Course: IBM Data Analyst
From Coursera

Cristo Daniel Alvarado

7 de noviembre de 2025

Índice general

Lista de Códigos

Capítulo 1

Introduction to Data Analytics

1.1. Data

So, data is important for an enterprise whose main concern is keep up with technological development. Interpreting the correct data can give us valuable information that's able to change the development of an enterprise.

In summary, data is relevant.

1.2. Roles in Data

There are several roles which are important in the process of analyzing data. These roles are the following:

- Data Engineer.
- Data Analyst.
- Data Scientist.
- Business Analyst.
- Business Inteligence Analyst.

Let's look into each one of them.

1.2.1. Data Engineer

It's the first one of the process line to work with data. He develop and maintains data arquitectures in order to make data available for business operations and analysis.

They work in the data ecosystem to work with data to:

- Extract, integrate and organize data from disparse sources (or different sources).
- Clean, transform and prepare data.
- Desing, store and manage data in data repositories (such as a database, etc...).

So, the role of a Data Engineer is one of the most fundamental roles in the Data Analysis. Without him, it would be much more complicated to work with data.

Observación 1.2.1
They make data available in several formats for different systems and processes that involve data.

His work serves Business Applications and Data Analysts and Data Scientists.

Observación 1.2.2 (**Skills**)
A data engineer needs the following skills in order to perform his job:

(1) Knowledge in programming.

(2) Sound knowledge of systems and technology architectures.

(3) In-depth understanding of databases (relational and no relational).

1.2.2. Data Analyst

In a few words, a Data Analyst translates information (such as tables, numbers and graphs) into plain language, so that organization can make desicions.

Observación 1.2.3 (**Responsabilities of a Data Anayst**)
A Data Analyst has the following responsabilities:

- Inspect and clean data for deriving insights, that is, for a specific purpose.
- Identify correlations, find patternsand apply statistical methods to analyze and mine data.
- Visualize data to interpret and present the findings of data analysis.

Basically, a Data Analyst answer questions that are made by the company.

What's the information we can infer from this data?

Observación 1.2.4 (**Skills**)
The skills a Data Analyst needs in order to perform his job are the following:

- Knowledge of spreedsheets, writing queries, using statistical tools to create charts and dashboards.
- Programming skills.
- Strong analytical and storytelling skills.

1.2.3. Data Scientist

Data scientist analyze data for:

- Actionable insights.
- Create predictive models using machine learning and deep learning.

They answer more complex questions, such as:

- How many new social media followers will we gain next month?
- Is this financial transaction fraudulent?

A data scientist needs the following skills:

- Knowledge of Mathematics and Statistics.
- Understanding of programming languages, databases and building data models.
- Domain knowledge.

1.2.4. Business Analyst and BI Analyst

They leverage the work of data analyst and data scientist to make strategic and tactical business desicions.

BI analyst focus on market forvces and external influcneces that shape their business, organizse and monitor data on different business functions.

They explore data to extract insights and actionables that improve business performance.

1.2.5. Summary

All the data roles are described in the following table:

Role	Function
Data Engineer	Converts raw data into usable data.
Data Analytics	Use data to generate insights.
Data Scientists	Use Data Analytics and Data Engineering to predict the future using data from the past.
Business Analysts and Business Ingelligence Analysts	Use insights and predictions to drive decisions that benefit and grow their business.

Cuadro 1.1: Summary of Data Roles

1.3. What is Data Analysis?

Definición 1.3.1 (**Data Analysis**)
Data Analysis is a process that consists of the following:

(1) Gather, clean, analyze and mine data.

(2) Interpret results.

(3) Report findings.

The objective of Data Analysis is to find patterns in data that can help to make better desicions.

With this signs and correlations we can make decisions. A Data Analyst helps a business to make better decisions using data:

- Understand past performance.
- Take informed decisions.
- Validate course action - saving time and resources, ensuring success.

