THE TOP FREE AND PAID WEB SCRAPING TOOLS AND SERVICES

“BAD” BOTS HAVE BECOME THE 21ST-CENTURY PLAGUE OF THE INTERNET. THEY MASQUERADE AS HUMANS AND EVADE CONVENTIONAL SECURITY MEASURES TO ATTACK WEB APPLICATIONS AND SITES. THEY PILLAGE PERSONAL DATA, TIE DOWN ONLINE INVENTORY AND DEGRADE APPLICATION/WEBSITE PERFORMANCE.

Web scraping has emerged as one of the predominant uses of bots. Although a percentage of web scrapers are used for legitimate business reasons (such as evaluating SEO rankings), many are now leveraged by competitors to undermine a competitive advantage or to steal your information. The programs that execute these content-stealing attacks are as diverse as they are popular. They range from simple, manually tuned scripts to highly automated, cloud-based services that use machine learning and automation to scrape more advanced websites and bypass bot mitigation tools like CAPTCHA. Other tools serve a supporting role.

This piece provides an overview of the top web scraping tools, cloud-based services and IP rotation solutions currently used to conduct web scraping attacks.

TOP WEB SCRAPING AND CRAWLING TOOLS

Web scraping and crawling tools are programs or automated scripts that browse websites and fetch new or updated information and store it and store it for easy access. Extracted data can typically be exported into a structured format including, but not limited to, Excel, HTML, CSV, etc.

Here are the top 15 web scraping tools currently available for use:

	NAME	DESCRIPTION	DATA EXPORT
1	DataMiner	DataMiner is a web scraping tool that is an extension for Google Chrome. It provides basic web scraping capabilities and can scrape data from webpages and export the data.	CSV or Excel
2	Scrapy	Scrapy is an open-source web scraping framework in Python used to build web scrapers. It provides users with tools to extract data from websites, process them and store them in your preferred structure and format. It's built on top of a Twisted asynchronous networking framework.	You can export data into JSON, CSV and XML formats.
3	Data Scraper	Data Scraper is a simple web scraping tool for extracting data from a single page. It is a personal browser extension that helps users transform data into a clean table format.	CSV and XSL data files
4	Scraper	Scraper is a Chrome extension for scraping simple webpages. It is simple to use and will help users scrape a website's content and upload the results to Google Docs. It can extract data from tables and convert it into a structured format.	Google Docs
5	ParseHub	ParseHub is a more advanced web-based scraping tool, which is built to crawl single and multiple websites that are using JavaScript, AJAX, cookies, sessions and redirects. The application can analyze and grab data from websites and transform it into meaningful data. It uses machine learning technology to recognize the most complicated documents.	Output file in JSON, CSV or Google Docs
6	Outwit Hub	Outwit Hub is a free data extractor built in a web browser, which is available as an extension or stand-alone application.	You can export the data into numerous formats (JSON, XLSX, SQL, HTML, CSV, etc.).
7	FMiner	FMiner is a visual web data extraction tool for web scraping and web screen scraping. In addition to the basic web scraping features, it also has AJAX/JavaScript processing and CAPTCHA solving. It can be run on both Windows and Mac OS, and it leverages the internal browser to accomplish the scraping.	Data can be saved into JSON and CSV formats.
8	Dexi.io	Dexi.io supports data collection from any website and provides different types of robots to scrape data — crawlers, extractors, autobots and pipes. The application offers anonymous proxies to hide a user's identity. Dexi.io also offers a number of integrations with third-party services.	CSV or Excel

