MODULE 2

Types of Data

You will learn about the 'Types of Data' in this module.

Module Learning Objectives

At the end of this module, you will be able to:

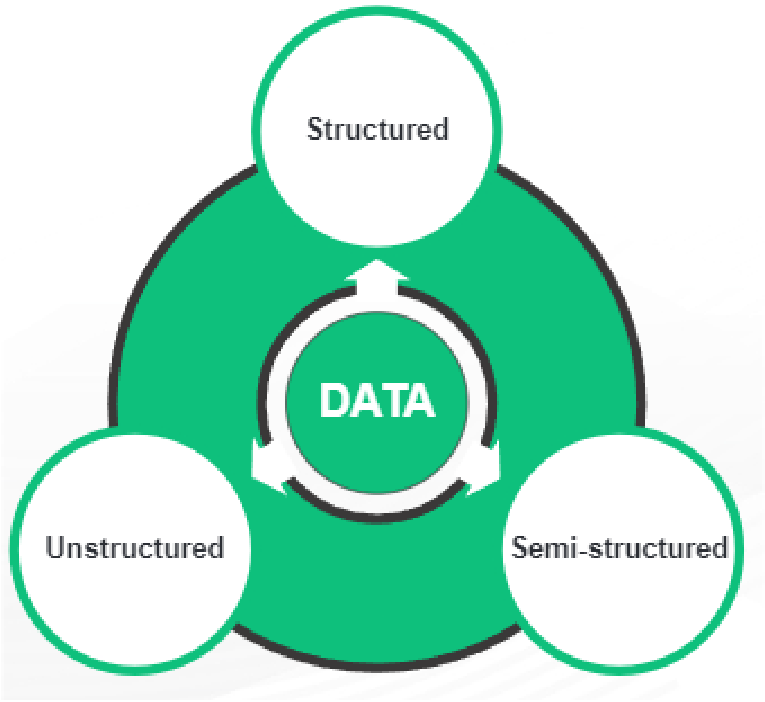
- List the different types of data.
- Define Structured Data and its organization.
- Define Unstructured Data.
- Define Semi-structured Data.
- Compare and contrast the different features of Structured, Unstructured and Semi-structured Data.
- Understand the concept of noise from a data viewpoint.

Module Topics

The following topics that will be covered in the module:

1. Structured Data
2. Semi-structured Data
3. Unstructured Data
4. Comparison of Structured, Unstructured and Semi-structured Data
5. Presence of Noise in Data.

1 Types of Data



In the previous module, you learnt about Data, its definition and sources of Data. In this module, you're going to learn about the different types of Data. There are three major types of Data.

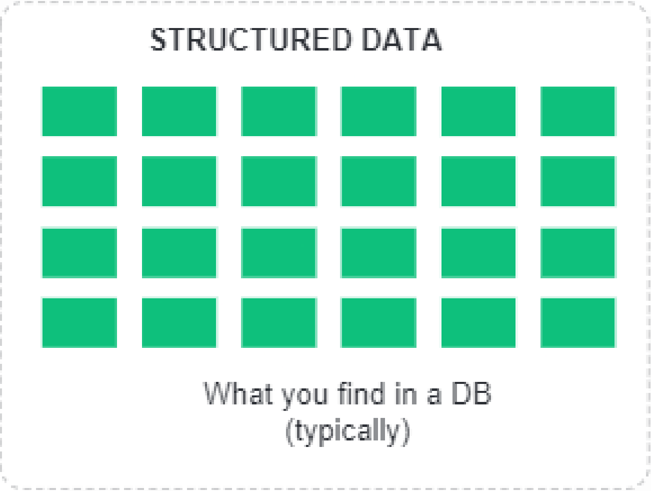
- **Structured Data:** Information stored relational databases belong to this category. Data has a strict structure and the schema has to be predefined before populating data in a relational database. It is the Database Management System (DBMS) that checks and ensures that the populated data follows the structure as specified in the schema.
- **Semi-structured Data:** Data that is collected in an ad-hoc manner before the storage and management methods are known. This data is portable and does not require the schema to predefined, though schema definition is not impossible. Data might have some structure, but all information will not have an identical structure. This type of data is also referred to as self-describing data.
- **Unstructured data:** Data that doesn't have any fixed structure. There is very less indication of the type of data. There is no predefined data model for this type of data. Since there is no predefined schema for storing unstructured data, complex queries are needed to search and retrieve information from unstructured data.

In the upcoming sections, we will learn about each of these data types in detail along with examples, their advantages and disadvantages.

1.1 Structured Data

Structured Data:

- Is highly organized and readily searchable by queries or algorithms.
- Can be quickly consolidated into facts.
- Follows a predefined schema.
- Usually resides in fixed fields.
- A typical example is a Relational Database Management System (RDBMS).
- A Schema is defined before the content is created and the data is populated.



Structured Data is a simple and highly organized type of Data. Data is written in a defined format, that is easy for the machines to understand and is easily searchable even by the basic algorithms. Structured Data is usually present in relational databases (RDBMS) and is retrieved using queries and algorithms using data-type and the field names.

Structured Data can be entered, stored, queried and analyzed using simple and efficient ways. Field-name and type of data should strictly be defined before populating in an RDBMS. Hence, Structured Data may face restrictions in terms of the number of characters or specific terminologies. Simple or complex queries in Excel spreadsheets or Structured Query Language (SQL) are used to retrieve information from a relational database.

1.2 Organization of Structured Data

The following diagram and tables illustrates the structured data.

