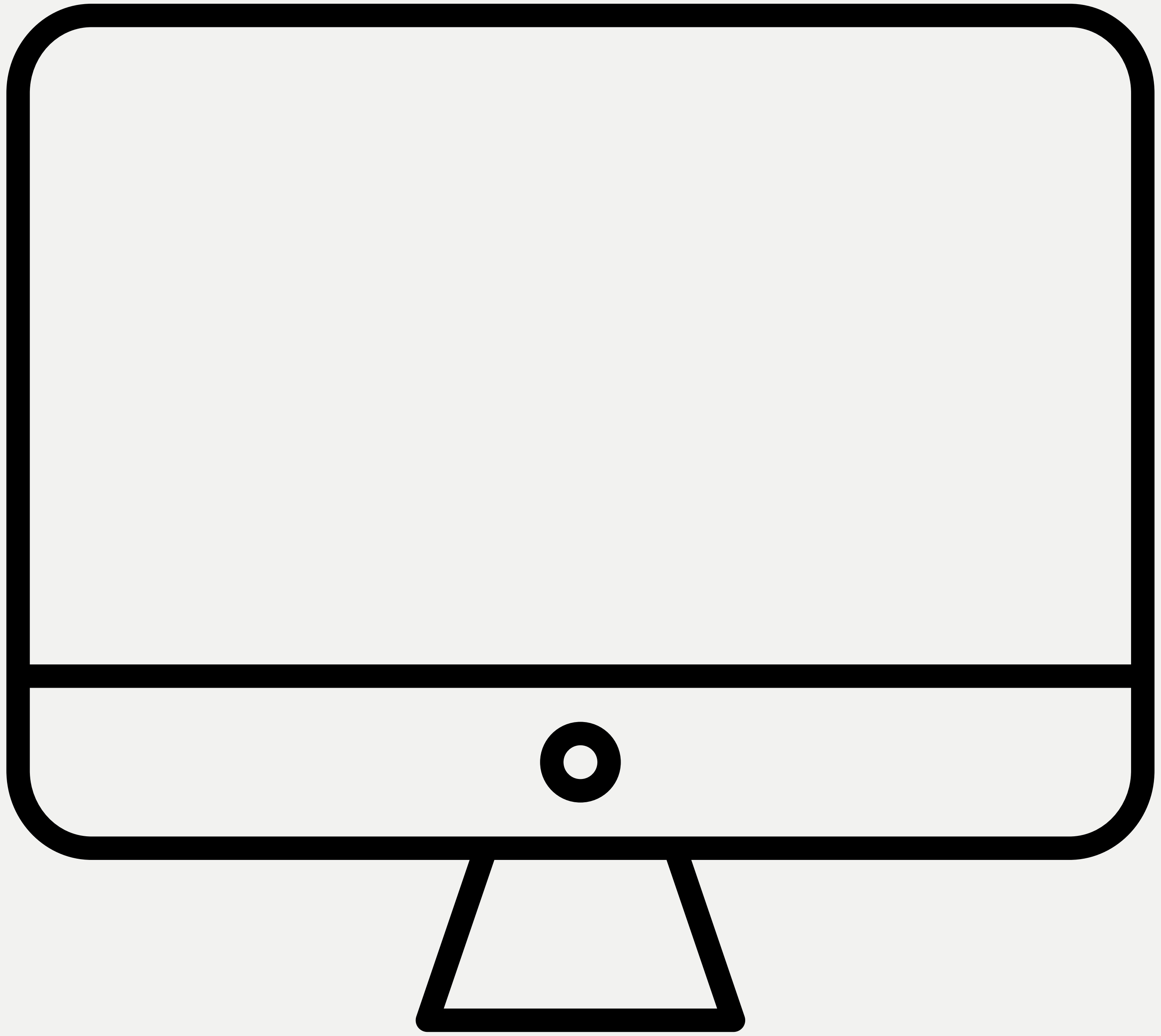
Web Scrapper

Skander RADHOUANE
Hamza BOUKHATEM
Omar ALLAL
Mohamed Ali BEN GHARBIA



Agenda

01

Introduction

- What's Web scrapping?
- How does it work?
- Web scrapper components

02

Overview

- Types of web scrapping solutions
- Tools used in web scrapping
- Use Case
- Workflow

03

Ethics

NEXT

INTRODUCTION

What's Web Scrapping ?

- Web Scraping is an automatic way to retrieve unstructured data from a website and store them in a structured format.
- Website scrapers are commonly used for data mining, web research, and competitive analysis.
- However, they can also be used for malicious purposes, such as stealing copyrighted content or personal information.
- Therefore, it is important to respect the website's terms of service and privacy policy.

NEXT

INTRODUCTION

How Does It Work?