TOP WEB SCRAPING AND CRAWLING TOOLS

	NAME	DESCRIPTION	DATA EXPORT
9	Octoparse	Octoparse is a visual scraping tool with a point-and-click interface that allows users to choose the fields they want to scrape from a website. The web scraper can handle both static and dynamic websites with AJAX, JavaScript, cookies, etc. The application also offers a cloud-based platform that allows you to extract large amounts of data.	You can export the scraped data in TXT, CSV, HTML or XLSX formats.
10	WebHarvy	WebHarvy is a visual web scraper that has a built-in browser that allows users to scrape data from webpages and provides minimal to no coding. It has a multilevel category scraping feature that can follow each level of category links and scrape data from listing pages.	The data can be saved into CSV, JSON and XML files or stored to a SQL database.
11	PySpider	PySpider is a web crawler written in Python. It supports Javascript pages and has a distributed architecture, so users can have multiple crawlers. PySpider can store the data on a back-end of your choosing such as MongoDB, MySQL, Redis, etc. You can use RabbitMQ, Beanstalk and Redis as message queues.	Data can be saved into JSON and CSV formats.
12	Apify	Apify is a Node.js library similar to Scrapy and positions itself as a universal web scraping library in JavaScript, with support for Puppeteer, Cheerio and more. With its unique features, like RequestQueue and AutoscaledPool, a user can start with several URLs and then recursively follow links to other pages and can run the scraping tasks at the maximum capacity of the system. It supports any type of website and has built-in support for Puppeteer.	Its available data formats are JSON, JSONL, CSV, XML, XLSX or HTML and available selector CSS.
13	Content Grabber	Content Grabber is a visual web scraping tool that has a point-to-click interface, which allows pagination, infinite scrolling pages and pop-ups. In addition, it has AJAX/JavaScript processing and a CAPTCHA solution and allows the use of regular expressions and IP rotation (using Nohodo). Intermediate programming skills are needed to use this tool.	You can export data in CSV, XLSX, JSON and PDF formats.
14	Mozenda	Mozenda is an enterprise cloud-based web scraping platform. It has a point-to-click interface. It comprises two parts: an application to build the data extraction project and a web console to run agents, organize results and export data. They also provide API access to retrieve the data and has built-in storage integrations such as FTP, Amazon S3, Dropbox and more. It is good for handling large volumes of data but requires more advanced coding capabilities to build a web scraper.	You can export data into CSV, XML, JSON or XLSX formats.
15	Cheerio	Cheerio is a library that parses HTML and XML documents and allows users to use the syntax of jQuery while working with the downloaded data. If crafting the web scraper in JavaScript, Cheerio is a fast option that makes parsing, manipulating and rendering efficient. It does not interpret the results as a web browser, produce a visual rendering, apply CSS, load external resources or execute JavaScript.	CSV and XML data files

TOP WEB SCRAPING CLOUD SERVICES AND PROVIDERS

Web scraping cloud-based services provide a simplistic solution for “self-service” scraping. Some provide a hosted platform for building a custom web scraper, while others offer fully managed services with customer support, product guides, etc.

Here are the top 10 web scraping tools currently available for use:

	TOOLS	DESCRIPTION	DATA EXPORT
1	Import.io	Import.io provides easy-to-use web data extraction for price monitoring, lead generation, market research, big data analysis and more. Users can clean, transform and visualize the data and can build a scraper using a web-based point-and-click interface. Like Diffbot (see below), Import.io can handle most of the data extraction automatically.	File Formats – CSV, JSON, Google Sheets Integrates with many cloud services Import.io APIs (premium feature)
2	Webscraper	Webscraper.io Cloud scraper is an online platform where a user can deploy scrapers using the free point-and-click Webscraper.io Chome Extension. Using the extension, users can create “sitemaps” that show how the data should be traversed and extracted.	CSV or CouchDB
3	Octoparse	Octoparse Cloud Service offers a cloud-based platform for users to run their extraction tasks built with the Octoparse desktop app.	File Formats – CSV, HTML, XLS and JSON Databases – MySQL, SQL Server, Oracle Octoparse API
4	ParseHub	ParseHub lets users build web scrapers to crawl single and multiple websites and supports JavaScript, AJAX, cookies, sessions and redirects using their desktop application and deploys them to their cloud service. ParseHub provides a free version where users receive 200 pages of data in 40 minutes, five public projects and limited support.	File Formats – CSV, JSON Integrates with Google Sheets and Tableau ParseHub API
5	Scrapy Cloud (Scrapinghub)	Scrapy Cloud is a hosted, cloud-based service by Scrapinghub, where users can deploy scrapers built using the Scrapy framework. Scrapy Cloud removes the need to set up and monitor servers and provides a nice UI to manage spiders and review scraped items, logs and stats.	File Formats – CSV, JSON, XML Scrapy Cloud API Write to any database or location using item pipelines