1.3.1. Types of Data Analysts

There are four types of primary Data Analysts:

Descriptive Analytics	**Diagnostic Analytics**	**Predictive Analytics**	**Prescriptive Analytics**
What happened?	Why did it happen?	What will happen next?	What should be done about it?
Privedis insights into past events	Takes the insights from descriptive analytics to dig deepear to finde the cause of the outcome	Leverages historical data and trends to predict future outcomes	Analyzes past decisiones and events to estimate the likelihood of different outcomes.

Cuadro 1.2: Primary Types of Data Analytics

Observación 1.3.1 (**Predictive Analytics**)
Predictive Analytics forecast what *may* happen in the future.

1.3.2. Data Analysis Process

As always, the process of a data analysts starts with **understanding the problem that needs to be solved and the desired result to be achieved** (where are we? and where do we want to be?).

Then decide **how we can mesure our goal** (such as KPIs, etc...). Deciding what will be measured and how it will be measured is crucial for the success of the data analysis.

Finally, **gathering the data necessary to perform the analysis**. Identify the data needed, the sources from which the data will be obtained, and the best tools and techniques to perform the anaylsis.

Finally, **cleaning the data** (fix quality issues that may affect the outcome of the analysis). Standarizing the data format coming from different sources.

Analyzing and mine data. Extracting, analyzing, manipulating data from different perspectives, understand trends, identify correlations, and find patterns and variations.

Interpret results. Evaluat the defenability of analyss and circumstances under which analysis may not hold true.

And, finally, **present results** in clear and impactful ways using data visualization tools and techniques.

In summary:

(1) **Understand the problem and define the goal**.

(2) **Decide how to measure the goal**.

(3) **Gather the data**.

(4) **Clean the data**.

(5) **Analyze and mine the data**.

(6) **Interpret results**.

(7) **Present results**.

Observación 1.3.2
Something important is to confirm a hypothesis and use data to make a story. Somethimes it is useful to break down information into subsets or smaller parts.

Observación 1.3.3 (**Data Analysis vs Data Analytics**)
During this course, the terms data analysis and data analytics mean the same thing. There is a subtle difference between them, but for the purpose of this course, they are used interchangeably.

The difference is the following:

- The **dictionary meanings** are:
  - **Analysis** - detailed examination of the elements or structure of something
  - **Analytics** - the systematic computational analysis of data or statistics

Analysis can be done without numbers or data, such as business analysis psycho analysis, etc. Whereas Analytics, even when used without the prefix "Data", almost invariably implies use of data for perfoming numerical manipulation and inference.

1.4. Responsabilities of a Data Analyst

The role of a data analyst may differ from one organization to another, but in general, a data analyst is responsible for the following tasks:

- Aquire data (from primary and secondary data sources).
- Creating queries to extract required data.
- Filtering, cleaning, standarizing and reotrganizing data.
- Using statistical tools to interpret data stes.
- Using statistical techniques.
- Analyze patterns.
- Preparing reports and charts.
- Cerating appropiate documentation

1.4.1. Skills of a Data Analyst

The data analysis processs requieres a set of thecnical, functional and soft skills. Such as:

- **Technical**:
  (1) Expertise in using spreedsheets.
  (2) Proficensy in statistic analysis and visualization and software tools, such as: IBM; Power BI; Tableau; Excel.
  (3) Proeficiency in prograaming languages, such as: R, Python, C++, Java, and MATLAB.

(4) Good knowledge of SQL and ability to work with data in relational and non-relational databases.

(5) Ability to access and extract data from repositories, such as Data Mars, Data Warehouses, Data Lakes and Data Pipelines.

(6) Familiarity with big data processing tools like Hadoop, Hive, and Spark.

- **Functional**:

  (1) **Proeficiency in statistics**: Analyze data, validate the analysis, identify fallacies and logical errors.

  (2) **Analytical skills**: Research and interpret data, theorize and make forecasts.

