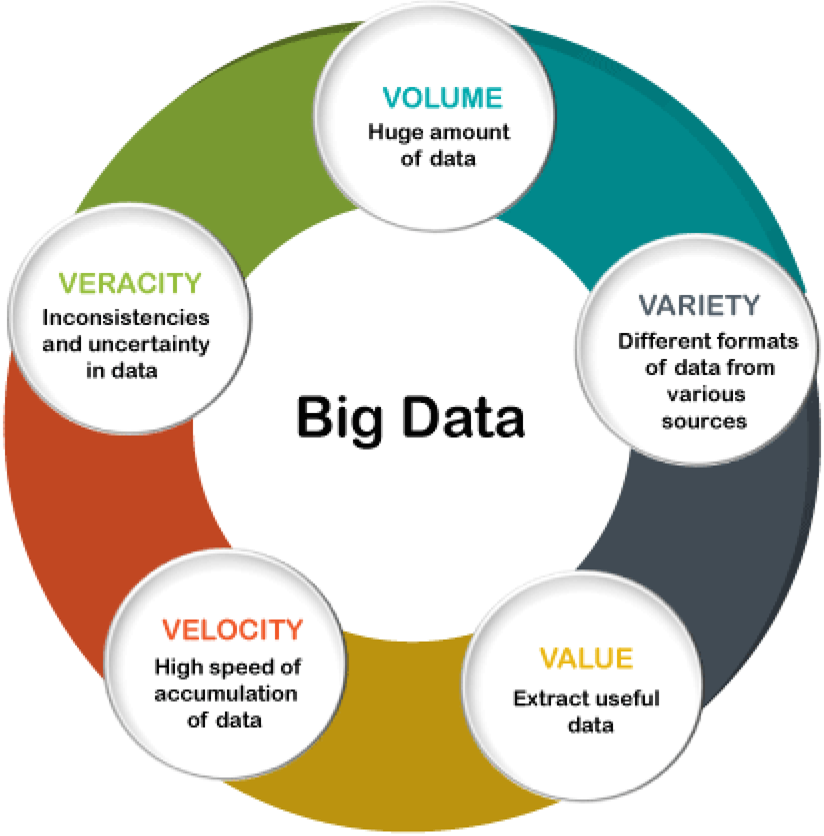
Volume

Veracity

Variety

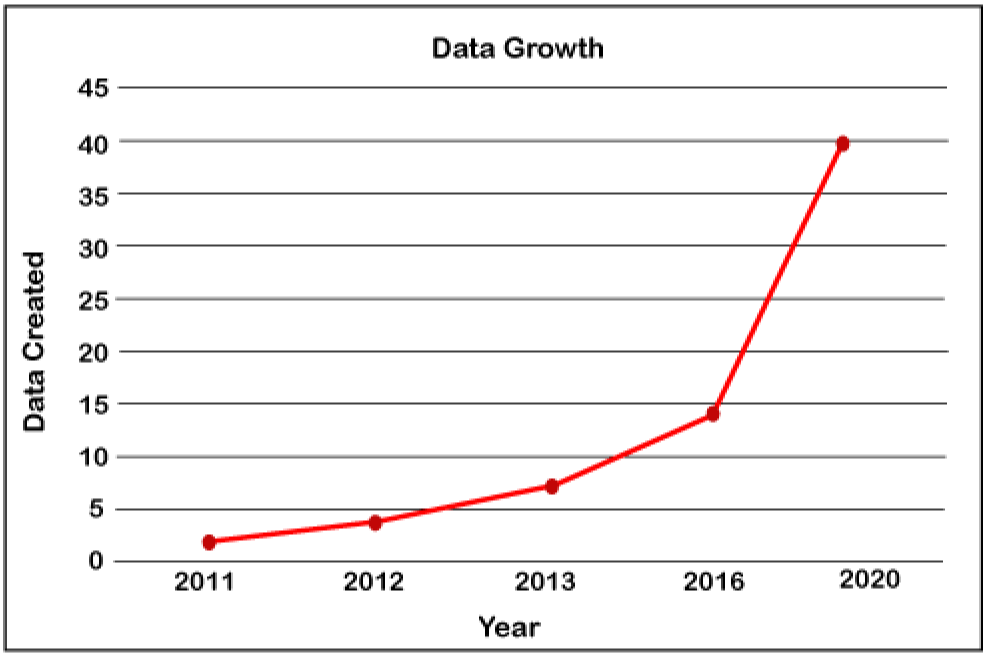
Value

Velocity


VOLUME
Huge amount of data
VARIETY
Different formats of data from various sources
VALUE
Extract useful data
VELOCITY
High speed of accumulation of data
VERACITY
Inconsistencies and uncertainty in data
Big Data

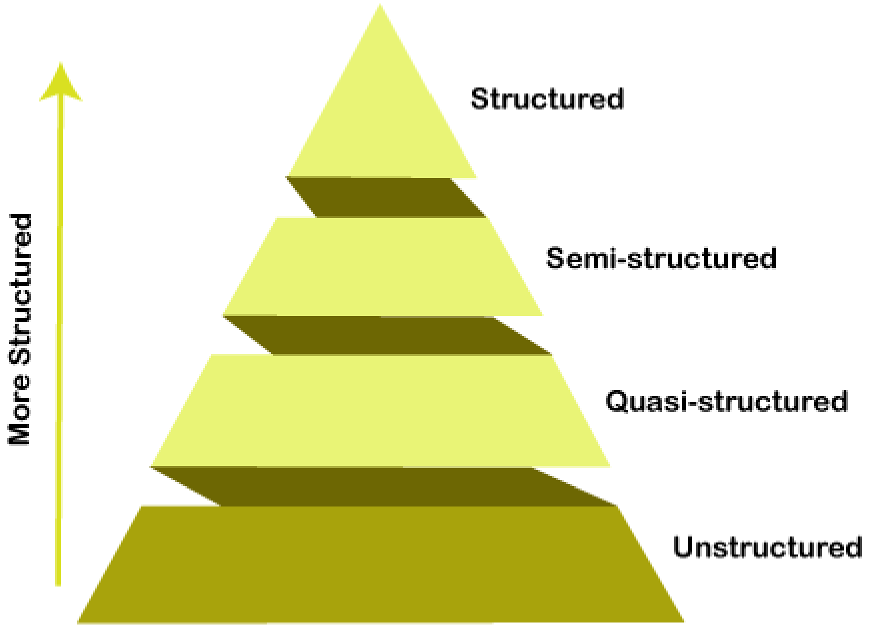
Volume

The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more. **Facebook** can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.



