



Extract

Prof. MSc. Fernando Prass

fprass@gmail.com

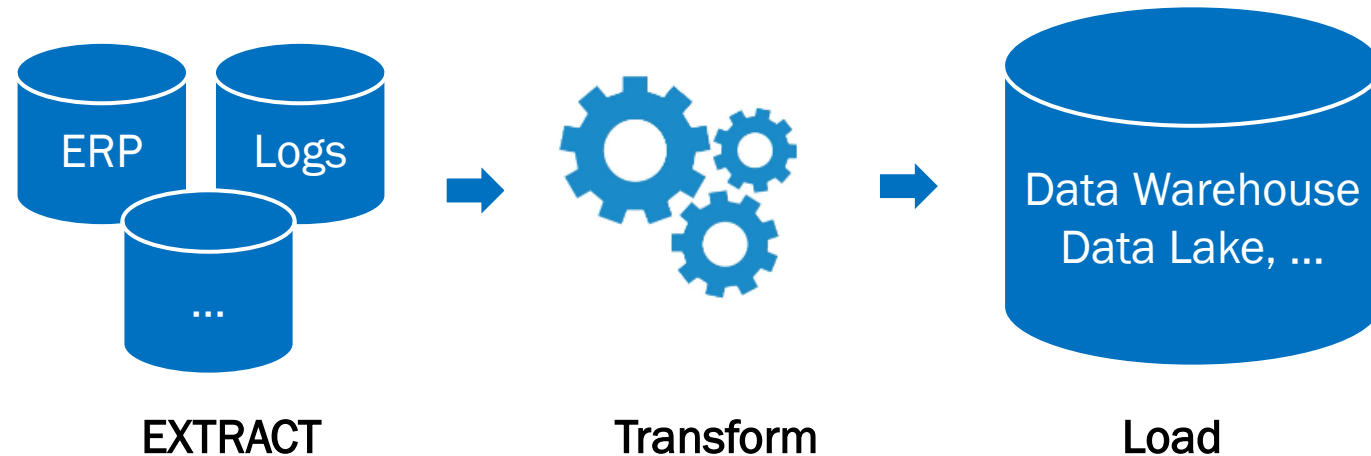
github.com/fernandoprass/etl

Who am I?

- Fernando Sarturi Prass
- Bachelor in Information Systems
- Master in Computer Science
- Senior Software Engineer (tech leader) at Workday
- fprass@gmail.com
- github.com/fernandoprass
- twitter.com/oFernandoPrass
- linkedin.com/in/fernandoprass
- lattes.cnpq.br/2919187023046130

EXTRACT, Transform, Load (ETL)

- **EXTRACT**, Transform, Load (ETL) *“is a data integration process that combines, cleans and organizes data from multiple sources into a single, consistent data set for storage in a data warehouse, data lake or other target system”*. (IBM)



Extract

- During data extraction, raw data is copied or exported from source locations to a staging area. Source locations can come from a variety of different sources, structured or unstructured:

Data extraction techniques

- Web scraping
- Database Extraction
- File Parsing
- API Integration
- Manual Extraction

Web scraping



- Web scraping, also known as web harvesting or web data extraction, is a technique used for extracting data from websites. It typically involves automated processes implemented using a bot or web crawler, which can access the Web using the HTTP Protocol or a web browser.
- While it can be done manually, the term usually refers to automated processes that fetch and extract specific data, often into a central local database or spreadsheet for analysis

Web scraping – How it works?

1. Using a web scraping tool to send an HTTP request to the target website's dedicated server. This request retrieves the HTML content of the web pages.
2. Once the website grants access to the scraper, the HTML markup is parsed. Parsing involves understanding the structure and arrangement of the HTML document and identifying specific HTML tags, attributes, or CSS selectors associated with the desired data.
3. The extracted and cleaned data is then stored in a structured format, such as CSV, JSON, or a database. This ensures that the data can be easily referenced for future analysis and other purposes.