  (3) **Problem-solving skills**: Identify issues, think critically and make data-driven decisions.

  (4) **Probin skills**: Identify and define the problem statement and desired outcome.

  (5) **Data visualization skills**: Create clear and compelling viisualizatoins to present the analysis.

  (6) **Project management skills**: manage the process, people, dependencies and timelines.

- **Soft skills**:

  (1) Work colaboratively in a team environment.

  (2) Communicate effectively with both technical and non-technical stakeholders.

  (3) Tell a compeling and convincing story.

  (4) Gather support and buy-in for your work.

  (5) Curiosity and eagerness to learn new tools and techniques.

  (6) Intiution to identify trends and patterns in data.

Some of the Technical, Functional and Soft skills will be reviewed during this course.

1.5. Generative AI

Definición 1.5.1 (**Generative AI**)
Generative AI refers to a *class of artificial intelligence models that create new content such as text, images, music, and more by learning patterns from existing data.*

Generative AI can respond naturally to human conversation and serve as a tool for customer service and personalization of customer workflows. For example, you can use AI-powered chatbots, voice bots, and virtual assistants that respond more accurately to customers for first-contact resolution.

1.5.1. Key techniques in generative AI

- **Generative adversarial networks (GANs)**: GANs consist of two neural networks: the generator and the discriminator. The generator creates new data, whereas the discriminator evaluates it. Over time, the generator improves to produce realistic data.
- **Variational autoencoders (VAEs)**: VAEs encode input data into a compressed format and then decode it back, generating new data points similar to the input data.
- **Transformers**: Used primarily in natural language processing (NLP), transformers generate human-like text by predicting the next word in a sequence. Generative Pre-trained Transformer 3 (GPT-3) is a notable example.

Generative AI can be applied in various use cases to generate virtually any kind of content. The technology is becoming more accessible to users of all kinds thanks to cutting-edge breakthroughs like GPT that can be tuned for different applications.

Some of the use cases for generative AI include the following:

- Implementing chatbots for customer service and technical support.
- Deploying deepfakes for mimicking people or even specific individuals.
- Improving dubbing for movies and educational content in different languages.
- Writing email responses, dating profiles, resumes, and term papers.
- Creating photorealistic art in a particular style.
- Improving product demonstration videos.
- Suggesting new drug compounds to test.
- Designing physical products and buildings.
- Optimizing new chip designs.
- Writing music in a specific style or tone.

Generative AI tools exist for various modalities, such as text, imagery, music, code, and voices. Some popular AI content generators to explore include the following:

- Text generation tools include GPT, Jasper, AI-Writer, and Lex.
- Image generation tools include Dall-E 2, Midjourney, and Stable Diffusion.
- Music generation tools include Amper, Dadabots, and MuseNet.
- Code generation tools include codeStarter, Codex, GitHub Copilot, and Tabnine.
- Voice synthesis tools include Descript, Listnr, and Podcast.ai.
- AI chip design tool companies include Synopsys, Cadence, Google, and NVIDIA.

Observación 1.5.1 (**Data Anaytics and AI**)
Generative AI has many applications that can enhance your data analytics work:

- **Data augmentation**: Create synthetic data to augment existing data sets, which is especially useful when data is scarce or imbalanced. This can improve predictive model performance.
- **Anomaly Detection**: Identify anomalies or outliers by understanding the distribution of normal data. This is valuable in fraud detection, network security, and quality control.
- **Text and image generation**: Generate realistic text and images for marketing, content creation, and customer engagement, such as automatic product descriptions and marketing visuals.
- **Simulation and forecasting**: Simulate scenarios and forecast future events by generating potential outcomes from historical data. This is crucial in financial planning, supply chain

management, and strategic decision-making.

Generative AI is a transformative technology that can significantly enhance your capabilities as a data analyst. By mastering generative AI techniques, you can unlock new possibilities in data augmentation, anomaly detection, content creation, and forecasting. As you embark on this journey, remember to balance innovation with ethical responsibility, ensuring that AI is used positively.