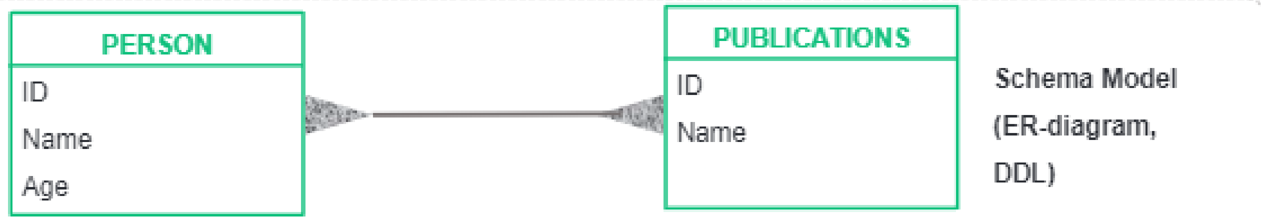


Table: Person		
ID	Name	Age
1	Rolf	28
2	Steffi	22

1	1
1	2
2	2

Table: Publications	
ID	Name
1	Social Tagging
2	Classification

Data table

Sample table in a relational database system

Structured Data usually resides in Relational databases or Data Warehouses. The Structured Data contains usually text files, displayed in titled columns and rows which can easily be queried and processed by Data Mining tools.

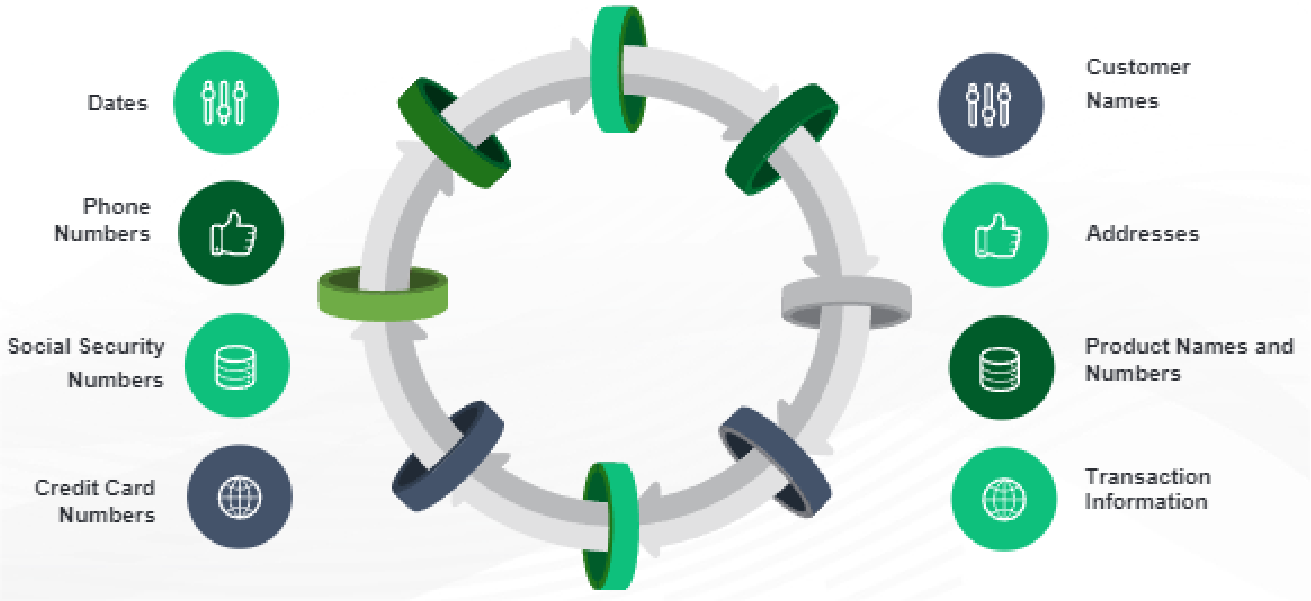
Let's take the example of a Relational Database System to understand the organization of structured data. The figure above gives an idea of an Entity-Relationship (ER) diagram and its concrete tables within a Relational Database Management System (RDBMS).

In a Relational Database System, the data is stored in tables with predefined columns. The Column title describes the type of information stored in a table. Designing a database schema is an elaborate process. Schema is the one which defines the type and structure of the data and its relations. In a relational database, the schema is defined well before content is created and the data is populated. The well-defined schema of a fully structured data helps the in-efficient data processing and improved storage and navigation of content.

In a Relational Database, extending previously-defined schemas that already have content, might be difficult. For example, to extend a single table row with a new attribute, another table column must be created. This will not be suitable for tables with thousands of other rows that do not need another attribute.

1.3 Examples of Structured Data

Some of the examples of Structured data:



You may note down a few examples of Structured Data.

1.4 Applications of Structured Data

Relational database applications with Structured Data:

- Airline reservation systems
- Inventory control
- Sales transactions
- ATM activity

Some of the applications where data is structured and well-organized are:

- Airline reservation systems
- Inventory control
- Sales transactions
- ATM activity

