

Ethical Web Scraping

Python Web Scraping

Introduction

Web scraping is an **automated** technique used to **extract data** from websites. It involves using scripts, software, or bots to crawl web pages, collect specific information, and store it in a structured format for further analysis or use.

Technical Synonyms:

- Web Data Extraction
- Web Crawling
- Data Mining

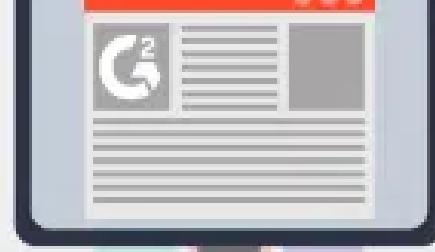
Web Scraping Process

1. Sending an HTTP Request
2. Receiving and Parsing the Response
3. Extracting the Desired Data
4. Storing the Data
5. Automating and Scaling

HTML
website

—
Web
scraping

—
Extracted
data



Where is Web Scraping Skills Needed?

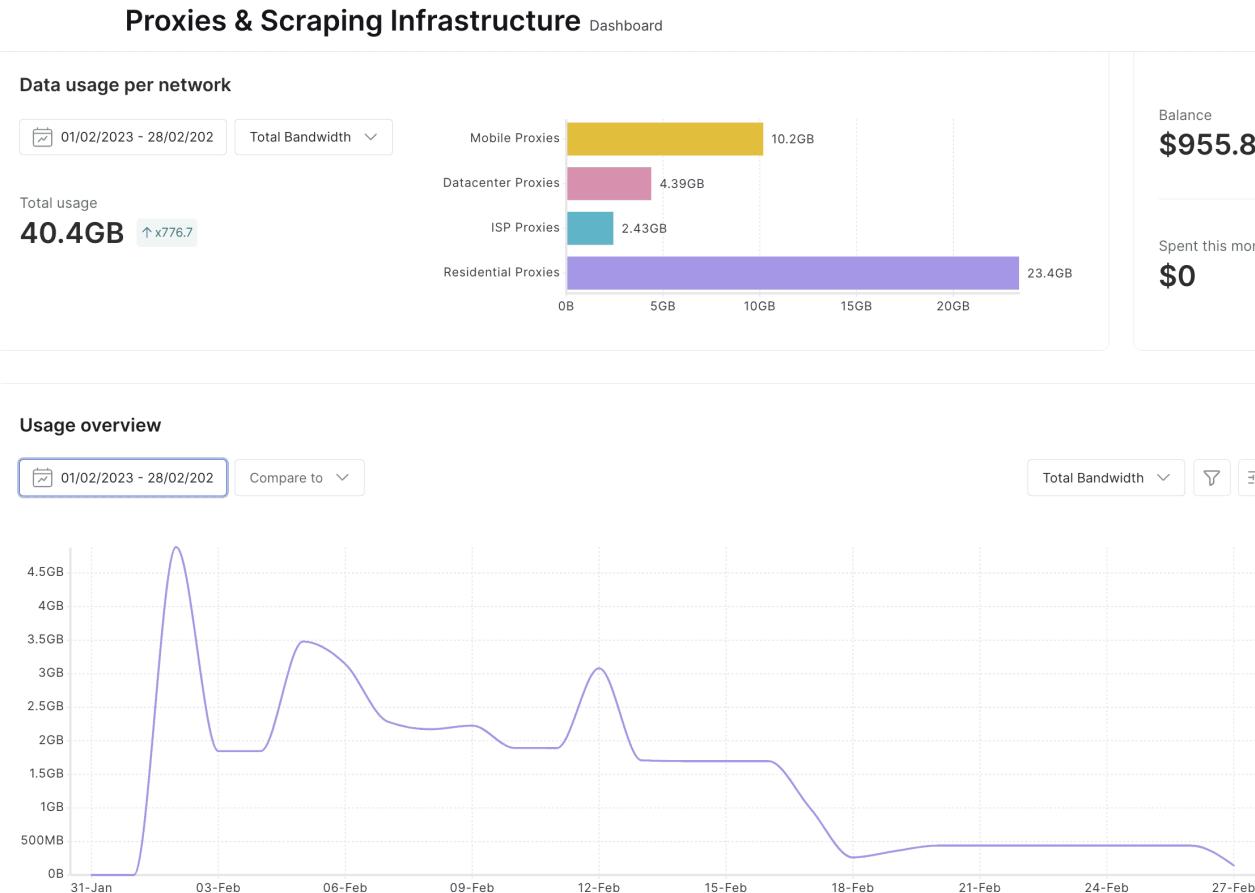
1. E-Commerce & Retail
2. Finance & Investment
3. Digital Marketing & SEO
4. Real Estate
5. Media & Journalism
6. Travel & Hospitality

Web Scraping Service Provider

Octoparse – A no-code desktop application that enables non-developers to scrape data from websites using a visual, drag-and-drop interface. Best for beginners and small-scale projects.

