




## Data Ingestion

### Data Ingestion Basics

 15 mins

#### Table of Contents

- [Introduction](#)
- [Types of Data](#)
- [Difference between data warehouse and a data lake](#)

Data Ingestion is usually one of the first steps in a data science endeavour. In the start of a project, it is generally done manually, but later on, in many cases it is automated to save time. Nonetheless, it is essential to understand the nuances of data ingestion.

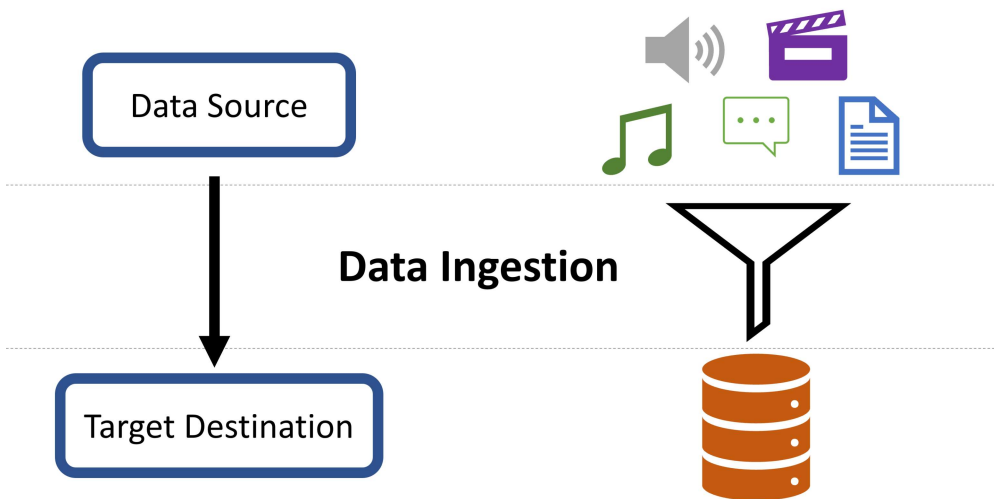
#### What you will learn in this module?

- What does data ingestion entail?
- What are the types of data?

- What is a data warehouse and a data lake?
- 

## Introduction

Data ingestion is the process of importing, transferring, loading and processing data for later use. Essentially, the data is transported from various sources to a target database. The source of data could range from a simple spreadsheet to real-time data from an online application. Whereas the target destination is generally a database, document store, data warehouse, or even a data lake. We will discuss more about these in a later section.



---

## Types of data

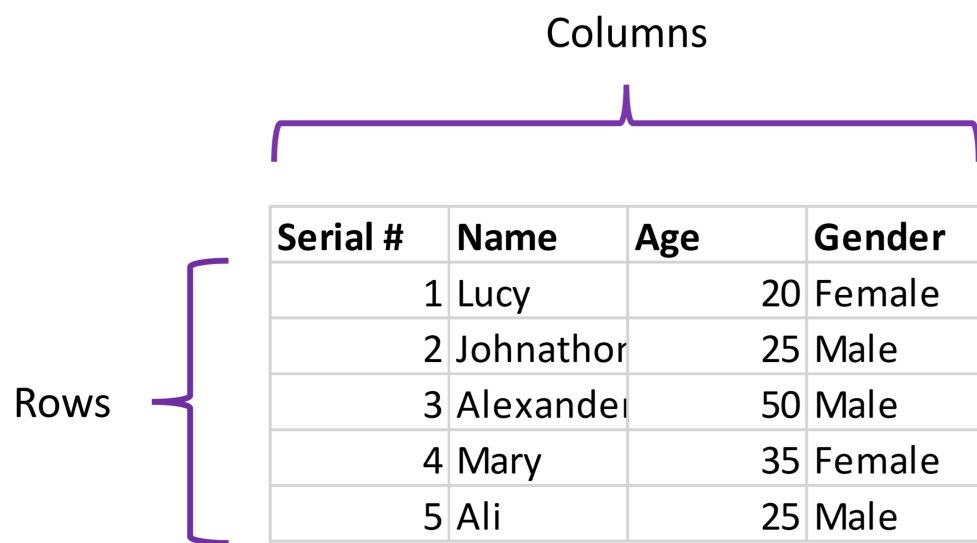
Every piece of information that we can perceive is in essence data. From the student's test score result to a snapchat video, all of it is data. In order to grasp the concept of ingesting data, it is imperative to differentiate between different types of data. We can generally make three categories of data:

1. Structured data
2. Semi-structured data
3. Unstructured data

Let's go over each of these.