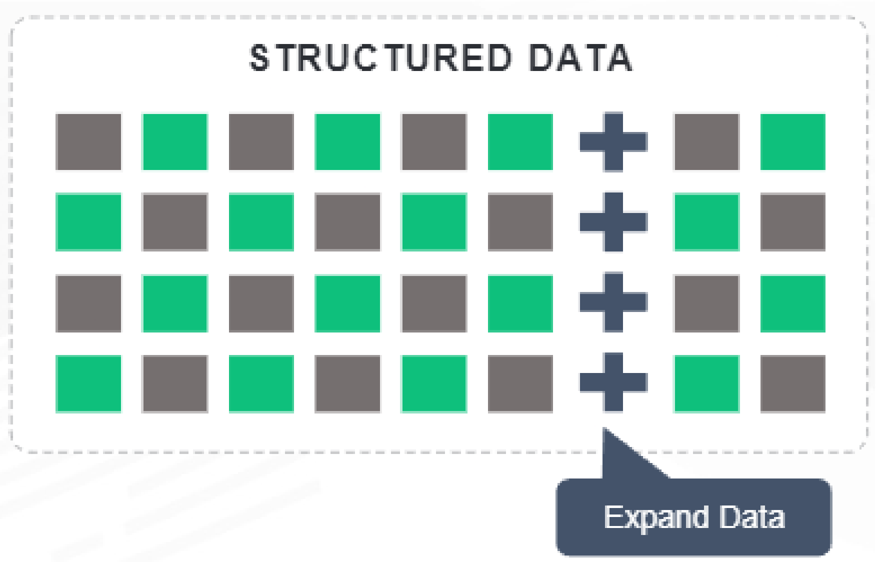
1.5 Expansion of Structured Data

The following are some reasons of expansion of Structured data:

- Creation of multiple backups of data and database.
- Having different databases like test database, development database, reporting database, production database etc.
- Having multiple copies of production database.
- Duplication of data: having multiple copies of the same data. One of the common causes of data expansion.

Techniques for controlling data expansion:

- Data Compression
- Deduplication
- Cloning
- Archiving



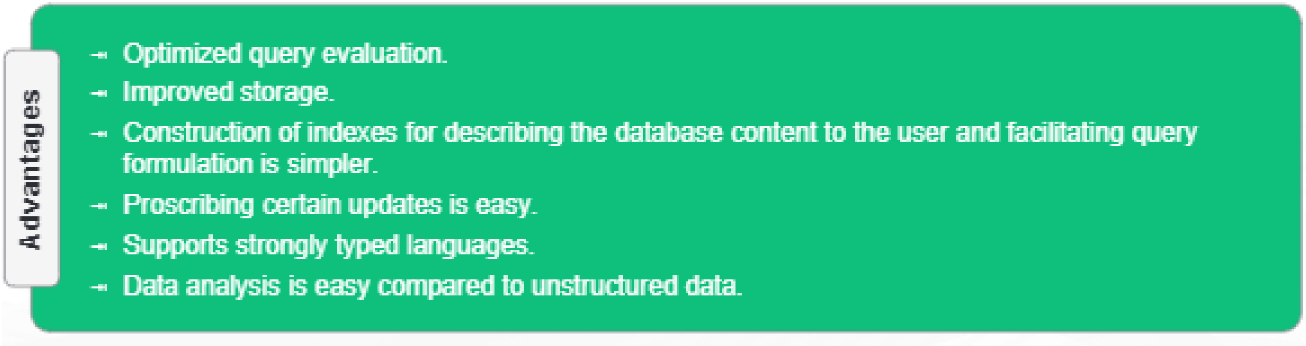
Understand how Structured Data expands. Data growth is unavoidable. Data expansion actually results in consumption of more storage space and there are associated complexities with data occupying more space. The cost of data storage and management also increases.

One of the major reasons for data expansion is unnecessary data duplication. Having multiple copies of the same data might be useful from a backup point of view. But this requires a lot of storage space and updating a copy of data doesn't mean that all copies of it will be updated. This results in data reliability issues as well. Data expansion can be controlled by implementing the following techniques:

- Data compression
- Deduplication
- Cloning
- Archiving

1.6 (a) Advantages and Disadvantages of Structured Data

The following are advantages of Structured data:



Advantages

- Optimized query evaluation.
- Improved storage.
- Construction of indexes for describing the database content to the user and facilitating query formulation is simpler.
- Proscribing certain updates is easy.
- Supports strongly typed languages.
- Data analysis is easy compared to unstructured data.

Structured Data has its own advantages and disadvantages.

Some of the advantages include:

- Query evaluation is optimized since data is well organized. Simple queries have the ability to fetch the desired information in less time.
- The Storage is simpler as the structure is defined.
- Describing the database content to the user becomes simpler and construction of queries and indexes to retrieve the information is also much simpler.
- Data analysis is much simpler compared to Unstructured Data.
- Strongly typed languages are supported.

1.6 (b) Advantages and Disadvantages of Structured Data

The following are disadvantages of Structured Data:

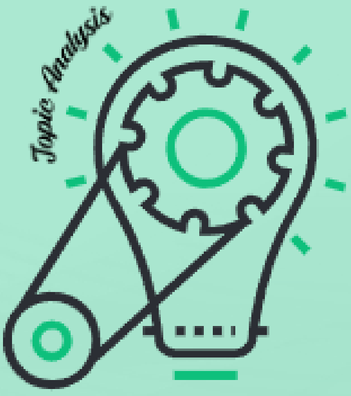
Disadvantages

- Schema needs to be predefined before populating data.
- Not suitable if the amount of data is very high.
- Cannot be used for data that keeps changing frequently.
- Subsequent extensions of a previously defined schema with content is difficult.

Some of the disadvantages include:

- Data cannot be populated to the database, without a predefined schema.
- This requires much time and effort even before data is loaded. In case of an update to the database, a complete database is required to be modified from the scratch, especially .
- If the amount of data is very high and from multiple sources, it is not possible to give a predefined structure to the data.
- Frequently changing data cannot be categorised under a defined structure.

What did You Grasp?



1. State True or False.

   Structured Data requires a predefined schema before the data is stored in database.