- Web scraping just works like a bot person browsing different websites and copy paste down all the contents.
- When you run the code, it will send a request to the server and the data is contained in the response you get.
- What you then do is parse the response data and extract out the parts you want.

INTRODUCTION

Web Scrapper Components

- Web spiders :

Web spiders, also known as web crawlers, are automated software programs that browse the World Wide Web in a systematic, automated manner

- Spider management:

Spider management refers to the process of controlling and managing the behavior of web spiders or crawlers that visit a website.

- Javascript rendering :

JavaScript can be used to dynamically generate content on a website, which can make it more difficult to scrape that content using traditional scraping methods. Allow you to programmatically control a browser, including executing JavaScript on the page and waiting for content to load before scraping it.

- Data quality:

The next building block is data quality assurance. All our scraping efforts are worth it only if the output data is the right data in the correct format.

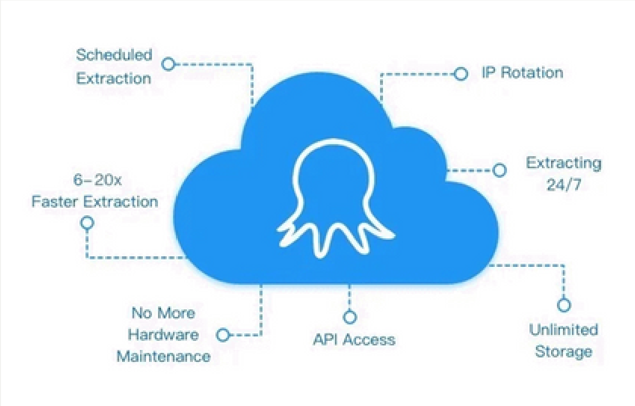
- Proxy management:

Proxy servers can be used in website scraping to help manage IP address restrictions, bypass geographical restrictions, and prevent the target website from blocking or detecting your scraping activity

NEXT

OVERVIEW

Types of web scrapping solutions



- Web Scraper Desktop-based :

EXAMPLE: Octoparse.

Web Scraper

- Web Scraping Plugins and Extensions

EXAMPLE:Webscraper.io

dexi.io

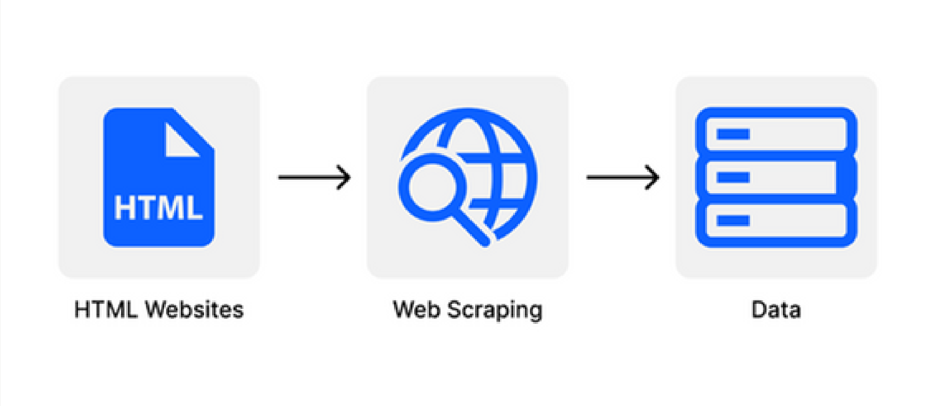
Digital Commerce Intelligence Suite

- Web-based Scraping Applications

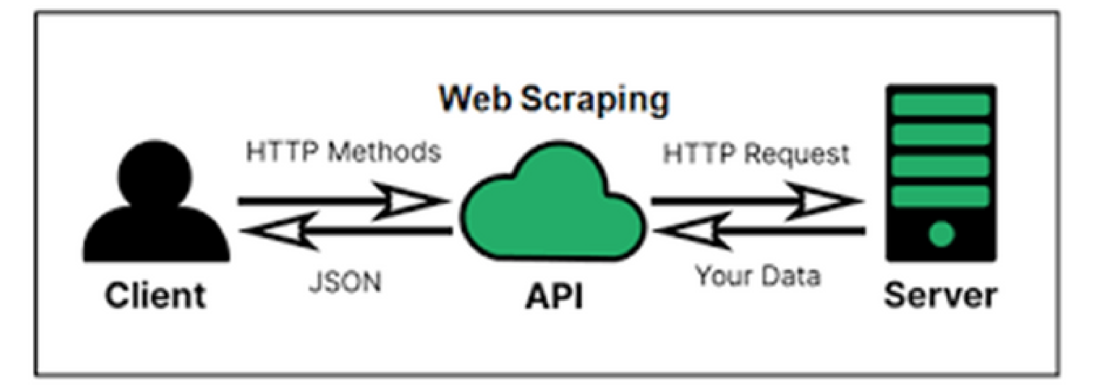
EXAMPLE: dexi.io / webhose.io

- Web scraping using programming languages:(exp: python)

1. Approach 1: HTML method



2. Approach 2: API method



NEXT

OVERVIEW

Tools used in web scrapping

Web scraping projects typically involve a combination of programming languages, libraries, and tools:

- 1. Programming Languages:

Python: Python is a popular language for web scraping due to its simplicity and the availability of various libraries.