Variety

Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos,** etc.



The data is categorized as below:

a. **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

b. **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.

c. **Unstructured Data**: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.

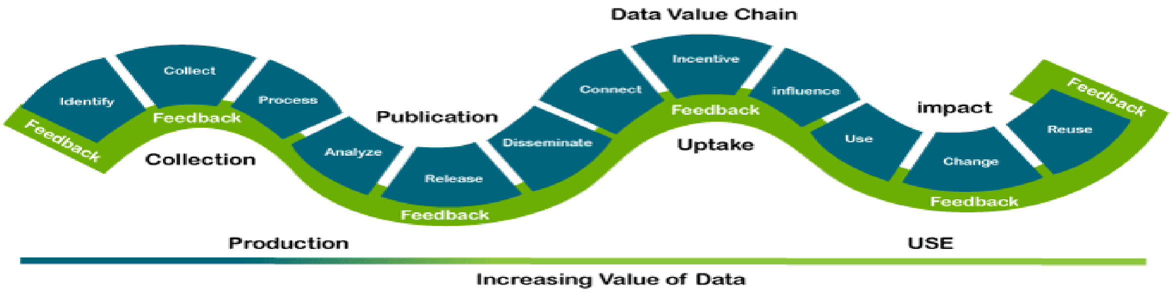
Example: Web server logs, i.e., the log file is created and maintained by some server that contains a list of **activities**.