Web scraping - Example

- An example of web scraping would be a price comparison service that uses web scrapers to read price information from several online stores. This data is then used to compare prices and present the best deals to users.


Web scraping - Pros

- **Data Availability:** It enables access to vast amounts of data from the internet, which can be used for various analyses and business intelligence.
- **Cost-Efficiency:** It is a cost-effective method of data collection, especially when compared to manual data gathering.
- **Automation:** Web scraping tools can automate the collection of data, saving time and resources.
- **Competitive Analysis:** Businesses can use web scraping to monitor competitors and market trends

Web scraping - Cons

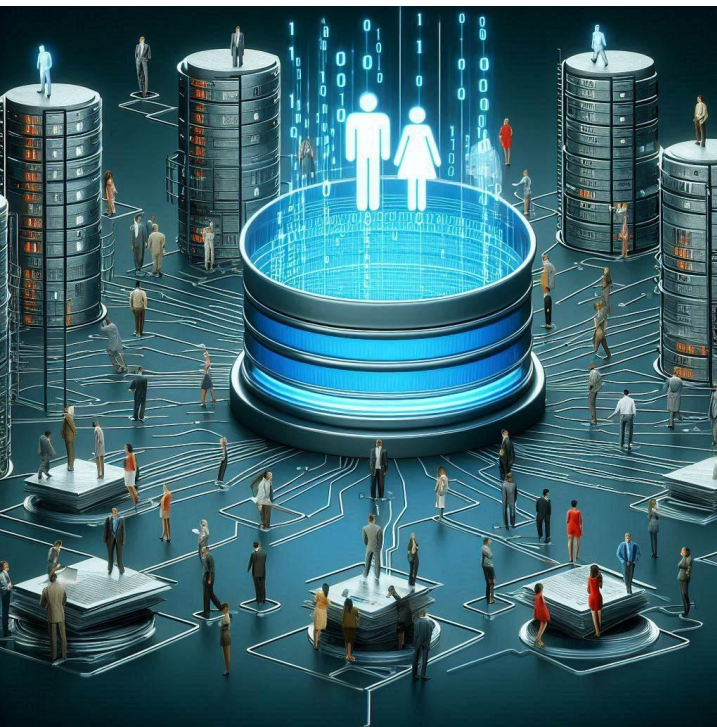
- **Legal and Ethical Concerns:** There are legal boundaries and ethical considerations to keep in mind, as some websites prohibit scraping of their data.
- **Data Quality:** The quality of scraped data can vary, and there may be a need for additional verification processes.
- **Technical Challenges:** Web scraping can be technically challenging, especially when dealing with complex websites or data formats.
- **Maintenance:** While low, there is still some maintenance required to keep scraping tools up-to-date with changes in web technologies.

Web scraping – References



- Advantages and Disadvantages of Web Scraping: The Good, the Bad, and the Scraped
 - coredevsltd.com/articles/advantages-and-disadvantages-of-web-scraping
- Web Scraping Python Tutorial – How to Scrape Data From A Website
 - www.freecodecamp.org/news/web-scraping-python-tutorial-how-to-scrape-data-from-a-website
- 75 Web Scraping Examples that Will Save Your Time
 - netpeaksoftware.com/blog/75-web-scraping-examples-that-will-save-your-time
- 10 FREE Web Scrapers That You Cannot Miss
 - www.octoparse.com/blog/9-free-web-scrapers-that-you-cannot-miss

Database Extraction



- Database extraction, also known as *Database querying*, is a process within the broader field of data extraction techniques, focusing on retrieving relevant information from databases. Here's a concise overview:
- It involves using queries to retrieve specific data from a database system for analysis, migration, or storage purposes.

Database Extraction – How it works?



1. **Query Definition:** Begin by defining the query based on the specific data you want to retrieve. Specify the relevant tables, columns, and any conditions or filters to narrow down the results.
2. **Query Formulation:** Once the query is conceptualized, write it in the appropriate syntax of the chosen database query language (e.g., SQL). Ensure that the query accurately represents your data retrieval requirements.
3. **Query Execution:** Execute the query against the database, which returns a result set containing data that matches the query criteria. Analyze, filter, sort, or aggregate the result set as needed for further processing.