- 2. Web Scraping Libraries:

BeautifulSoup: BeautifulSoup is a Python library for parsing HTML and XML documents.

- Selenium:

Selenium is a web testing framework that allows you to automate browser interactions and extract data from websites that heavily rely on JavaScript.

- 3. HTTP Client Libraries:

Requests: Requests is a popular Python library for making HTTP requests. It simplifies the process of sending HTTP requests and handling responses in web scraping projects.

- 4. Data Parsing and Manipulation:

Pandas: It providesconvenient data structures and functions for cleaning, transforming, and analyzing scraped data.

JSON/XML parsers: ElementTree to parse and extract relevant information from JSON or XML responses.

- 5. Proxy Tools:(conditional)

Proxies: In some cases, you may need to use proxies to scrape websites without getting blocked or to bypass certain restrictions.

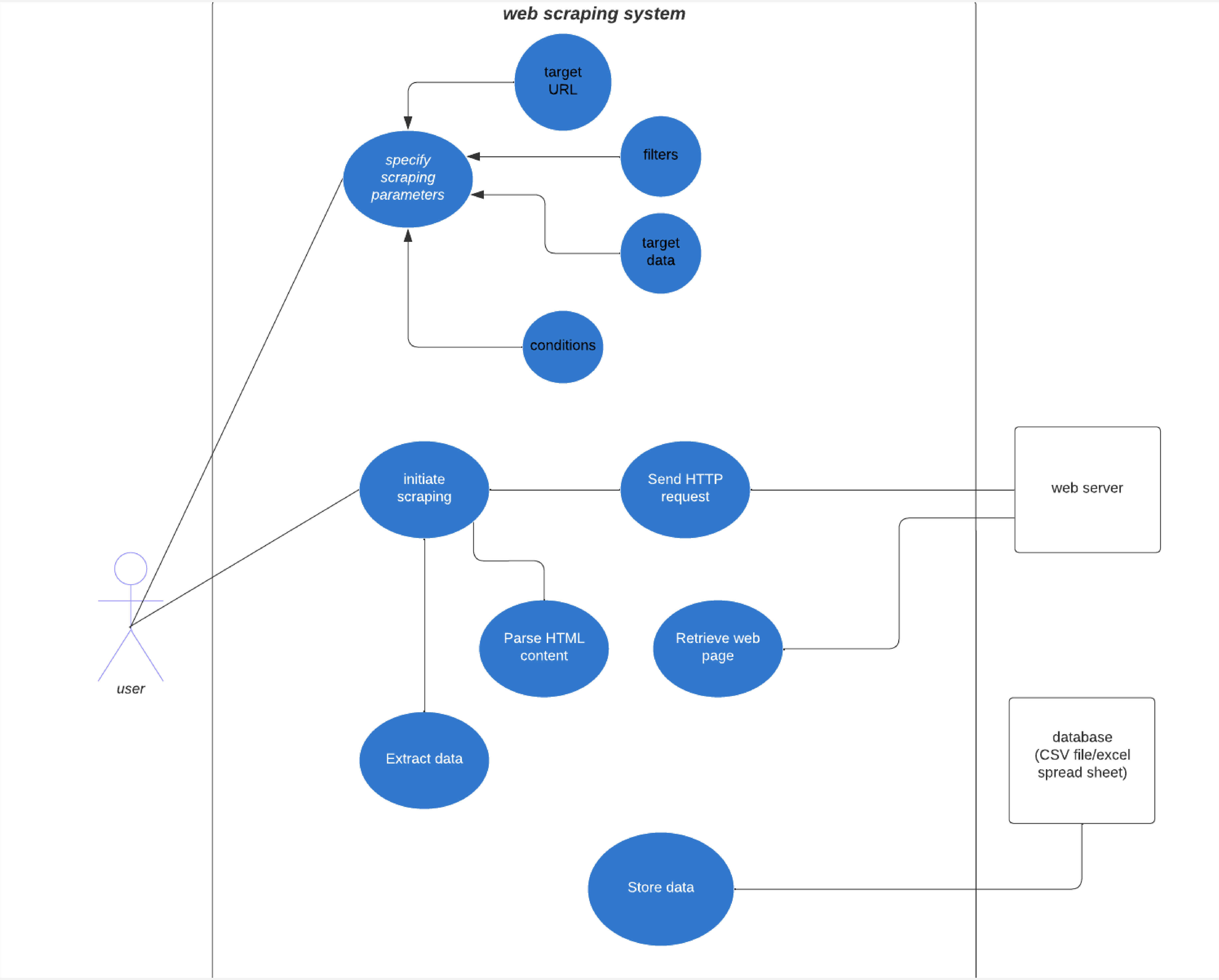
- 6. Data Storage:

Store the scraped data in simple file formats like CSV, JSON, or XML.