1.6. Average Process of a Data Analyst

Some of the responsabilities of a Data Analyst are:

(1) Acquiring data from varied sources.

(2) Creating quieries for fetching data from data repositories.

(3) Looking for insights into data.

(4) Interacting with stakeholders for gathering information and presenting findings.

(5) Cleaning and preparing the data for data analysis.

The last part is one of the biggest responsabilities of a data analyst and it is one of the most important ones.

Usually, a data analyst is going to be presented with some problem (the rising of the price of a product, etc...), what he has to do is obtain information about the probem (complaint data, subscriber information data and biling data).

Observación 1.6.1 (**Hypothesis**)
The use of an initial hypothesis could be useful in order to try to answer some of the questions presented before.

Once the questions aries and hypothesis are created, we have to identify the datasets that are going to be isolated and analyze them in order to validate or refute the hypothesis.

Capítulo 2

Data Ecosystem

A Data Analyst's ecosystem includes the following:

- Infraestructure
- Software
- Tools
- Frameworks
- Processes used to:
  - Gather,
  - Clean,
  - Mine,
  - and Visualize

  data.

2.1. Overview of the Ecosystem

Data can be categorized into the following types:

1. **Structured**: Data that follows rigid format and can be organized into rows and columns.

   This data is typically seen in databases and spreedsheets.

2. **Semi-structured**: Mixed of data that has consisten characteristics and data that does not conform to a rigid structure.

   For example, emails (strucured data such as subject, email, etc. . . and unstrucutred as the message).

3. **Unstrucutred**: Data that is complex and mostly qualitative information that cannot be structured into rows and columns.

   For example, photos, videos, files and social media contant.

Observación 2.1.1 (**Importance of Data Types**)
The type of data determines the type of data repositories where data can be collected, stored in and, tools used to query or process data.

2.1.1. Data Formats

Data can come in a variety of file formats, such as:

1. Relational.
2. Non-relational databases.
3. APIs.
4. Web Services
5. Data streams.
6. Social Platforms.
7. Sensor Devices.

Definición 2.1.1 (**Data Repositories**)
A **Data Repository** is a container of data. These include:

1. Databases.
2. Data Warehouses.
3. Data Marts.
4. Data Lakes.
5. Big Data Stores.

A Data Repository collects types of data, file formats and are sources of data. A data repository is dependent upon this characteristics mentioned earlier, to collect, store and mine data.

Ejemplo 2.1.1
A Big Data we will need data warehouses, where we need to store and process large volume and high velocity data. Also, frameworks that allow perform complex analytics in big data.

2.1.2. Languages

The languages available in the Data Analyst Ecosystem are:

- **Query Languages**: Such as SQL for quering data and manipulating data.
- **Programming Languages**: Such as Python for data applications.
- **Shell and Scripting Languages**: For repetitional and operational tasks.

2.1.3. Automated Tools

Automated tools, frameworks, and processes for all stages of the analytics process are part of the Data Analysis Ecosystem. These tools allow to:

- **Gather, extract, transfrom and load data**.
- **Data wrangling and cleaning**.
- **Data analysis and mining**.
- **Data visualization**.

2.2. Data

Definición 2.2.1 (**Data**)
Data is unorganized information that is processed to be meaningful. This can be conformed of:

- Facts, observations, perceptions.
- Numbers, characters, symbols.
- Images.

Data can be interpreted to have a meaning.

2.2.1. Categorization of Data

We can organize data in the following three types:

1. **Structured**.
2. **Semi-structured**.
3. **Unstrucutred**.

2.2.2. Structured Data

The characteristics of Structured Data are the following:

- Well-define structure.
- Can be stored in well-defined schemas.
- Can be represented in a tabular maner using rows and columns.

Uses facts and numbers which can be:

- Collected.
- Exported.
- Stored.
- Organized.

in databases.