Veracity

It refers to the quality and accuracy of data. Gathered data could have missing pieces, may be inaccurate or may not be able to provide real, valuable insight. Veracity, overall, refers to the level of trust there is in the collected data.

For example, **Facebook posts** with hashtags.

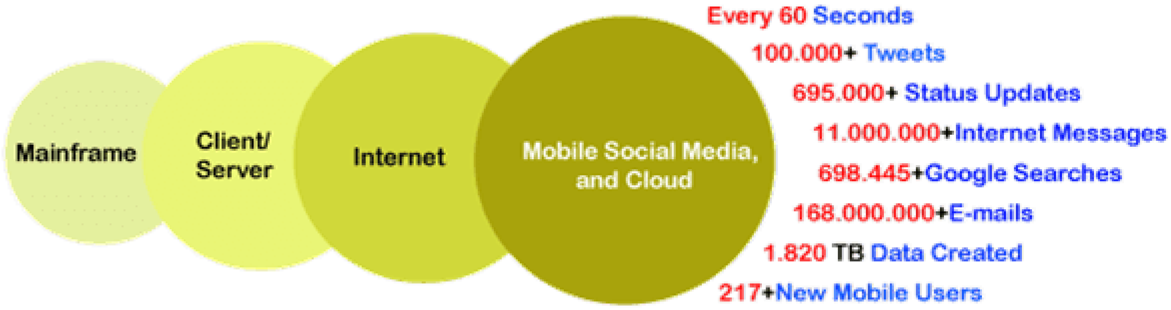
Value: This refers to the value that big data can provide, and it relates directly to what organizations can do with that collected data.



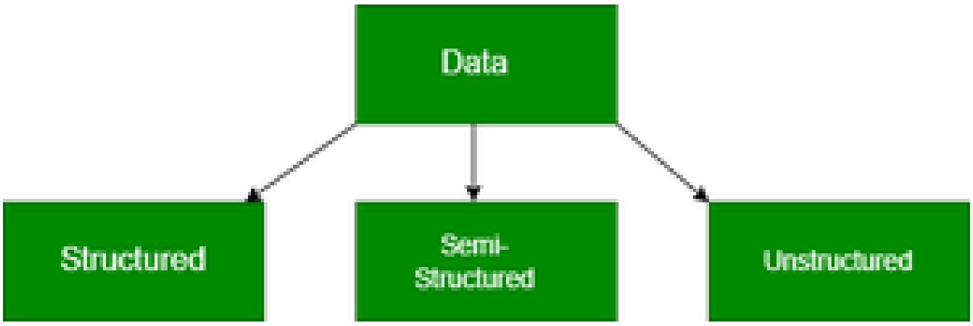
Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in **real-time**. It contains the linking of incoming **data sets speeds, rate of change**, and **activity bursts**. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like **application logs, business processes, networks, and social media sites, sensors, mobile devices,** etc.