Database Extraction - Example

- An SQL query like `SELECT * FROM customers WHERE purchase_date >= '2021-01-01'`; can be used to extract all customer records with purchases made in the year 2021 from an e-commerce database.

Database Extraction - Pros

- **Targeted Retrieval:** Extracts specific data, enhancing efficiency.
- **Analytical Aid:** Facilitates data analysis and informed decision-making.
- **Migration Support:** Assists in data migration and backup processes.

Database Extraction - Cons

- **Technical Knowledge:** Requires understanding of database languages.
- **Time Intensive:** Potentially slow for large datasets.
- **Accuracy Risks:** Poorly designed queries may yield incorrect data.
- **Security Concerns:** Mishandling sensitive data can lead to security issues.

Database Extraction – References



- Extraction in Data Warehouses
 - docs.oracle.com/cd/B10500_01/server.920/a96520/extract.htm
- Techniques For Periodically Extracting Data From Relational Databases
 - alirezasadeghi1.medium.com/techniques-for-periodically-extracting-data-from-relational-databases-323d97cac326

File Parsing



- File parsing is a crucial technique in data extraction that involves processing a file's content into a structured, usable form. This technique is essential for making sense of and utilizing the information contained within various file formats.
- File parsing transforms complex data into manageable parts, allowing systems to recognize and use the data effectively. It goes beyond simple data extraction by not only copying data but also understanding it.

File Parsing - How it works?

1. Using a web scraping tool, send an HTTP request to the target

File Parsing - Example

- A common example of file parsing is converting a dense CSV file into a more manageable JSON format. For instance, using a Python script to read a CSV file and convert it into a JSON object which can then be easily manipulated or sent over a network.

File Parsing - Pros

- **Facilitates Data Integration:** Helps knit varied data streams together, crucial for organizations using multiple systems.
- **Enables Advanced Analytics:** With neatly parsed and structured data, organizations can tackle complex analytical tools for deeper insights.
- **Cost Reduction:** Efficient file parsing lowers operating costs by reducing the need for human data entry.

File Parsing - Cons

- **Complexity:** Creating parsers can be complex and requires programming knowledge.
- **Maintenance:** Parsers require maintenance to keep up with changes in file formats.
- **Potential for Errors:** If not designed correctly, parsers can misinterpret data, leading to incorrect outcomes.

File Parsing - References

- Advantages and Disadvantages of Web Scraping: The Good, the Bad, and the Scraped
 - coredevsltd.com/articles/advantages-and-disadvantages-of-web-scraping

API Integration



- API Integration is a process that involves connecting two or more software systems using Application Programming Interfaces (APIs) to facilitate the seamless transfer of data.
- APIs are sets of protocols and tools that allow different software applications to communicate with each other

API Integration - How it works?

1. **Initiate the process** by verifying the user's identity using their API key. Refer to the provided API documentation or the user guide to execute the necessary API requests to acquire the specific data you need.
2. **Process the incoming data** by parsing the API's response. Isolate the pertinent details and standardize the data format to align with your system's requirements for further analysis or storage.
3. **Incorporate the data** into your existing analytics or business intelligence platforms or store it in your data warehouse. Merge this data with other datasets to conduct a thorough analysis, derive actionable insights, and develop comprehensive reports or visualizations.

API Integration - Example

- A common example of API integration is the synchronization between a Customer Relationship Management (CRM) system and an email marketing service.
- For instance, when a new contact is added to the CRM, the API integration can automatically update the email marketing service's contact list, ensuring that the new contact receives upcoming marketing campaigns.

API Integration - Pros

- **Efficiency:** Automates and streamlines workflows by allowing systems to communicate directly.
- **Real-time Data Sync:** Provides up-to-date information across platforms, supporting time-sensitive decisions.
- **Scalability:** Facilitates the growth of systems and services without the need for extensive redevelopment.
- **Enhanced User Experience:** Allows for the integration of third-party services, adding value to the existing software without starting from scratch.