The screenshot shows the Octoparse desktop application interface. At the top, there's a header with 'Octoparse Demo Site' and various control buttons like 'Show Browser', 'Task Settings', 'Save', and 'Run'. Below the header, a central panel displays a circular progress bar with the number '18' and the status 'Running'. It also shows statistics: 'Duplicates: 0 line(s)', 'Time Spent: 34s', and 'Avg. Speed: 32 lines/min'. Below this, a 'Data List' table shows 18 rows of scraped data with columns for '#', 'Title', 'Title_URL', 'Description', 'Time', 'Title1', 'Text', and 'Field'. The first row of data is: #8, Title: 'A Full Guide on Scrapi...', Title_URL: 'https://demo.octopars...', Description: 'Are you sometimes tro...', Time: '2022-08-05 16:00', Title1: '8. Lorem Ipsum', Text: '8. This is a dummy t...', Field: 'Author: Octoparse'. To the right of the data list is a 'Visual Workflow Editor' window showing a step-by-step process: 'Go to Webpage' → 'Pagination' → 'Loop Item' → 'Extract Data' → 'Click URLs in th...' → 'Extract data on th... ***' → 'Click to Paginate'. At the bottom left, there's a 'Data Preview' section showing a single row of data: No. 1, Title1: '1. Lorem Ipsum', Text: '1. This is a dummy text article for testing purpos...', and Field: 'Author: Octoparse'. The bottom right has an 'Apply' button.

BrightData – A robust proxy and web scraping service offering powerful tools for large-scale data extraction, including residential, mobile, and datacenter proxies. Ideal for enterprise-level scraping.



Zyte – A managed web scraping service that provides high-quality, ready-to-use data using AI-powered extraction. It also offers smart proxy solutions and anti-bot bypassing.

The screenshot shows the Zyte Jobs Dashboard interface. At the top, there's a navigation bar with the Zyte logo, a search bar, and user information: "Welcome back, Jane Doe!". Below the navigation is a breadcrumb trail: "Jobs / Dashboard". On the right side of the header, there are buttons for "Watch" and "Run", with "Run" being highlighted by a purple box. The main content area is divided into two sections: "Next" and "Running".
Next Section: Shows 0 results found. It has columns for Job, Spider, Units, and Priority. Buttons for "Cancel" and "Edit" are present.
Running Section: Shows 0 results found. It has columns for Job, Spider, Items, and a status icon. Buttons for "Stop" and "Edit" are present.
On the far right, there's a sidebar titled "Units Breakdown" for the "Default Group". It shows usage statistics:

- This Project: 0
- Others: 0
- Free: 1

Below this, under "Job Status", there are four toggle switches:

- Next: 0
- Running: 0
- Completed: 0
- Deleted: 0

The screenshot shows the Apify web scraping platform interface. On the left is a sidebar with navigation links: Home, Store, Actors (selected), Saved tasks, Runs, Schedules, Storage, Proxy, Settings, and Billing. Below these are Documentation, Help & resources, Memory (0 MB / 8 GB), and Free usage (\$0.07 / \$5.00). At the bottom are Upgrade and Build buttons, with a red arrow pointing to the Build button.

The main area displays the 'my-test-actor' page. It includes a header with the actor name, a personal account dropdown, and buttons for Create task, API, and more. Below the header are tabs for Source, Information, Runs, Builds, Integrations, Monitoring, Issues, Saved tasks, and Settings. A Version dropdown shows '0.0 (latest)'. A 'Publication' section has a 'Tips' button. Below these are tabs for Code, Last build, Input, and Last run. The Source type is set to 'Web IDE'. The code editor contains the following JavaScript code:

```
// Apify SDK - toolkit for building Apify Actors (Read more at https://docs.apify.com/sdk/js/).
import { Actor } from 'apify';
// Web scraping and browser automation library (Read more at https://crawlee.dev)
import { PuppeteerCrawler } from 'crawlee';
// this is ESM project, and as such, it requires you to specify extensions in your relative imports
// read more about this here: https://nodejs.org/docs/latest-v18.x/api/esm.html#mandatory-file-extensions
import { router } from './routes.js';

// The init() call configures the Actor for its environment. It's recommended to start every Actor with an init().
await Actor.init();

// Define the URLs to start the crawler with - get them from the input of the Actor or use a default list.
const input = await Actor.getInput();
const startUrls = input?.startUrls || [{ url: 'https://apify.com' }];

// Create a proxy configuration that will rotate proxies from Apify Proxy.
const proxyConfiguration = await Actor.createProxyConfiguration();

// Create a PuppeteerCrawler that will use the proxy configuration and handle requests with the router from routes.js file.
const crawler = new PuppeteerCrawler({
    proxyConfiguration,
    requestHandler: router,
});

// Run the crawler with the start URLs and wait for it to finish.
await crawler.run(startUrls);

// Gracefully exit the Actor process. It's recommended to quit all Actors with an exit().
await Actor.exit();
```