TOP WEB SCRAPING CLOUD SERVICES AND PROVIDERS

TOOLS	DESCRIPTION	DATA EXPORT
6 Dexi	Dexi.io is similar to ParseHub and Octoparse, except that it has a web-based point-and-click utility instead of a desktop-based tool. It lets users develop, host and schedule scrapers like many other services.	File Formats – CSV, JSON, XML Can write to most databases through add-ons Integrates with many cloud services Dexi API
7 Diffbot	Diffbot lets users configure crawlers that can go in and index websites and then process them using its APIs for automatic data extraction from various web content. Users can also write a custom extractor if an automatic data extraction API doesn't work for the websites being targeted.	File Formats – CSV, JSON, Excel Cannot write directly to databases Integrates with many cloud services through Zapier Diffbot APIs
8 SunTec	SunTec is a more advanced web scraping service that utilizes tools to ensure that data can be mined from sites that leverage JavaScript, MooTools frameworks, etc. It also provides capabilities to navigate sites that leverage CAPTCHA to thwart bots.	File formats — CSV, HTML, XLS and JSON
9 Connotate	Now a part of import.io, this tool provides two versions: either as a software-as-a- service solution or as a fully managed service. Its advanced features include the ability to handle dynamic sites built with JavaScript and AJAX, and it is capable of creating agents by simply browsing websites. It also leverages machine learning.	Once extracted, data can be exported into an array of different formats within the product.
10 Agenty	Agenty is a software-as-a-service company that offers point-and-click web scraping tools for users who want to web scrape with little to no coding knowledge.	File formats — CSV, HTML, XLS and JSON

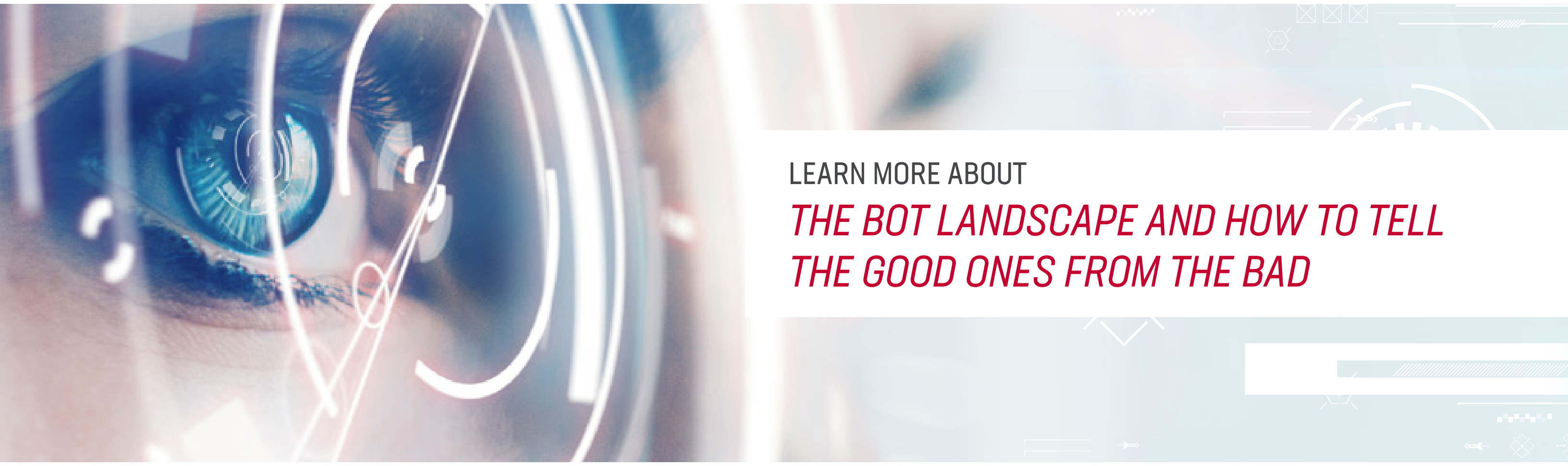
IP ROTATION TOOLS

To stop web scraping, many websites block traffic from certain IP addresses. IP rotation proxy tools and services are the workaround. Residential proxies, backconnect proxies, rotating proxies or other IP rotation proxy services avoid having block scrapers blocked.