   A) True
   B) False

2. Which of the following options is an example of Structured Data?

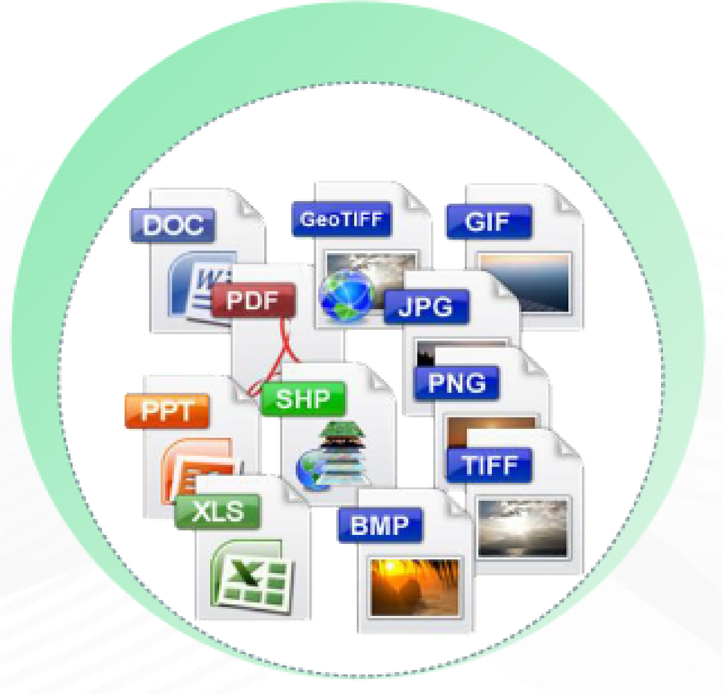
   A) Email
   B) Videos
   C) Images
   D) Phone numbers

2 What is Unstructured Data?

Unstructured Data:

- No identifiable structure for this kind of data
- Have internal structure, but no predefined schema

- Cannot be stored in rows and columns like a relational database
- No fixed data model, a massive unorganized conglomerate of various information
- Require more storage space than structured data



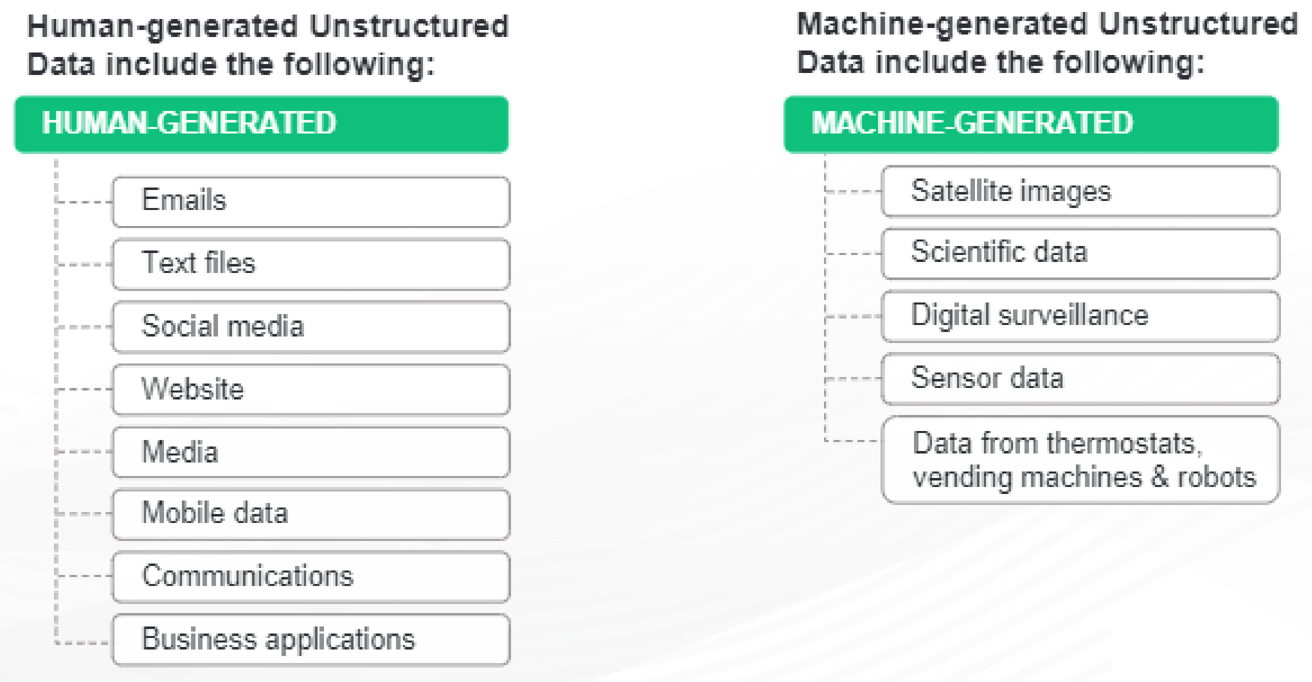
Around 80% of business data is Unstructured Data. Unstructured Data is defined as data that has no identifiable structure. There is an internal structure to unstructured data, but it is not structured by pre-defined data models or schema. Unstructured Data may be textual or non-textual and human-generated or machine-generated. This primarily involves text files, audio, video and images. Without preprocessing, we cannot store unstructured data in a table.

The term 'Unstructured data' needs to be understood properly. We can call a data 'unstructured' if it has some form of structure but is not helpful to processing. For example, the email messages have data with some implied structure, still, normal data mining tools cannot parse the information in an email. Therefore, it comes under Unstructured Data category.

Searching information from the Unstructured Data is not straightforward, as compared to Structured Data. Complex queries and algorithms are needed to retrieve information from an unstructured data.

Full-text search is one of the ways of searching text documents, but this is not suitable for pictures or videos.

2.1 Examples of Unstructured Data



Human-generated Unstructured Data include the following:

- **Text files:** Word processing, spreadsheets, presentations, email, logs.
- **Email:** Email has some internal structure, thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
- **Social Media:** Data from Facebook, Twitter, LinkedIn.
- **Website:** YouTube, Instagram, photo sharing sites.
- **Mobile data:** Text messages, locations.
- **Communications:** Chat, IM, phone recordings, collaboration software.
- **Media:** MP3, digital photos, audio and video files.
- **Business applications:** MS Office documents, productivity applications.