Ejemplo 2.2.1 (**Sources of Structured Data**)
Structued data can come from:

1. SQL Databases.
2. Online Transaction Processing.
3. Spreadsheets.
4. Online forms.
5. Sensors GPS and RFID text.

6. Network and webserver logs.

2.2.3. Semistructured Data

Characteristics of semi-structured data are the following:

- Has some organizational properties but lacks a fixed or rigid schema.
- Cannot be stored in rows and columns.
- Contains tangs and elements or metadate used to group and organize it in a hierarchy.

Ejemplo 2.2.2 (**Sources of Semi-structured Data**)
Include:

- Emails.
- XML and other markup languages.
- Binary executables.
- TCP/IP packets.
- Zipped files.
- Integration of data from different sources.

Observación 2.2.1 (**Use of XML and JSON**)
`XML` and `JSON` allow users to define tags, associate atributes and store data in a hierarchy form and are used widely to store and exchange semi-structured data.

2.2.4. Unstructured Data

Characteristics of unstructured data are the following:

- Does not have an easily identificable structure.
- Cannot be organized ina mainstream relational database in the form of rows and columns.
- Does not follow any particular format, sequence, semantics and rules.

Unstrucutred data can deal with a heterogeneity of sources and has applications in business.

Ejemplo 2.2.3 (**Sources of Unstructured Data**)
Sources of unstructured data include the following:

- Web pages
- Social media feeds
- Images in a varied file formats
- Video and audio files
- Documents and PDF files

- PowerPoint presentations
- Media logs
- Surveys

Can be stored in files and documents, such as:

- **Files and Docs** for manual analysis.
- **NoSQL Databases** for the use of analysis tools.

2.3. File Structures

Al told earlier, data has to be saved and we have plenty of file formats in order to store data. Turns out important to understand the structure of file formats in order to choose between their benefits and limitations.

We will see the following file formats:

- **Delimited file formats or** `.CSV`**.**
- **Microsoft Excel Open XML spreedsheet or** `.XLSX`**.**
- **Extensible Markup Langauge, or** `.XML`**.**
- **Portable Document Format, or** `.PDF`**.**
- **JavaScript Object Notation, or** `.JSON`**.**

2.3.1. Delimited Text Files

Definición 2.3.1 (**Delimited Text Files and Delimiters**)
Delimited Text Files are *files used to store data as text in which each line and row has a value separated by a delimiter.*

A **delimiter** is a *sequence of one or more characters for specifying the boundary between independent entities or values.*

Most common delimeters are coma, tab, colon, vertical bar and space.

Ejemplo 2.3.1 (**Commonly Used Delimited Text Files**)
Two delimited text files are Comma-separated values and Tab-separated values (`.CSV` and `.TSV`, respectively) are the most commonly used.

Ejemplo 2.3.2
A `.CSV` looks like this:

```
1 id,nombre,edad,puesto,salario
2 1,Cristo Alvarado,24,Data Analyst,45000
3 2,Ana Lopez,29,Software Engineer,62000
4 3,Marco Gomez,31,Project Manager,70000
5 4,Laura Ruiz,26,QA Tester,40000
6 5,Carlos Perez,35,DevOps Engineer,75000
```

Código 2.1: .CSV File Content Example.

And, a .TSV like this:

```
id  nombre  edad  puesto  salario
1 Cristo Alvarado 24  Data Analyst  45000
2 Ana Lopez 29  Software Engineer 62000
3 Marco Gomez 31  Project Manager 70000
4 Laura Ruiz  26  QA Tester 40000
5 Carlos Perez  35  DevOps Engineer 75000
```

Código 2.2: .TSV File Content Example.

The first row are the names of the variables of each column.

Observación 2.3.1
Delimiters represent one of variuous means to specify bonudaries in a data stream.

2.3.2. Microsoft Excel Open XML Spreadsheet or .XLSX

Is a Microsoft Excel Open XML file format that falls under the spreadsheet file format. It is an XML-based file format created by Microsoft.