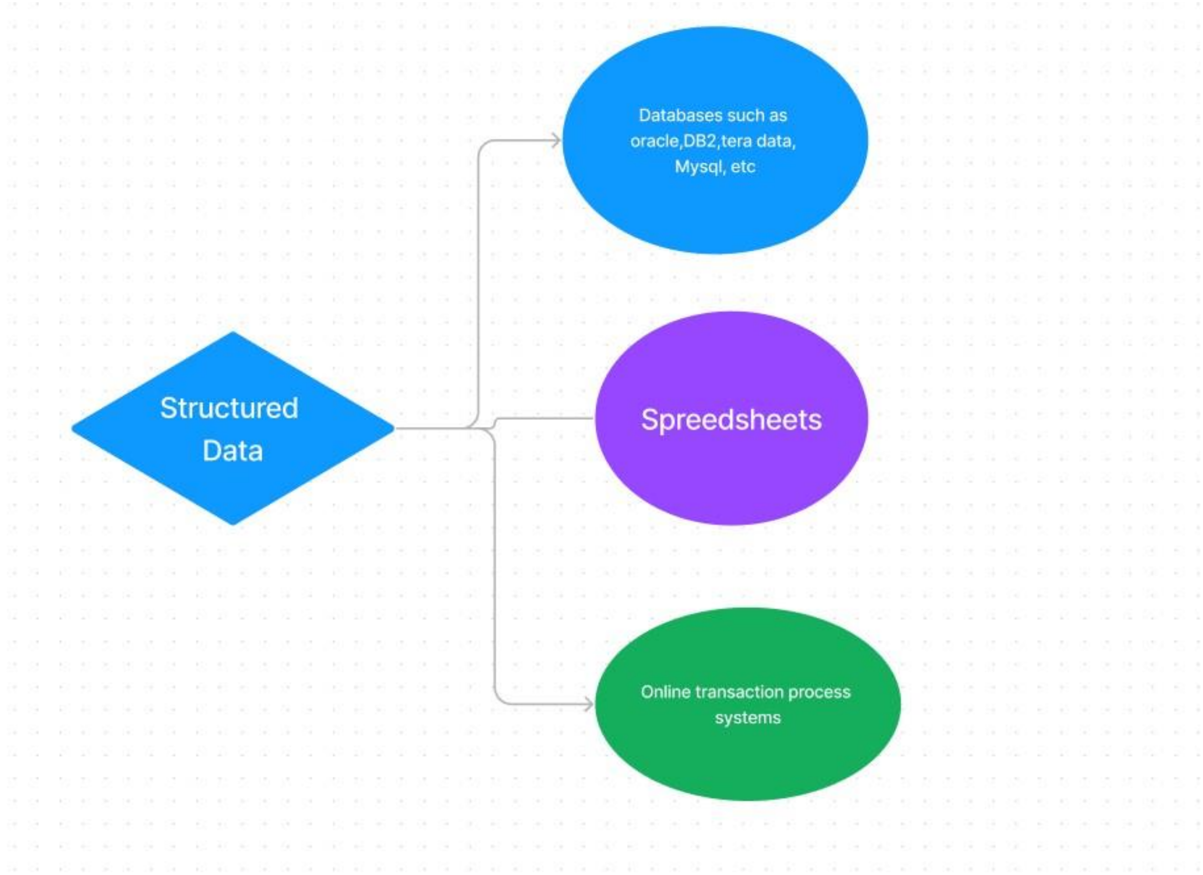


Types of Big Data



Structured Data

- Structured data can be defined as the data that resides in a fixed field within a record.
- It is type of data most familiar to our everyday lives. for ex: birthday, address
- A certain schema binds it, so all the data has the same set of properties. Structured data is also called relational data. It is split into multiple tables to enhance the integrity of the data by creating a single record to depict an entity. Relationships are enforced by the application of table constraints.
- The business value of structured data lies within how well an organization can utilize its existing systems and processes for analysis purposes.


Structured Data
Databases such as oracle,DB2,tera data, Mysql, etc
Spreedsheets
Online transaction process systems

