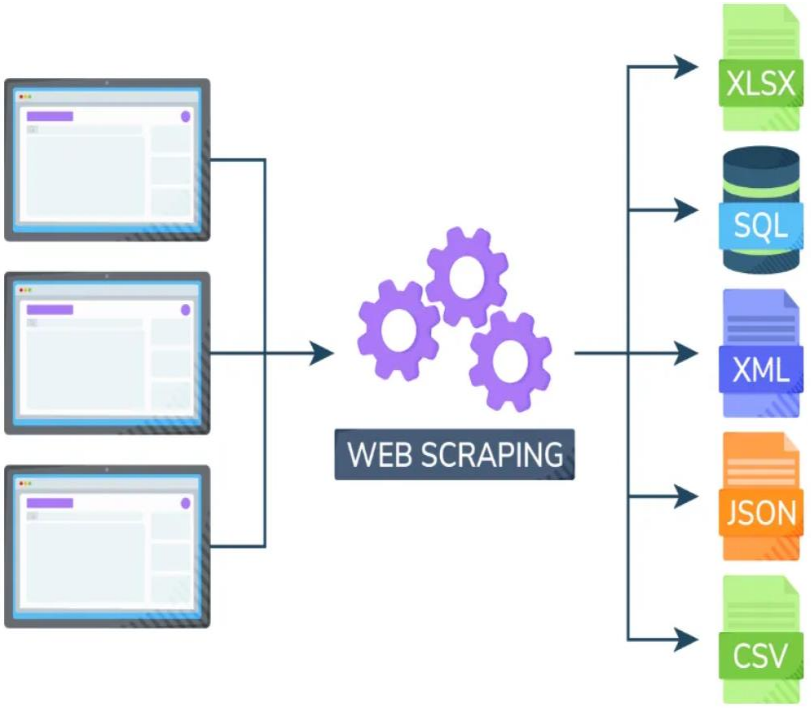
Web Scraping

- **Web scraping** refers to the automatic extraction of data from websites. It is also sometimes referred to as web harvesting.



- **How Does It Work?**
    - Usually, we send multiple HTTP requests to the website we are interested in and then receive the HTML content of the website. This content is then parsed, throwing away irrelevant/unnecessary

content and keeping only the filtered data. It is to be noted that the data can be in the form of text or visuals (images/videos). This process can be done either in a semi-automated way where we copy the data from the website ourselves, or automated, in which we use tools and configure data extraction.

- **How Do Web Scrapers Work?**

    - **1.HTTP requests:** The web scraper commences by sending an HTTP request to a designated URL, with the objective of retrieving the web page's content. This procedure mirrors the way a web browser fetches a web page.

    - **2.Acquiring HTML:** The server hosting the website responds to the request by transmitting the HTML content of the web page. This HTML code encompasses all components like text, images, links, and other elements constituting the web page.

- **3.HTML parsing:** Subsequently, the web scraper engages in HTML parsing, a process of analysing and interpreting the HTML content to locate sections of the web page containing the desired data. This entails utilizing tools like HTML parsing libraries to navigate the structural aspects of the HTML code.

- **4.Data extraction:** Once the pertinent segments of the HTML are pinpointed, the scraper proceeds to extract the targeted data.

- **5.Data cleansing:** Depending on the quality of the HTML code and the page's structure, the extracted data might necessitate cleaning and formatting. This phase involves eliminating extraneous tags and special characters, ensuring that the data is formatted in a usable manner.

- **6.Data storage:** After the cleansing phase, the cleaned data can be organized into a structured format. This could involve storing the data in mediums like CSV files, databases.

- **7.Iterating through pages:** In cases where the scraper needs to accumulate data from multiple pages (such as scraping search results), it iterates through the process by sending requests to distinct URLs, extracting data from each individual page.

- **8.Handling dynamic content:** Websites employing JavaScript to load content dynamically subsequent to the initial HTML retrieval necessitate more sophisticated scraping techniques. This involves utilizing tools like a headless browser or resources like Selenium to interact with the page as a user would, thereby extracting dynamically loaded content.

- **9.Rate limiting:** To avert overwhelming a website's server with an excessive number of requests in a short span, the scraper might integrate rate-limiting mechanisms. These mechanisms are designed to ensure responsible and restrained scraping.