Here are the 10 most popular residential and backconnect rotating proxy services right now:

NAME		DESCRIPTION
1	Scrapoxy	Scrapoxy hides a web scraper using its cloud-based solution by starting a pool of proxies to send requests. It is an open-source solution and integrates with Amazon AWS/EC2.
2	ProxyMesh	ProxyMesh provides 15 rotating IP addresses for each proxy server, each with 10 IP addresses rotating twice daily, giving the user a total of 300 IP addresses per day. When you make a request through one of these 15 anonymous proxy servers, your request will be randomly routed through one of 10 different proxy IP servers.
3	Luminati	Luminati is one of the world's largest proxy services, which provides an array of services for enterprise data centers, residential networks and mobile networks.
4	Oxylabs.io	OxyLabs Proxy Rotator is an IP address rotator meant for enterprise-level data mining activities. It eliminates exposure to IP-based blocking by misleading data representation or CAPTCHA coming from targeted websites.
5	Squid Proxy	Squid Proxy is a caching and forwarding HTTP web proxy. It has a wide variety of uses, including accelerating a web server by caching repeated requests and caching web, DNS and other computer network lookups. It provides both proxy and IP address rotation.
6	ProxyRack	ProxyRack offers rotating, residential rotating and datacenter rotating proxies. The company has worldwide server locations in over 40+ countries and provides access to 1,200,000+ IPs monthly. It provides multifunction rotating ports and randomly selected new IPs on each connection port and uses both HTTP(S) and SOCKS5 protocols.
7	Proxy-Connect	This service offers an automatic worldwide IP rotation/backconnect proxy with public HTTP proxy servers ideal for web data scraping/extraction tools, data mining applications and SEO proxy tools such as XRumer, SEnuke, GSA SER, GScraper, ScrapeBox, Hrefer and any other web scrapers.
8	ParseHub	ParseHub is a cloud-based web scraper that provides IP rotation functionality as part of a broader product solution, which offers a GUI-based tool for extracting data, images, text and more.
9	GeoSurf	GeoSurf provides an unblockable proxy network for businesses, which enables users to access over 2 million IP address across 130 global locations.
10	Proxy Rotator	The Proxy Rotator network consists of millions of proxies to provide geographic coverage globally for the majority of cities. The network embodies anonymous and elite proxies as well as a balanced mixture of residential, private and public IP addresses for the best experience.

Malicious bots now comprise nearly 25% of total internet traffic, of which web scrapers are a significant percentage. They represent one of the fastest-growing threats to your website. In addition to understanding the aforementioned product landscape, stopping them requires an overall understanding of the evolution of bots and the various options available for detecting and mitigating them.



LEARN MORE ABOUT

THE BOT LANDSCAPE AND HOW TO TELL THE GOOD ONES FROM THE BAD

About Radware

Radware® (NASDAQ: RDWR) is a global leader of cybersecurity and application delivery solutions for physical, cloud and software-defined data centers. Its award-winning solutions portfolio secures the digital experience by providing infrastructure, application and corporate IT protection and availability services to enterprises globally. Radware's solutions empower more than 12,500 enterprise and carrier customers worldwide to adapt quickly to market challenges, maintain business continuity and achieve maximum productivity while keeping costs down. For more information, please visit www.radware.com.

Radware encourages you to join our community and follow us on: Radware Blog, LinkedIn, Facebook, Twitter, SlideShare, YouTube, Radware Connect app for iPhone® and our security center DDoSWarriors.com that provides a comprehensive analysis of DDoS attack tools, trends and threats.

This document is provided for information purposes only. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law. Radware specifically disclaims any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. The technologies, functionalities, services or processes described herein are subject to change without notice.