**What is Unstructured Data?**

* Last Updated : 10 Oct, 2021

**Unstructured data** is the data which does not conforms to a data model and has no easily identifiable structure such that it can not be used by a computer program easily. Unstructured data is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database.

**Characteristics of Unstructured Data:**

* Data neither conforms to a data model nor has any structure.
* Data can not be stored in the form of rows and columns as in Databases
* Data does not follows any semantic or rules
* Data lacks any particular format or sequence
* Data has no easily identifiable structure
* Due to lack of identifiable structure, it can not used by computer programs easily

**Sources of Unstructured Data:**

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

**Advantages of Unstructured Data:**

* Its supports the data which lacks a proper format or sequence
* 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 type of data have a variety of business intelligence and analytics applications.

**Disadvantages Of Unstructured data:**

* It is difficult to store and manage unstructured data due to lack of schema and structure
* Indexing the data is difficult and error prone due to unclear structure and not having pre-defined attributes. Due to which 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 to store unstructured data.
* 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](https://www.geeksforgeeks.org/structured-data/)
* Indexing the unstructured data is difficult

**Possible solution for storing Unstructured data:**

* Unstructured data can be converted to easily manageable formats
* using Content addressable storage system (CAS) 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 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 organising 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

To read Differences between Structured, Semi-structured and Unstructured data refer the following article:

EXAMEL

## **What are some examples of unstructured data?**

Unstructured data can be created by people or generated by machines.

Here are some examples of the human-generated variety:

* Email: Email message fields are unstructured and cannot be parsed by traditional analytics tools. That said, email metadata affords it some structure, and explains why email is sometimes considered semi-structured data.
* Text files: This category includes word processing documents, spreadsheets, presentations, email, and log files.
* Social media and websites: Data from social networks like Twitter, LinkedIn, and Facebook, and websites such as Instagram, photo-sharing sites, and YouTube.
* Mobile and communications data: Text messages, phone recordings, collaboration software, Chat, and Instant Messaging.
* Media: Digital photos, audio, and video files.

Here are some examples of unstructured data generated by machines:

* Scientific data: Oil and gas surveys, space exploration, seismic imagery, and atmospheric data.
* Digital surveillance: Reconnaissance photos and videos.
* Satellite imagery: Weather data, land forms, and military movements.

…………………………………………………………………………………………………

Unstructured data is information that is not arranged according to a pre-set data model or schema, and therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia are two common types of unstructured content. Many business documents are unstructured, as are email messages, videos, photos, webpages, and audio files.

## **Unstructured data**

From 80 to 90 percent of data generated and collected by organizations, is unstructured,, and its volumes are growing rapidly — many times faster than the rate of growth for structured databases.

Unstructured data stores contain a wealth of information that can be used to guide business decisions. However, unstructured data has historically been very difficult to analyze. With the help of AI and machine learning, new software tools are emerging that can search through vast quantities of it to uncover beneficial and actionable business intelligence.

## **Unstructured data vs. structured data**

Let’s take structured data first: It’s usually stored in a relational database or RDBMS, and is sometimes referred to as relational data. It can be easily mapped into designated fields — for example, for zip codes, phone numbers, and credit cards, respectively. Data that conforms to RDBMS structure is easy to search, both with human-defined queries and with software.

Unstructured data, in contrast, doesn’t fit into these sorts of pre-defined data models. It can’t be stored in an RDBMS. And because it comes in so many formats, it’s a real challenge for conventional software to ingest, process, and analyze. Simple content searches can be undertaken across textual unstructured data with the right tools .

Beyond that, the lack of consistent internal structure doesn’t conform to what typical data mining systems can work with. As a result, companies have largely been unable to tap into value-laden data like customer interactions, rich media, and social network conversations. Robust tools for doing so are only now being developed and commercialized.

## **What is structured, semi-structured, and unstructured data?**

As we’ve already seen, structured data is organized in ways that make for easy searching. Unstructured data — comprising most other types — exists in formats such as audio, video, and social media postings, and is not easy for conventional tools to search.

The contrasting of one type “versus” the other should not be thought of as a conflict. You simply choose one or the other based on the applications you’re interested in. Relational databases handle structured data, and just about all other kinds of systems can house unstructured data.

Common RDBMS applications using structured data include airline reservation systems, inventory control, sales transactions, and ATM activity. Typical unstructured use cases are media viewing and editing tools, presentation software, and word processing.

There is also a third category called semi-structured data. While not stored in relational databases, this type of information has some organizing properties, making it easier to parse and analyze. Specifically, semi-structured data contains internal tags and markings that allow for grouping and hierarchies.

Email is a common semi-structured data application. While detailed email analysis requires sophisticated tools, its native metadata allows for basic classification and keyword searches. Semi-structured data is only a 5% to10% slice of the total enterprise data pie, but it has some critical use cases. Examples include the XML markup language, the versatile JSON data-interchange format, and databases of the NoSQL or non-relational variety. These last are a good choice for storing information such as text with variable lengths. The most widely-used non-relational database, MongoDB, accommodates semi-structured documents by natively storing them in the JSON format.

## **How is unstructured data structured?**

Unstructured types of data can actually have internal structural elements. They’re considered “unstructured” because their information doesn’t lend itself to the kind of table formatting required by a relational database. As noted earlier, unstructured data can be textual or non-textual (such as audio, video, and images), and generated by people or by machines. Non-relational databases such as MongoDB are the preferred choice for storing many kinds of unstructured data.

## **What is unstructured data used for?**

Simple content searches can be performed on textual unstructured data. Traditional analytics tools are optimized for highly structured relational data, so they’re of little use for unstructured sources such as rich media, customer interactions, and social media data.

Big Data and unstructured data often go together: [IDC](https://www.pcmag.com/news/90-percent-of-the-big-data-we-generate-is-an-unstructured-mess) estimates that 90% of these extremely large datasets are unstructured. New tools have recently become available to analyze these and other unstructured sources. Powered by AI and machine learning, such platforms function at near real-time speed and educate themselves based on the patterns and insights they uncover. These systems are being employed against large unstructured datasets to enable never-before-possible applications like:

* Analyzing communications for regulatory compliance
* Tracking and analyzing customer social media conversations and interactions
* Gaining reliable insights into widespread customer behavior and preferences

## **How is unstructured data stored?**

Unstructured data can be stored in a number of ways: in applications, [NoSQL (non-relational) databases](https://www.mongodb.com/nosql-explained), data lakes, and data warehouses. Platforms like [MongoDB Atlas](https://www.mongodb.com/cloud/atlas) are especially well suited for housing, managing, and using unstructured data.