Machine-generated unstructured data include the following:

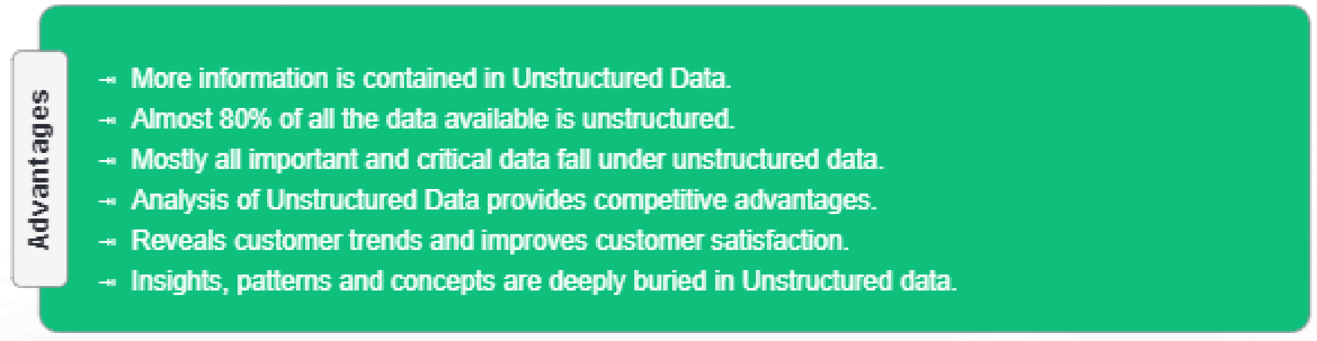
- **Satellite imagery:** Weather data, landforms, military movements.
- **Scientific data:** Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- **Digital surveillance:** Surveillance photos and video.
- **Sensor data:** Traffic, weather, oceanographic sensors.

Common examples of the types of files considered as unstructured data:

- Books
- Health records
- Satellite images
- Adobe PDF files
- A warranty request created by a customer service representative
- Notes in a web form
- Objects from presentations
- Blogs
- Text messages
- Word documents
- Videos
- Photos and other images

2.1 (a) Advantages and Disadvantages of Unstructured Data

The following are advantages of Unstructured Data:



Advantages

- More information is contained in Unstructured Data.
- Almost 80% of all the data available is unstructured.
- Mostly all important and critical data fall under unstructured data.
- Analysis of Unstructured Data provides competitive advantages.
- Reveals customer trends and improves customer satisfaction.
- Insights, patterns and concepts are deeply buried in Unstructured data.

Like structured data, unstructured data also has several advantages and disadvantages. Let's look at some of these.

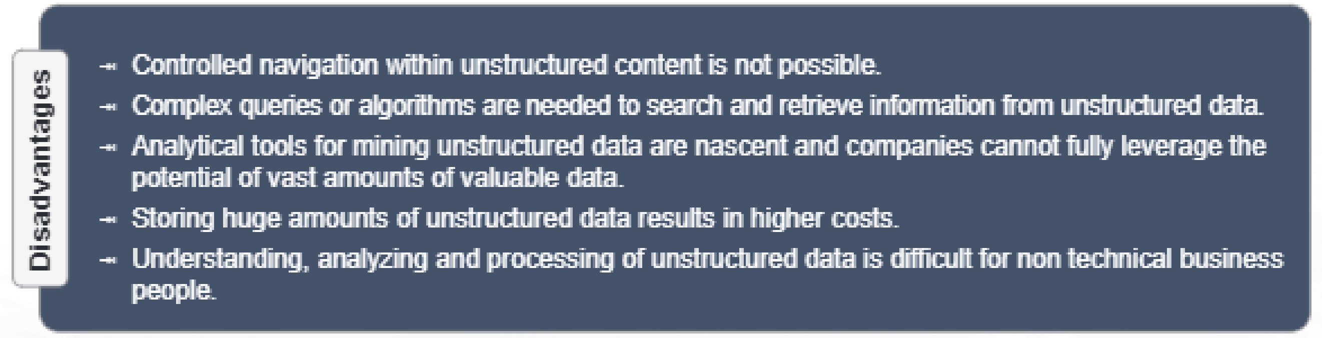
Advantages of Unstructured Data are:

- The information here is not organized into a defined 'Schema'. Almost all that is outside a Relational database is Unstructured Data. There are many types and sources of unstructured data. This means more information is contained in them
- Amount of unstructured data is astonishing. This category makes up nearly 80% of all digitally stored data which is very advantageous for data-driven businesses.

- Unstructured Data critical to businesses are commonly present in websites, presentations or documents. All the medical records are unstructured. The test reports and values are mostly stored as unstructured data.
- Despite the fact that unstructured data is not well-organized or easy to access, analyzing this data can offer competitive advantages to businesses. Especially in this era of Artificial Intelligence and Machine Learning, data is everything. The way data is visualized, analyzed, processed and insights are derived is critical for businesses. This information can give a clear picture of the customer behaviour and requirements, which is a major driver for business growth.
- Analyzing social trends such as tweets, Facebook posts and transcripts from support calls will give a clear view of how customers perceive a value regarding the products. Understanding the issues proactively and acting upon them can dramatically improve customer satisfaction. The requests for new features can be captured, grouped and prioritized in ways that were previously not possible.
- Mining the sales information, social media sentiments, news and survey data helps us to understand the pattern of sales, why it went up or down. Insights, patterns, and concepts are usually buried deep in human-language communication data like, emails, notes, surveys, social network posts, tweets, etc. which will not be available to businesses otherwise.