API Integration - Cons

- **Complexity:** Can be technically complex to set up and maintain, especially when dealing with multiple integrations.
- **Security Risks:** Exposes systems to potential security vulnerabilities, requiring robust security measures.
- **Dependence on Third Parties:** Relies on external services, which can lead to issues if the third-party service experiences downtime or changes its API.
- **Data Quality:** Requires consistent and accurate data formats, which can be challenging to manage across different systems.

API Integration - References

- A guide to the API integration process
 - www.merge.dev/blog/api-integration-process

Manual Extraction

- Manual Extraction is a data extraction technique that involves manually copying and pasting data from the source into your database. This method is typically used when the amount of data is small and doesn't change frequently



Manual Extraction - How it works?

1. Identify the Data Source (this could be a document, a website, a database, or any other)
2. Access the Data (open the document, navigate to the website, or access the database)
3. Select the Data e Highlight the data you want to extract (this could involve selecting text, cells in a spreadsheet, or rows in a database)
4. Copy a Paste (ctrl C + ctrl V) the Data

Manual Extraction - Example

- Suppose you have a small list of contacts in a physical directory, and you want to transfer this data into an electronic spreadsheet. You would open the directory, read each contact's details, and then manually type this information into the spreadsheet.

Manual Extraction - Pros

- **No technical skills required:** Anyone who can read the data source and operate the destination (like a spreadsheet or a simple database) can perform manual extraction.
- **No special tools needed:** Manual extraction doesn't require any special software or hardware, just the source of the data and the place where you want to store it.

Manual Extraction - Cons

- **Time-consuming:** Manual extraction can be very slow, especially for large amounts of data.
- **Prone to human error:** Since the process relies on human accuracy, errors can easily be introduced into the data.
- **Not feasible for large amounts of data:** Manual extraction becomes impractical when dealing with large datasets, as the time and effort required would be enormous.

Manual Extraction - References

- Advantages and Disadvantages of Web Scraping: The Good, the Bad, and the Scraped
 - coredevsltd.com/articles/advantages-and-disadvantages-of-web-scraping

Data extraction techniques - Resume

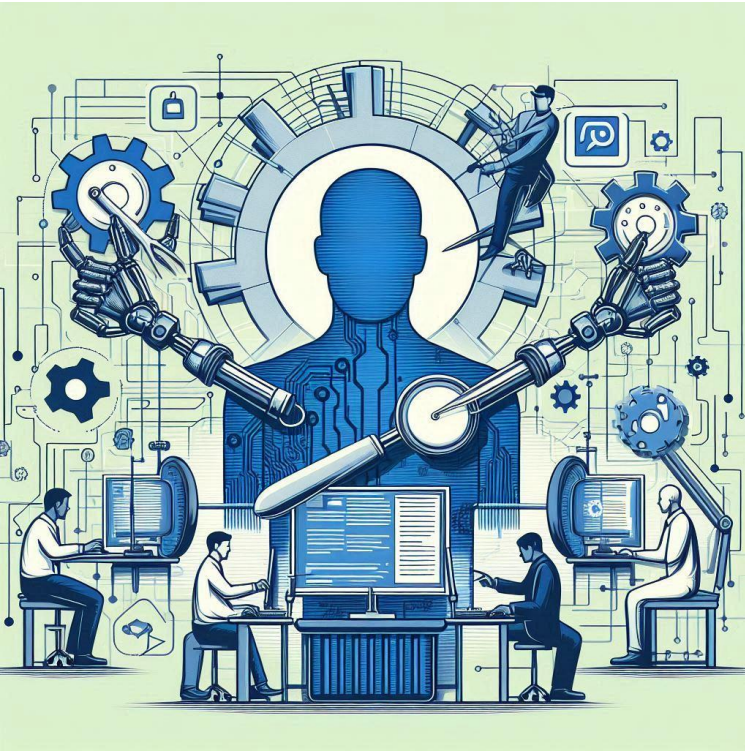
	Description	Pros	Cons
Web Scraping	Involves extracting data from websites. It's done by making HTTP requests to the URLs of the websites and then parsing the HTML response to retrieve the data.	Can extract data from any website; No need for API access.	Can be complex depending on the website structure; May violate terms of service of some websites.
API Integration	Involves interacting with a server through a series of API calls. The server then responds with the requested data, usually in a structured format like JSON or XML.	Structured data; Usually faster and more reliable than web scraping.	Requires API access which may not always be available or free.