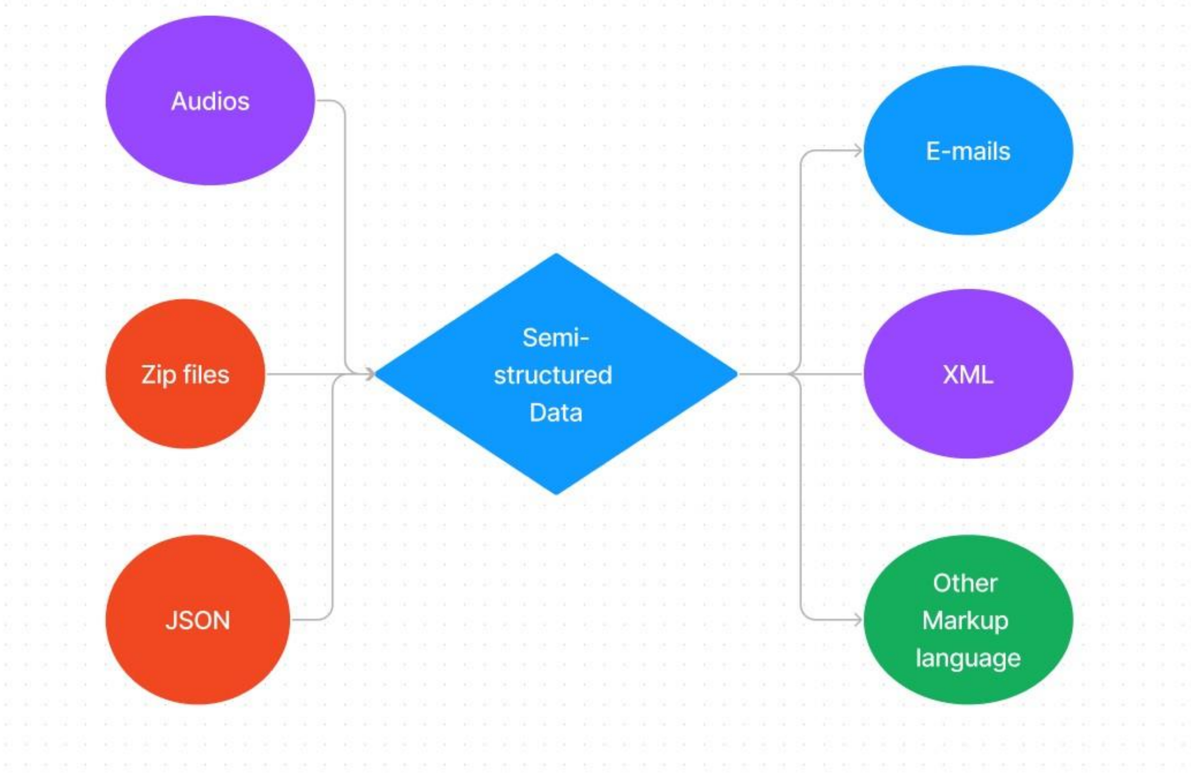
Sources of structured data

A *Structured Query Language (SQL)* is needed to bring the data together. Structured data is easy to enter, query, and analyze. All of the data follows the same format. However, forcing a consistent structure also means that any alteration of data is too tough as each record has to be updated to adhere to the new structure. *Examples* of structured data include numbers, dates, strings, etc. The business data of an e-commerce website can be considered to be structured data.

Semi-Structured Data

- Semi-structured data is not bound by any rigid schema for data storage and handling. The data is not in the relational format and is not neatly organized into rows and columns like that in a spreadsheet. However, there are some features like key-value pairs that help in discerning the different entities from each other.

- Since semi-structured data doesn't need a structured query language, it is commonly called *NoSQL data*.
- A data serialization language is used to exchange semi-structured data across systems that may even have varied underlying infrastructure.
- Semi-structured content is often used to store metadata about a business process but it can also include files containing machine instructions for computer programs.
- This type of information typically comes from external sources such as social media platforms or other web-based data feeds.



Semi-Structured Data

Data is created in plain text so that different text-editing tools can be used to draw valuable insights.

1. XML– XML stands for *eXtensible Markup Language*. It is a text-based markup language designed to store and transport data. XML parsers can be found in almost all popular development platforms. It is human and machinereadable. XML has definite standards for schema, transformation, and display. It is self-descriptive. Below is an example of a programmer's details in XML.

- XML

```
<ProgrammerDetails>
    <FirstName>Jane</FirstName>
    <LastName>Doe</LastName>
    <CodingPlatforms>
        <CodingPlatform Type="Fav">GeeksforGeeks</CodingPlatform>
<CodingPlatform Type="2ndFav">Code4Eva!</CodingPlatform>
        <CodingPlatform Type="3rdFav">CodeisLife</CodingPlatform>
   </CodingPlatforms>
</ProgrammerDetails>
```

XML expresses the data using *tags (*text within angular brackets*)* to shape

the data (for ex: FirstName) and *attributes* (For ex: Type) to feature the data. **2. JSON**– JSON (JavaScript Object Notation) is a lightweight open-standard file format for data interchange. JSON is easy to use and uses human/machine-readable text to store and transmit data objects.