2.1 (b) Advantages and Disadvantages of Unstructured Data

The following are disadvantages of Unstructured data:



Disadvantages

- Controlled navigation within unstructured content is not possible.
- Complex queries or algorithms are needed to search and retrieve information from unstructured data.
- Analytical tools for mining unstructured data are nascent and companies cannot fully leverage the potential of vast amounts of valuable data.
- Storing huge amounts of unstructured data results in higher costs.
- Understanding, analyzing and processing of unstructured data is difficult for non technical business people.

Disadvantages of Unstructured data:

- Since data is diverse and has no defined structure, controlled navigation is not possible with unstructured data.
- Since data is not grouped or organized using a defined schema, it is difficult to retrieve data using simple queries. Complex algorithms need to be developed to search and retrieve the necessary information from the huge pile of data.
- Unless data is analyzed, there is no value to it. Analyzing unstructured data requires sophisticated tools. Many enterprises lack these tools and are not able to leverage the potential of data completely.
- Storing unstructured data is a complex and an expensive exercise for the organizations. Though the cost of storage devices have come down, the associated cost of maintaining data centres is still a nightmare to organizations. Apart from the cost, highly skilled personnel is also required for storing, maintaining and analyzing the data. Many organizations still lack this.

- Non-technical business people still find the concept of unstructured data difficult. A rich technical expertise is required for analyzing and processing unstructured data. The business people will not be able to understand the intricacies, consequently, they will not be able to explain this to their customers.

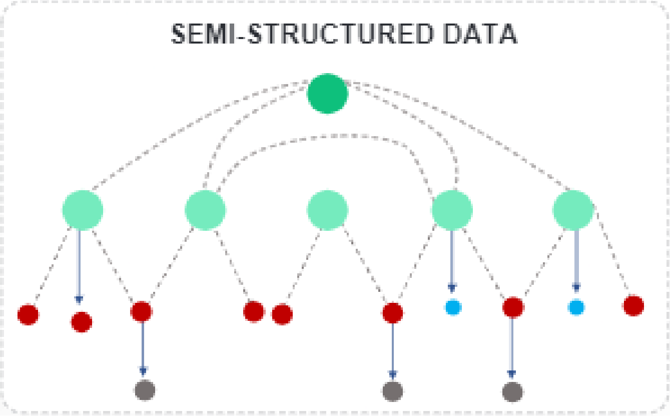
What did You Grasp?

1. Which of the following options is true with respect to unstructured data?
   - A) There is no internal structure to this type of data.
   - B) The data model is fixed.
   - C) Requires less storage space than structured data.
   - D) Preprocessing is required to store unstructured data.
2. Which of the following options is/are human-generated data?
   - A) Email
   - B) Sensor data
   - C) Text files
   - D) Scientific data
   - E) Videos

3 What is Semi-structured Data?

Semi-structured data:

- Often explained as schema-less or self-describing
- Contains semantic tags, but do not comply with the standards or structure of typical relational databases
- No separate description of the type or structure of the data
- Does not require a schema definition, the definition is not impossible, but it is optional
- Data can have different attributes, and new attributes can be added anytime



Semi-structured data is the data that does not reside in a relational database but has some organizational properties that make it easily analyzable. When processed, it can be stored in a Relational database. Though Semi-structured data do not comply with the standards of a Relational database, tags and the other types of markups in them helps in the identification of the individual, distinct entities within the data.

Semi-structured data can have n-level hierarchies that differentiate it from the structured data, where data is represented as a flat table. Documents and databases can be semi-structured. The Semi-structured data also form only 5-10% of all informative data but has critical business use cases.

In semi-structured data, some of the aspects are structured, while the others are not. The Semi-structured data serves as an important source for Big Data analytics but may not be suitable for traditional databases.

3.1 Examples of Semi-structured Data

Examples of Semi-structured Data include:

- **Markup language XML** - a set of documents encoding rules that define a human- and machine-readable format.
- **Open Standard JSON** - a lightweight, plain-text, data-interchange format based on a subset of the JavaScript programming language.
- **NoSQL** - some noSQL databases contain semi-structured data.

Examples of semi-structured data include:

- **XML:** XML is a semi-structured document language. XML defines the set of rules for encoding documents. XML's initial focus was on the documents, but later it found many use cases like a representation of arbitrary data structures and serving as the base language for communication protocols. The tag-driven structure of XML is highly flexible, and developers can use it for globalizing the data structure, storage and transportation on the web.
- **JSON:** JSON is a lightweight, plain-text, data-interchange format based on a subset of the JavaScript programming language. One of the semi-structured data interchange formats. JSON structure can contain name/value pairs (or object, hash table, etc.) and an ordered value list (or array, sequence, list). The structure of JSON is interchangeable among languages and is very helpful in transmitting data between web applications and servers.
- **NoSQL:** Many NoSQL ("not only SQL") databases contain semi-structured data. In NoSQL databases, a schema is not separated from the data and is very flexible. For example, the text with varying lengths cannot be stored in a relational database but stored in a NoSQL DB. Some newer NoSQL databases like MongoDB and Couchbase store data in the JSON format which is semi-structured.