Data extraction techniques - Resume

	Description	Pros	Cons
Database Extraction	Involves extracting data directly from a database using SQL queries or other database operations.	Can access all data in the database; Can use powerful SQL queries for data extraction.	Requires direct database access; Can be slow for large databases.
File Parsing	Involves extracting data from files such as CSV, Excel, JSON, XML etc.	Can handle a variety of file formats; Can work offline.	Requires parsing which can be complex for some file formats.
Manual Extraction	Involves manually copying data from the source and pasting it into your database.	No technical skills required.	Time-consuming; Prone to human error; Not feasible for large amounts of data.

Other Data extraction techniques

- Batch Extraction
- Image Recognition
- Optical Character Recognition (OCR)
- Text pattern matching
- Natural Language Processing (NLP)



Batch Extraction

- *Retrieving large volumes of data at scheduled intervals.*
- **Example:** Extracting daily sales data every night from a retail database.
- **Pros:** Efficient for large datasets; minimizes system impact.
- **Cons:** Not real-time; may lead to outdated information.

Image Recognition

- *Identifying objects or features in digital images or videos.*
- **Example:** Facial recognition software for security systems.
- **Pros:** Automates and speeds up data processing.
- **Cons:** Requires significant computational resources.

Optical Character Recognition (OCR)

- *Converting scanned documents into editable text.*
- **Example:** Digitizing printed legal documents for text searching.
- **Pros:** Reduces manual data entry; enables text search.
- **Cons:** Accuracy can be affected by document quality.

Text Pattern Matching

- *Finding strings that match a predefined pattern.*
- **Example:** Using regex to locate email addresses in a document.
- **Pros:** Highly efficient for specific pattern searches.
- **Cons:** Complex patterns can be difficult to construct.

Natural Language Processing (NLP)

- *Enabling computers to understand human language.*
- **Example:** Chatbots that interpret and respond to customer inquiries.
- **Pros:** Enhances human-computer interaction.
- **Cons:** Can struggle with nuances and context in language.

Myths



Myths

- AI-powered web scraping is illegal
 - The legality of web scraping depends on respecting website terms of service and ethical considerations.
- AI makes web data extraction easy, and anyone can do it
 - While some user-friendly tools exist, AI scraping often requires technical expertise.
- All online data is free range and up for grabs
 - Many websites have restrictions or require authentication for access.