NEXT

OVERVIEW

Use Case



NEXT

OVERVIEW

Workflow

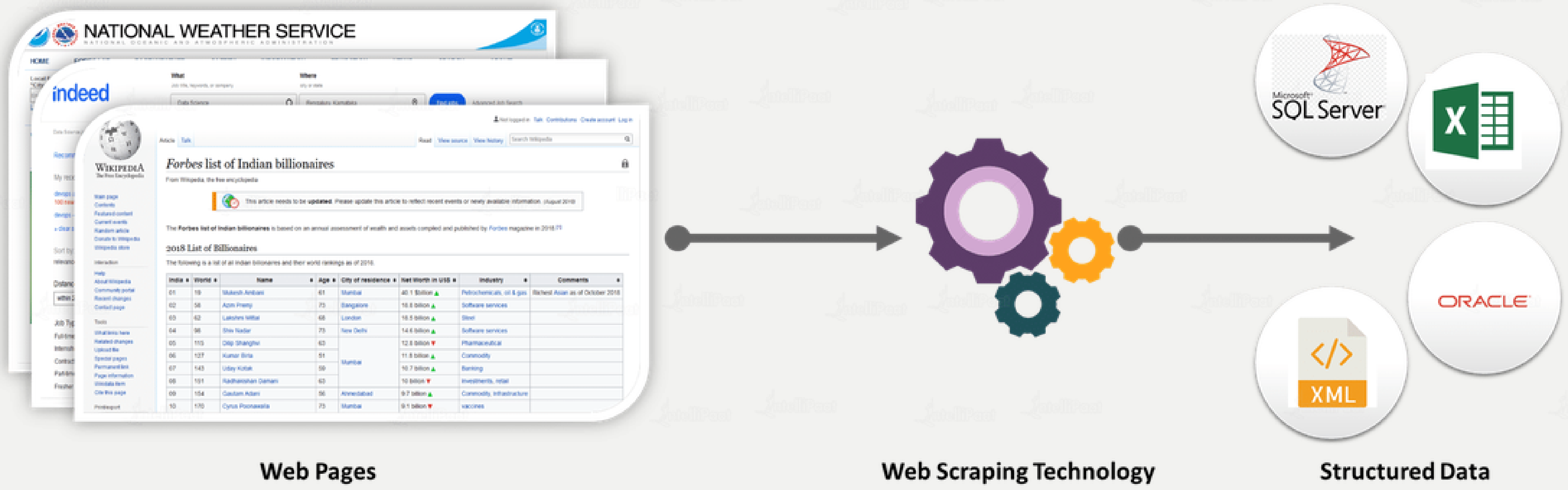
- Can you scrape from all the websites?

Scraping makes the website traffic spike and may cause the breakdown of the website server. Thus, not all websites allow people to scrape. How do you knowwhich websites are allowed or not? You can look at the 'robots.txt' file of the website. You just simply put robots.txt after the URL that you want to scrape and you will see information on whether the website host allows you to scrape the website.

- Web scraping workflow

A web scraping project workflow is commonly categorized into three steps:

1. First, fetch web pages that we want to retrieve data from;
2. Second, apply web scraping technologies,
3. And finally, store the data in a structured form.



NEXT

ETHICS

Ethics

It’s important to talk about web scraping ethics. When
you scrape a website, you have to make sure you also respect it. Here are some best practices you can follow to scrape respectfully.

- Don't be a burden

The most important rule when you scrape a website is not to harm it. Do not make too many requests. Making requests too frequently could make it hard for the website server to serve other visitors.

- Robots.txt

Before scraping, always inspect the robots.txt file first. This will give you a good idea of what parts of the website you are free to visit and what pages you should not.

- User-agent

Define a user-agent that clearly describes you or your company. Also, it’s best to include contact information in your user-agent as well, so they can let you know if they have any issues with what you’re doing.

- Pages behind a login wall

There are cases when you can only access certain pages if you are logged in. If you want to scrape those pages, you need to be very careful. By logging in and/or explicitly agreeing to the website's terms and conditions that state you cannot scrape, then you CAN NOT scrape. You should always honor the terms of any contract you enter into, including website terms and conditions and privacy policies.

Responsible Web Scraping:
gathering web data without becoming Dr. Evil
SOSHACE

NEXT