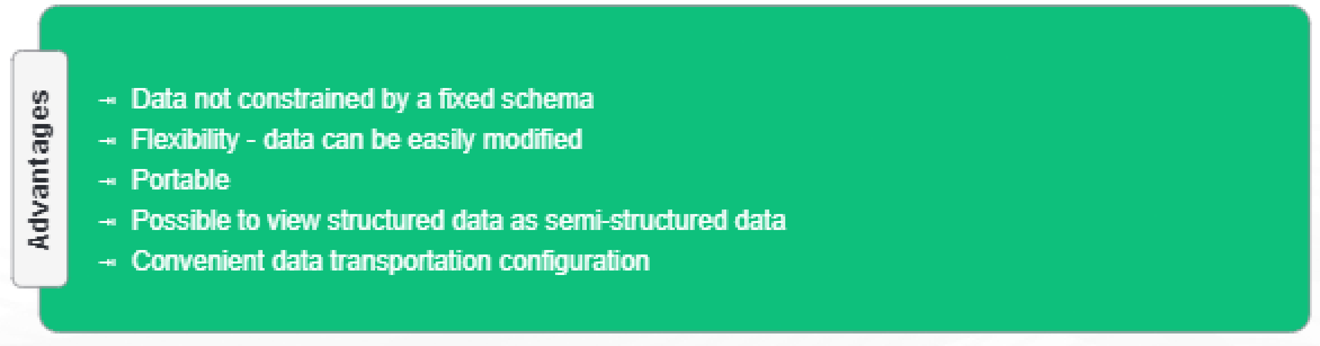
3.2 Applications of Semi-structured Data

Applications containing semi-structured data include:

- Big Data Infrastructure.
- Web applications
  - LinkedIn
  - Salesforce
  - Reader recommendations in Amazon

3.3 (a) Advantages and Disadvantages of Semi-structured Data

The following are advantages of Semi-structured data:



Advantages

- Data not constrained by a fixed schema
- Flexibility - data can be easily modified
- Portable
- Possible to view structured data as semi-structured data
- Convenient data transportation configuration

Semi-structured data is an evolving form and is still under research. Let's look at some of the advantages and disadvantages of semi-structured data.

Advantages of Semi-structured data are:

- Data is not constrained by a fixed schema like that in a structured data and is very flexible. This type of data can be used to represent information that cannot be constrained by a schema.
- Semi-structured data offers greater flexibility and it can be easily modified even if the data changes frequently.
- This type of data is portable that the data exchange is also possible between contrasting databases.
- Semi-structured data are very supportive in screening structured data as semi-structured data.
- This type of data is convenient for transportation configuration.

3.3 (b) Advantages and Disadvantages of Semi-structured Data

The following disadvantages of Semi-structured data:

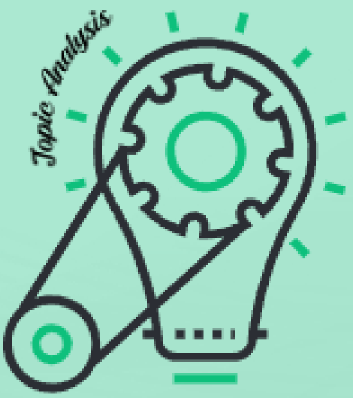
Disadvantages

- Queries are less efficient than in a constrained structure
- Data diversity
- Extensibility
- Storage

Disadvantages of Semi-structured data are:

- Records in semi-structured database are stored with only one of a kind IDs that are referenced with indicators to their specific locality on a disk. Despite the fact that queries are very well-organized, this approach is not practical while doing searches over scores of records, for the reason that it is forced to seek in the various regions of the disk by following the indicators.
- In federated systems, data diversity is a serious issue. This involves complexities such as unit and semantic incompatibilities, grouping incompatibilities, and non-consistent overlapping of sets.
- It is important that extensibility from the context of data is an indication of data presentation and not data processing. The Data processing should be able to happen without the aid of database updates.
- Transfer formats like XML are universally in text or in Unicode. They are helpful for data transfer but not for storage. The presentations are instead, stored by deep-seated and accessible systems that support such standards.

What did You Grasp?



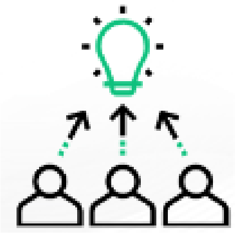
1. Which of the following options is/are true about semi-structured data?

   A) Semantic tags need to comply with the standards or a defined structure.
   B) Tags and other markups help in identification of entities in the data.
   C) Data attributes cannot be modified once added.
   D) Structured data cannot be viewed as Semi-structured data.

2. State True or False.
   Schema definition cannot be done at all, for Semi-Structured data.

   A) True
   B) False

Activity



Your facilitator will give different examples of Data. Discuss among yourselves and identify which category they belong to. Justify your answer, why they belong to that category.

4 Comparison of Structured, Unstructured and Semi-structured Data

The following table describes the comparison between all types of data.

	Structured data	Unstructured data	Semi-structured data
Technology	Relational database tables.	Character and binary data like image, text, video, audio etc.	XML/RDF.
Transaction management	Matured transaction management, various concurrency techniques.	No transaction management, no concurrency.	Transaction management adapted from RDBMS, not matured.
Version management	Versioning over tuples, rows, tables, etc.	Versioned as a whole.	Not very common; versioning over triples or graphs possible.
Flexibility	Schema-dependent, very flexible.	Very flexible, absence of schema.	Flexible, tolerant schema.
Scalability	Scaling DB schema is difficult.	Highly scalable.	Schema scaling is simple.
Robustness	Very robust, enhancements since 30 years.	-	New technology, not widely spread.
Query performance	Structured query, allows complex joins.	Only textual queries possible.	Queries over anonymous nodes are possible.
Content type	Sensitive Content & System Maintenance Data, Core Component Data.	Textual Content.	Non-sensitive Core Component Data, Flexible Content & Individual Data.

Comparison of Structured, Unstructured and Semi-structured data will give you a clear idea of which data will be insightful under what scenario.

What did You Grasp?