A note at the bottom of the code editor says: "⚠ To apply your changes, you need to build the Actor. [Build now](#)".

Apify – A cloud-based web scraping and automation platform that allows developers to create, run, and manage web scrapers using "Actors" (custom scraping scripts). Ideal for automation and large-scale data extraction.

Choosing Web Scraping Provider

For no-code beginners → Octoparse

For managed high-quality data → Zyte

For large-scale scraping & proxy needs → BrightData

For coding & automation → Apify

Web Scraping Cases

- X vs Bright Data (2024): X Corp. sued Bright Data for unauthorized data scraping, but the case was dismissed.
- Meta vs Ekrem Ateş (2022): Meta sued Ateş for scraping Facebook & Instagram, resulting in a **permanent ban and legal injunction**.
- HiQ Labs vs LinkedIn (2017-2023): LinkedIn sued HiQ for scraping public profiles, leading to **years of legal battles** over data ownership. **HiQ filed for bankruptcy**, effectively ending the legal battle.
- Aaron Swartz Case (2011): Swartz faced **35 years in prison & \$1M in fines for scraping academic papers from JSTOR**.

Web Scraping Don'ts

1. Don't Scrape Data Without Checking Legal & Ethical Boundaries

Before scraping, always check:

- Robots.txt** – contains instructions for web crawlers and bots, specifying which parts of a website they are allowed or disallowed to access and scrape.
- Privacy Laws (GDPR, CCPA, etc.)** – contain provisions that regulate how personal data can be collected, stored, and used, which directly impacts web scraping activities.

2. Don't Scrape Personal or Sensitive Data

Avoid collecting:

- ✗ Usernames, passwords, or login credentials
- ✗ Private emails or phone numbers
- ✗ Medical records or financial data

Avoid websites:

- ✗ Betting/Gambling Websites
- ✗ Adult Content Websites (Pornographic Sites)
- ✗ Payment Platforms

3. Don't Overload a Website's Server

Sending too many requests too fast can:

- ✗ Crash the website (denial of service).
- ✗ Get your IP banned.
- ✗ Harm the website owner's business.

Best Practice:

- ✓ Use time delays.
- ✓ Implement **rate limiting & exponential backoff**.

4. Don't Ignore API Alternatives

Many platforms provide **official APIs**, which are:

-  Legal & reliable
-  Faster and structured
-  Less likely to get blocked

5. Don't Scrape Paywalled or Copyrighted Content

 Avoid scraping **premium articles, academic papers, or licensed content.**

-  Instead, check for **public datasets or official APIs.**



Web Scraping Career Path

If you're interested in:

- **Web Development**, a natural progression would be to explore roles like **backend developer**. You can focus on building scalable web scraping systems, handling data processing, and integrating APIs to collect valuable information from websites.
- **Data and analysis**, you can aim to become a **data engineer** or **data scientist**. In this role, you'll take the data you scrape and clean it, then analyze it for trends or build predictive models using tools like Pandas and machine learning.

- If **automation** excites you, a great option is to dive into **automation engineering**. You'll work on streamlining web scraping processes and building automated workflows to collect data continuously without manual intervention, using tools like Selenium or Puppeteer.
- If you're leaning toward **business and marketing**, you can look into **lead generation, market research** or becoming a **web scraping consultant**. In these roles, you'll apply web scraping to gather data for market research, competitor analysis, and lead generation. You'll also need to ensure that the scraping is compliant with privacy laws like GDPR.

Class Activity

List all the possible challenges you may encounter when performing web scraping, along with potential solutions. (15 minutes)

Reflection

Apart from the existing web scraping companies, how do you plan to compete with them?

Exploration

Apify offers a platform where you can provide web scraping and automation services to clients worldwide. As a freelancer, you can build and sell custom scrapers, automate data extraction tasks, and collaborate with businesses in need of structured data. Check out more details and get started here: [Apify Freelancer Program](#).

Thank you

Any Question?