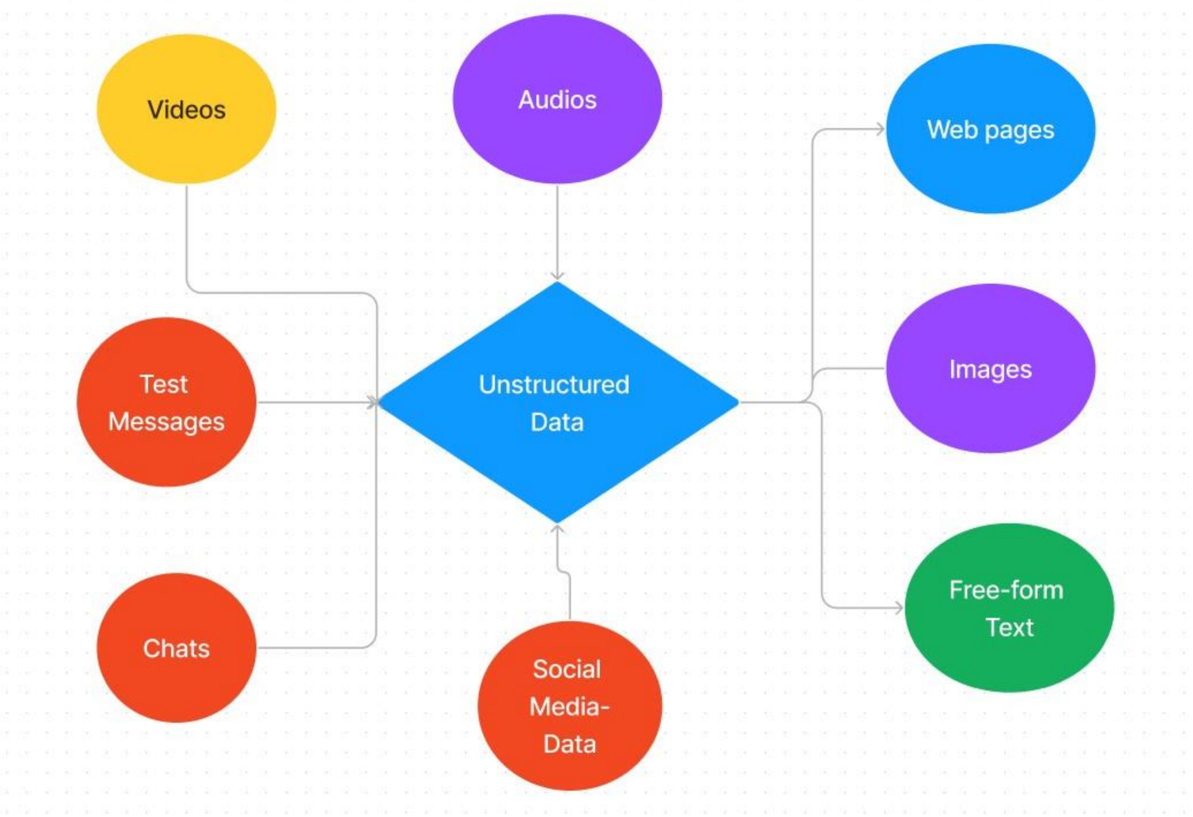
- Javascript

```
{
    "firstName": "Jane",
"lastName": "Doe",
    "codingPlatforms": [
        { "type": "Fav", "value": "Geeksforgeeks" },
        { "type": "2ndFav", "value": "Code4Eva!" },
{ "type": "3rdFav", "value": "CodeisLife" }      ]
}
```

This format isn't as formal as XML. It's more like a key/value pair model than a formal data depiction. Javascript has inbuilt support for JSON.

Unstructured Data

- Unstructured data is the kind of data that doesn't adhere to any definite schema or set of rules. Its arrangement is unplanned and haphazard.
- Photos, videos, text documents, and log files can be generally considered unstructured data. Even though the metadata accompanying an image or a video may be semi-structured, the actual data being dealt with is unstructured.
- Additionally, Unstructured data is also known as "dark data" because it cannot be analyzed without the proper software tools.



Un-structured Data