Is a secure file format since it cannot contain malicious malware.

2.3.3. Extensible Markup Language or .XML

Definición 2.3.2 (**Extensible Markup Language (XML)**)
Extensible Markup Language is a markup language with a set of rules for encoding data.

- This format is readable by humans and machines.
- Self-descriptive language.
- Similar to .HTML in some respects.
- Does not use predefined tags like .HTML does.
- Platform independent.
- Programming langauge dependent.
- Makes it simpler to share data between systems.

Ejemplo 2.3.3
An example of a .XML file is the follwing:

```
<?xml version="1.0" encoding="UTF-8"?>
<employees>
  <employee>
    <id>1</id>
    <nombre>Cristo Alvarado</nombre>
    <edad>24</edad>
```

```xml
    <puesto>Data Analyst</puesto>
    <salario>45000</salario>
  </employee>
  <employee>
    <id>2</id>
    <nombre>Ana Lopez</nombre>
    <edad>29</edad>
    <puesto>Software Engineer</puesto>
    <salario>62000</salario>
  </employee>
  <employee>
    <id>3</id>
    <nombre>Marco Gomez</nombre>
    <edad>31</edad>
    <puesto>Project Manager</puesto>
    <salario>70000</salario>
  </employee>
  <employee>
    <id>4</id>
    <nombre>Laura Ruiz</nombre>
    <edad>26</edad>
    <puesto>QA Tester</puesto>
    <salario>40000</salario>
  </employee>
  <employee>
    <id>5</id>
    <nombre>Carlos Perez</nombre>
    <edad>35</edad>
    <puesto>DevOps Engineer</puesto>
    <salario>75000</salario>
  </employee>
</employees>
```

Código 2.3: .XML File Content Example.

2.3.4. Portable Document File (PDF)

Definición 2.3.3 (**Portable Document File (PDF)**)
A **Portable Document File** .PDF is a file format developed by adobe to present documents independent of application software, hardware, and operating systems.

- Can be viewed the same way on any device.
- Is frequently used in legal and financial documents.
- Can also be used to fill data for forms.

2.3.5. JavaScript Object Notation (JSON)

Definición 2.3.4 (**JavaScript Object Notation (JSON)**)
A **JavaScript Object Notation** .JSON is a text-based open standard designed to transmit

data over the web.

- Langauge-independent data format.
- Can be read by any programming language.
- Easy to use.
- Comptaible with a wide range of browsers.
- Best tools for sharing data.

Ejemplo 2.3.4
An example of a `.JSON` file is the follwing:

```
{
  "employees": [
    {
      "id": 1,
      "nombre": "Cristo Alvarado",
      "edad": 24,
      "puesto": "Data Analyst",
      "salario": 45000
    },
    {
      "id": 2,
      "nombre": "Ana Lopez",
      "edad": 29,
      "puesto": "Software Engineer",
      "salario": 62000
    },
    {
      "id": 3,
      "nombre": "Marco Gomez",
      "edad": 31,
      "puesto": "Project Manager",
      "salario": 70000
    },
    {
      "id": 4,
      "nombre": "Laura Ruiz",
      "edad": 26,
      "puesto": "QA Tester",
      "salario": 40000
    },
    {
      "id": 5,
      "nombre": "Carlos Perez",
      "edad": 35,
      "puesto": "DevOps Engineer",
      "salario": 75000
    }
```

```
    ]
}
```

Código 2.4: .JSON File Content Example.

2.4. Sources of Data

Common sources of data are:

- Relational databases.
- Flat files and XML Datasets.
- APIs and Web Services.
- Web Scraping.
- Data Streams and Feeds.

2.4.1. Relational Databases

Typically, an organizational unit posess a store of data ot his:

- Business activities.
- Customer transactions.
- Human resource activities.
- Workflows.