### Structured data:

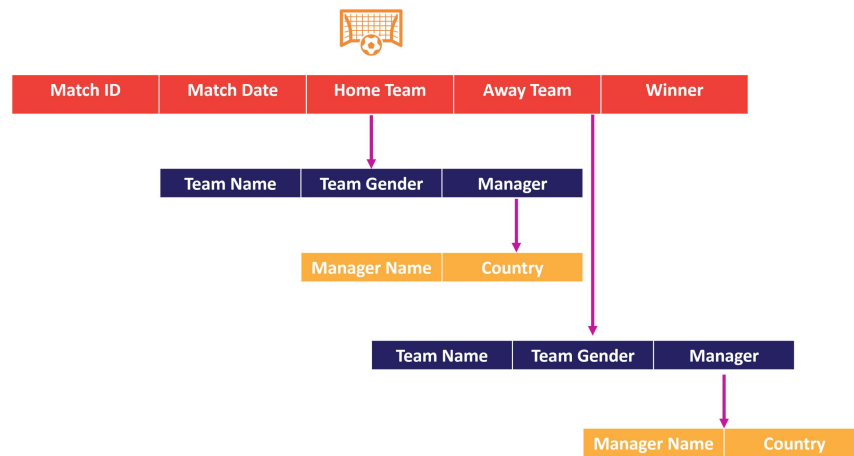
As the name suggests such data have a structure or a shape. The ideal example is a table which has columns and rows that make it very easy to read. Below is an example of a simple table that has four columns to represent different attributes and five rows each containing a data entry.



Serial #	Name	Age	Gender
1	Lucy	20	Female
2	Johnathon	25	Male
3	Alexander	50	Male
4	Mary	35	Female
5	Ali	25	Male

We could also have more complicated data structures which are sometimes referred to as complex data. In these datasets we have columns nested into columns. For example, in a soccer tournament data, instead of having one giant table including all the information of every detail, we can have columns/tables nested. (See the figure below). Home team details are nested inside the home team column. And inside the home team table, manager details are nested in the manager column.

## Complex Data For Soccer Tournament



Such type of data are a little tricky to ingest and we will talk more about them in the next lesson.

**Interesting Fact:** The architecture of the nested data which explain the heirarchy is called schema.

### Semi-structured data:

Semi-structured is slightly different than structured data and does not have nicely defined rows and columns. This type of data does not conform to a data model but has *some* structure. It lacks the rigid 'schema' of relational databases, but has some organization which make it easier to analyze.

Some examples of semi-structured data could be:

- Emails
- HTML code
- PDF documents with graphs

It is usually stored in XML and JSON file-types.

Some advantages of semi-structured data are:

- Flexibility of storing varying data
- Can easily deal with heterogeneity of sources

Disadvantages:

- Difficult to store
- Not performant as structured data

### **Unstructured data:**

This type of data neither conforms to a data model, nor does it have any structure. Such type of data cannot be stored in a relational database. We are seeing an uptick in the generation of unstructured data, and interest in analyzing it.

Processing or analyzing unstructured data is a lot more challenging than processing or analyzing structured/unstructured data. Unstructured data cannot be queried using regular SQL and it needs special tools for searching and other operations.

Some examples of unstructured data could be:

- Text documents
- Audio files
- Sensor data
- Video files

Generally, unstructured data is stored in data lake. We will talk more about these in the coming section.

---

## **Data Warehouse & Data Lake**

Data warehouse(DW) is a system used for analysis and reporting of data. DW are essentially central repositories where all the data of an organization is stored and then processed for further analysis. It is considered a core component of business intelligence.

A key element of WH is that it has some structure which is called schema. Just like an actual warehouse which racks and aisles, the DW also has compartmentalization and organization. Because of this schema, fetching and analyzing data is pretty quick and intuitive.

Another characteristic of DW is that it also has tiers for different purposes. The top most tier is the most accessible which can be retrieved very quickly, whereas the data in the bottom layer will be the one which needs to be accessed less frequently.

On the other hand, Data Lake is also a repository for data, but it lacks schema/structure. But that gives data lake a lot flexibility. Any type of data can be stored in its raw form to be retrieved later. As mentioned previously, unstructured data can only be stored in Data Lakes.

Usually, all raw data is stored in a Data Lake, and then DW extracts relevant data from a Data Lake to process and store it in the DW, which is then analyzed and reported.

### Data Warehouse



### Data Lake



Author Name: Umar Farooq Ghumman

Author Contact: [umarfarooq.ghumman@vertica.com](mailto:umarfarooq.ghumman@vertica.com)

## Resources

- [Data Lake wiki](#)