1. State True or False.
   Structured data is highly scalable compared to unstructured data.

   A) True
   B) False

2. Which of the following options is right, with respect to queries in Structured data?

   A) Structured queries that allow complex joins
   B) Only textual queries are possible
   C) Queries over anonymous nodes are possible
   D) None of the above

5 Noisy Data

Noisy Data:

- Contains a large amount of meaningless information, hence considered to be corrupted.

Noise:

- An anomaly and is considered to be an error. According to Zhu et al. (2004), noise in data negatively affects data mining results.
- A persistent phenomenon that is superimposed on desired information.
- Unnecessarily occupies more storage space.
- Two major sources of noise:
  - Implicit errors
  - Random errors

We've learned about the different types of data. The amount of data generated every day is growing in an exponential phase. In order to derive value out of data, it needs to be processed and analyzed. It is also important to distinguish between useful and not-so-useful data before it is processed. Unless useful data is identified early in the process, organizations will end up spending a lot of time, money and effort, for the value output that is much less than the expected. One such phenomenon that affects the results of Data Mining is the presence of Noise in the data.

Noisy data is thus defined as the data that contains a large amount of meaningless information, i.e noise. Noisy data is considered to be the corrupted data. Noise is an anomaly and thus considered to be an error in the data. Zhu et al, in their research article titled "Class Noise vs. Attribute Noise: A Quantitative Study", which was published in the journal, Artificial Intelligence Review in 2004, explained that noise negatively affects data collection and data preparation processes, hence Data Mining results are affected. According to them, there are two major sources of Data Noise:

- **Implicit errors:** Implicit errors are introduced by data measurement tools, such as different types of sensors.
- **Random errors:** Random errors are the ones that are generated by Batch Processing Systems or experts who collect them. One example is an error that is introduced during Data Digitization process.

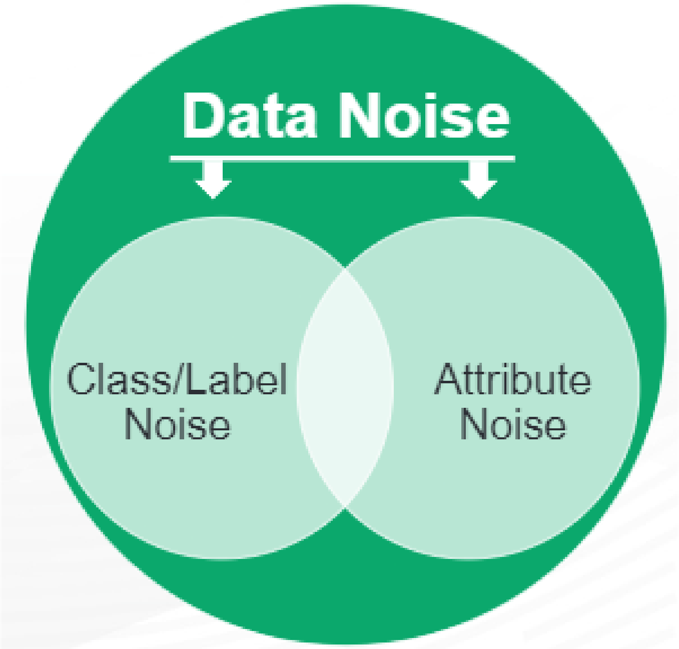
The performance of data classifiers depends, to a larger extent, on the quality of data used for training the machine learning models. Thus, noise in data may have a direct impact on the predictions made by machine learning models.

There is another phenomenon called 'outliers' which should not be confused with noise. The outliers are not meaningless but are out of range data. They are still legitimate, but they do not comply with the general behaviour of data. They sometimes are considered while data processing, according to the need. One example of their application is fraud detection systems.

Detecting noise in data is very important for improving machine learning performance. Understanding of data and its knowledge can be expanded by means of identifying the exceptional cases in data.

5.1 Types of Data Noise

There are two major categories of data noise as depicted in the illustration below:



There are two major categories of Data Noise:

1. **Class/Label Noise:** Class noise occurs when data is incorrectly labelled. There are several reasons for class noise. Some of the common reasons include: Subjectivity during labelling, data entry errors, insufficient information used to label a data, etc. There are two types of class noise:

   a) **Contradictory examples** - refers to duplicate data that have different class labels.

   b) **Misclassifications** - data that are labelled as a class that is different from the real one.

2. **Attribute Noise:** Attribute noise refers to the corruption that occurs in one or more of attribute values. Some of the examples of attribute noise are:

   a) Erroneous attribute values

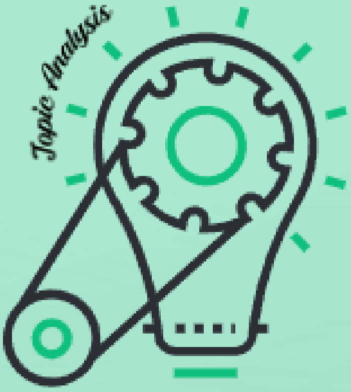
   b) Missing or unknown attribute values

   c) Incomplete attributes or "do not care" values

Zhu et al discovered that attribute noise is more harmful than class noise. They also found that eliminating class and attribute noise will improve classifier performance.

Usage of noise filters can greatly reduce noise in data.

What did You Grasp?



1. State True or False.

   Implicit errors are generated by batch processors.

   A) True
   B) False

2. Which of the following options is an example for class noise?

   A) **Missing attribute values**
   B) **Misclassifications**
   C) **Incomplete values**
   D) **Erroneous values**

In a nutshell, we learnt:

1. Data types - Structured, Unstructured and Semi-structured Data.
2. Structured Data, its organization, examples, advantages and limitations.
3. Unstructured Data, examples, advantages and limitations.
4. Semi-structured Data and examples.
5. Comparison between Structured, Unstructured and Semi-structured Data.
6. Noise in data, sources of noise, types of Data Noise and ways to minimize