These systems use Relational Databases, such as SQL Server, Oracle, MySQL and IBM DB2. In this relational databases they store structured data.

These relational databases could be used for analysis.

2.4.2. Flat File and XML Datasets

External to the organization, there are other datasets.

Ejemplo 2.4.1
Goverment can give datasets which are either public or private. For example, demographic or economic dataset are released on a regular time period.

Also, this type of data could be a point of sale, financial or weather.

Idea 2.4.1
This data could be used in companies to define a strategy, predict demand, and make distribution decisions, among other things.

This data is often available in form of:

- Flat files.
- Spreadsheet files.
- XML documents.

Definición 2.4.1 (**Flat Files**)
Flat Files store data in plain text format. Each line, or row is one record. Each value is separated by a delimiter (comma, semicolon, tabs, etc. . .).

Flat Files only have one table to organize all of his data. A common example of a Flat File are `.CSV` and `.TSV` files.

Definición 2.4.2 (**Spreadsheet Files**)
Spreadsheet files are a special type of flat files, which can organizse data in a tabular format, can contain multiple worksheets.

Common examples are `.XSL` or `.XLSX` spreadsheet formats.

Also, Google sheets, apple numbers and libreoffice calc.

Definición 2.4.3 (**XML Files**)
XML Files contain data values that are identified or marked up using tags (as we saw earlier). Can support more complex data structures.

The uses we have for XML files are online surveys, bank statements, and other unstructured datasets.

2.4.3. APIs and Web Services

Definición 2.4.4 (**Application Program Interface API**)
An **Application Programming Interface (API)**, is a *set of rules and protocols that allows different software applications to communicate and interact with each other.* It acts as *an intermediary, enabling one application to request data or functionality from another without needing to understand the internal workings of the other application.*

Definición 2.4.5 (**Web Service**)
A **Web Service** is a *software that enables machine-to-machine communication over the internet using standardized protocols like HTTP, allowing different applications to exchange data and function together regardless of their underlying programming languages or platforms.*

So, basically we can obtain data using APIs and Web Services.

Observación 2.4.1 (**Calling an API or Web Service**)
Typically, APIs and Web Services listen for upcoming requests, which can be in form of Web requests or Network requests. After a request has been made, they can return data in form of a JSON, XML, Media Files, etc. . .

Ejemplo 2.4.2 (**Examples of APIs**)
Some popular APIs are the following:

- Twitter and Facebook APIs.
- Stock Market APIs.

- Data Lookup and Validation APIs.

2.4.4. Web Scraping

Definición 2.4.6 (**Web Scraping**)
Web Scraping is *the automated process of using bots to extract data from websites and save it into a structured format, like a database or spreadsheet.*

Instead of copying and pasting by hand, software is used to parse a website's HTML code to pull specific information, which can then be used for analysis, research, or other applications.

Observación 2.4.2 (**Use of Web Scraping**)
Web Scraping is useful to:

- Extract data from unstructured sources.
- Also known as screen scraping, web harvesting, and data straction.
- Download specific data based on defined parameters.
- Can extract text, contact information, images, videos, product items, and more...

Ejemplo 2.4.3
Popular uses of Web Scraping are:

- Providing price comparasions by collecting product details from retailer, manufacturers, and eCommerce websites.
- Generating sales leads though public data sources.
- Extracting data from posts and authors on varios forums and communities.
- Collecting training and testing datasets for machine learning models.

Observación 2.4.3 (**Popular Web Scraping Tools**)
Some popular web scraping tools are:

- BeautifulSoup.
- Scrapy.
- Pandas.
- Selenium.

2.4.5. Data Streams

Definición 2.4.7 (**Data Stream**)
A **data stream** *is a continuous, ordered sequence of data generated by a source in real-time. Unlike data that is stored and processed in batches, a data stream is processed as it arrives,*

allowing for immediate analysis and action.

Examples include sensor data from IoT devices, website clickstreams, or financial transactions.