Myths

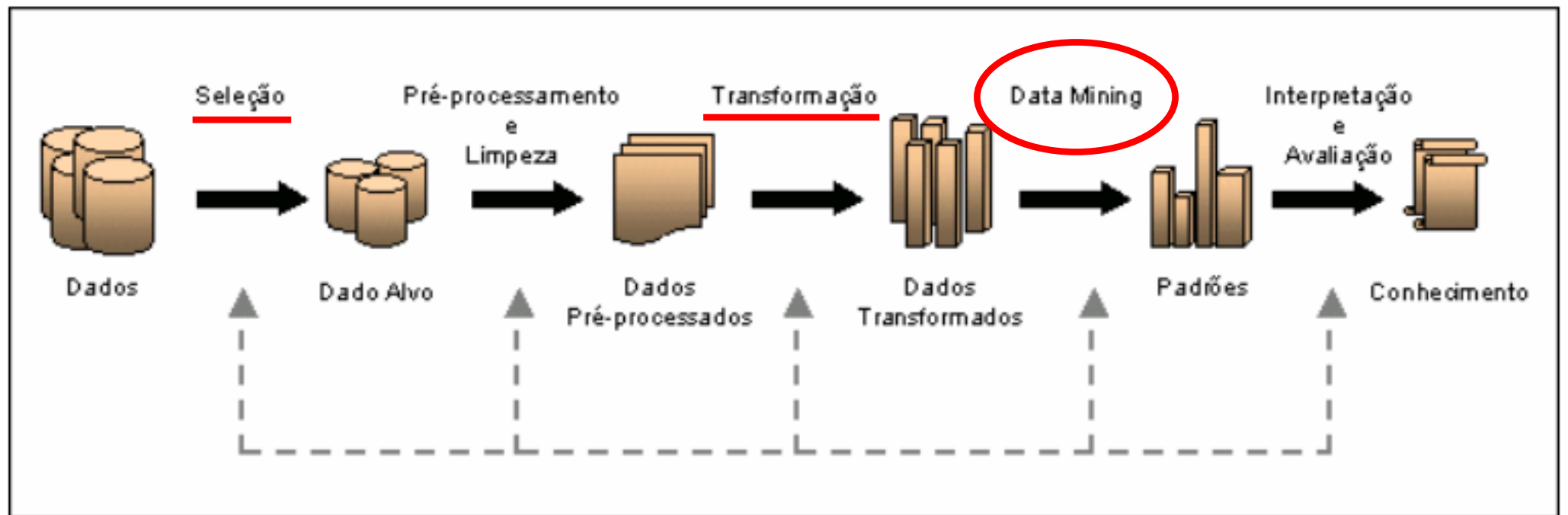
- AI can magically clean up any messy data
 - IA can be a powerful tool; but needs clean and well-structured data to work effectively.
- Web scraping is illegal
 - Web scraping is not illegal, but it's important to respect the website's terms of service.
- Any website or data can be scraped
 - There are numerous limitations and challenges associated with web scraping.
- You need to know how to code to scrape data
 - This is not necessarily true, there are tools available for non-developers.

Data Extraction = Data Mining?!?! ---

- **Data extraction** is the process of extracting data from one or more sources and converting it into a usable format for further analysis. It retrieves structured, semi-structured, and unstructured data from diverse sources, such as documents, websites, databases, etc.
- **Data mining**, also referred to as Knowledge Discovery in Database (KDD), is a technique often used to analyze large data sets with statistical and mathematical methods to find hidden patterns or trends and derive value from them.



Data Mining



O ciclo do processo de KDD. Fonte: Adaptação de FAYYAD et al. (1996)

Data Mining Techniques

- **Classification:** It's a data analysis task, where a model or classifier is constructed to predict categorical labels
- **Clustering:** It's a technique that groups similar instances on the basis of attributes into clusters
- **Neural Networking:** Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns
- **Regression Analysis:** It's a predictive modeling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor)

Data Extraction vs Data Mining

	Data Extraction	Data Mining
Goal	Involves gathering data for storage and further analysis.	works to extract data that can generate relevant insights. Its goal is to help businesses get insights they had previously ignored.
Pros	Provides data that you can build into blocks for various analytical structures.	Organizes and cleans data to offer a clear picture. By automating the mining process, data mining tools can sweep through the databases and identify hidden patterns efficiently.
Cons	Without proper data extraction, businesses lose sight of the bigger picture and cannot fully leverage the information cloaked in the data.	For data to be effectively mined, it first needs to be structured and cleaned up.

Other References

- 7 common data extraction techniques for efficient information retrieval
 - www.docsumo.com/blog/data-extraction-techniques
- 10 Best Data Extraction Software and Tools for 2024
 - microblink.com/resources/blog/data-extraction-software
- 11 Most Common Myths About Data Scraping Debunked
 - medium.com/grepsr-blog/11-most-common-myths-about-data-scraping-debunked-1f96978d6ebd
- Data Extraction or Data Mining? Understanding the Differences for Effective Data Strategy
 - www.docsumo.com/blogs/data-extraction/vs-data-mining



Thank you

Fernando Prass
fprass@gmail.com