How To Prevent Scraper Bots:
A guide for retailers

NETACEA

WHAT IS WEB SCRAPING?

A scraper bot or web scraper is a bot trying to precure, aggregate and parse data, publicly available or otherwise, from an internet-enabled source.

Many scrapers take copies of content you have made freely available on your website, faster than any human is able. Scrapers are vital to business success, but some attempt to access more private data which may be on your site without you realizing. Most of the time scrapers are not inherently bad; the likes of Googlebot and Duolingo extract data from websites for good or benign reasons. Good bot activity includes content aggregation for display on aggregation sites or content scraping by affiliates to help them market your products and services.

Common good or benign reasons for web scraping:

- Search engine indexing (bots that crawl your website and put different pages of your site on Google)
- News aggregation
- Pricing intelligence (price floors are set by manufacturers and pricing intelligence will alert the manufacturer if your retail price drops below the pricing floor)
- Sentiment aggregation
- Investment decision making
- Predictive analysis

However, for all the good scraper bots, there are bots that scrape your site for malicious purposes. Malicious web scraping can cause a business to suffer severe financial losses if the data is extracted without consent. Two frequently used methods of malicious web scraping are price scraping and content theft.

Why we block bad bots:

- **Scrapers can be overly aggressive.** While websites can scale for this, server stress can cause downtime leading to poor user experience and brand damage.
- **Scrapers use up precious resource.** Each unit of compute that scraper bots take up is compute not consumed by your users – meaning no profit. Paying CDNs and analysis tools per traffic levels means paying for bots that aren't doing anything for your site.
- **Scrapers can share private information.** Where scrapers are concerned, information security becomes a risk. Scrapers make public the data on your site intended for private use, as they scrape and utilize anything available. This information then goes to humans looking for intelligence that gives them an advantage over your business.

POPULAR SCRAPING TECHNIQUES IN ECOMMERCE

Price scraping

Price scraping is a technique used to extract pricing data from websites where bots search, find and copy information.
This price monitoring technique is used to track valuable information, notably on eCommerce websites. Scraper bots target the pricing information of competing businesses to undercut rivals and increase their own sales. By doing this, competitors can attract price-sensitive buyers by setting their own prices lower than others in the market. Price scraping bots allow adversaries to maintain real-time monitoring over an entire product catalogue, altering their own pricing accordingly. As price scraping is a common problem across the industry, many eCommerce websites use tools to prevent this.

Content scraping

Content scrapers are automated bots that steal and aggregate content from websites and mobile apps for their own use. Content scrapers typically copy all the content from a webpage and portray it as their own content, including public website information such as text, images, HTML and CSS code. Using sophisticated techniques, content scraper bots can illegally send a series of HTTP requests to the website to be copied. The bots gather content like journalism or paid-for data in a matter of seconds to be used elsewhere without consent; for eCommerce sites this could be thousands of product pages.

Content scraping is also used as a way for attackers to gain the relevant resources to go on and clone a website as part of a wider business logic attack. Cloning is when an adversary copies a website to create a replica of said site. This is generally used when the attacker is trying to impersonate a legitimate site for illegitimate purposes.

According to the BLADE framework:[1]

Sources:
1. BLADE framework

WHAT'S THE RISK TO ECOMMERCE BUSINESSES?

Price scraping is one of the most costly bad bot threats to online retailers. It can jeopardize the overall security of eCommerce websites, customer loyalty and brand reputation. Scraping is often used to gather prices and product information from retail websites to allow competitors to undercut prices and offers, driving customers and profits away from the target websites. The more successful the retailer, the more likely they are to become a target of scraping.

Skewed analytics

With potentially over half of your website traffic coming from bots, monitoring your genuine traffic levels becomes difficult. Unless you have visibility over how much of your traffic is automated, scraper bots causing high volumes and velocity of traffic will lead to misinformed business decisions as a result of skewed analytics. Scraping can also affect SEO and web authority rankings as copied content can outrank the original owner's site on Google.

Brand damage

Losing sales and suffering a slow site – or worse, downtime – can lead to poor customer experience and potential brand damage. Losing customer data results in fines from data protection organizations, whilst dealing with the repercussions of customers leaving for competitors can be costly to rectify in marketing and PR expenses.

Lost sales

Price scraping is a significant threat to retailers as their pricing strategies are exposed to competitors. When scrapers overwhelm an eCommerce website and cause slowdowns for genuine customers, it can lead to abandoned shopping carts as users give up and head to competing websites for a faster, cheaper deal.

Slow site

Serving requests to these bots uses up server resources, which can slow down or even crash a website, as well as pushing up infrastructure costs significantly for no commercial benefit. Of course, a completely crashed site can't function at all. That means no sales until the issue is fixed, plus the added cost of fixing the issue, which could be costly in hours, too, especially during crucial retail events like Black Friday.

HOW TO PREVENT MALICIOUS WEB SCRAPING WITH SOPHISTICATED BOT MANAGEMENT

Being aware is the first step

Understand that scrapers exist and of what they are capable. Monitoring how much of your website traffic is scraper bots is crucial to forming an effective detection and mitigation strategy. Put a plan in place to deal with high traffic spikes – such as using a scalable platform or a virtual waiting room tool – so your website can cope with an overflow of traffic to other servers.

Use sophisticated scraper bot management

The complexity and range of web scrapers hitting eCommerce websites means that we need to look at more than just the behavior that indicates a visitor is carrying out scraping activity, such as the frequency of requests, or whether they identify themselves as a Googlebot.

Netacea's Intent Analytics™ engine uses advanced machine learning techniques to detect scrapers and categorize them based on the scraping activity, for instance, the information they are collecting and the patterns emerging in the collection methods.

Once the activity has been successfully identified, to prevent further web scraping we combine information about the unique attack with data from a wide range of industry sources. This adds an additional layer of insight to the activity categorization and allows us to successfully establish appropriate bot management policies.

Protect your website and revenue from the threat of scraping with Netacea's advanced bot detection technology

Visit netacea.com/impact-of-bots-calculator to find out how much scraper attacks are costing your business.