evolvingweb

Surviving the Swarm: Bot and Crawler Protection for Drupal Sites

2026.03.13

Rich Lawson

Associate Director of Technology,
Evolving Web

/in/rich-lawson

evolvingweb.com

Higher Ed

Healthcare

Culture & Tourism

Transit & Infra

Financial

Government

evolvingweb

Our clients exist to make an impact.

We exist to help them do just that.

- A top 10 Drupal contributor
- Diamond Certified Drupal Partner
- Part of the Drupal CMS leadership team
- Pantheon's Partner of the Year in 2025
- 5,000+ Drupalists trained since 2011
- Organizers of the EvolveDigital event series

New York City | September 2026

See you in NYC?

Get 35% off tickets:
NJPromo35

EvolveDigital

evolvedigital.com

New York City | September 2026

See you in NYC?

Get 35% off tickets:
NJPromo35

EvolveDigital

evolvedigital.com

Agenda

01. When the swarm hits
02. The current bot landscape
03. What happens to Drupal?
04. Layered defense strategy
05. Case study
06. Actionable playbook
07. Questions

01

When the swarm hits

What do you do?

Editors encounter 503 errors when trying to update their landing pages

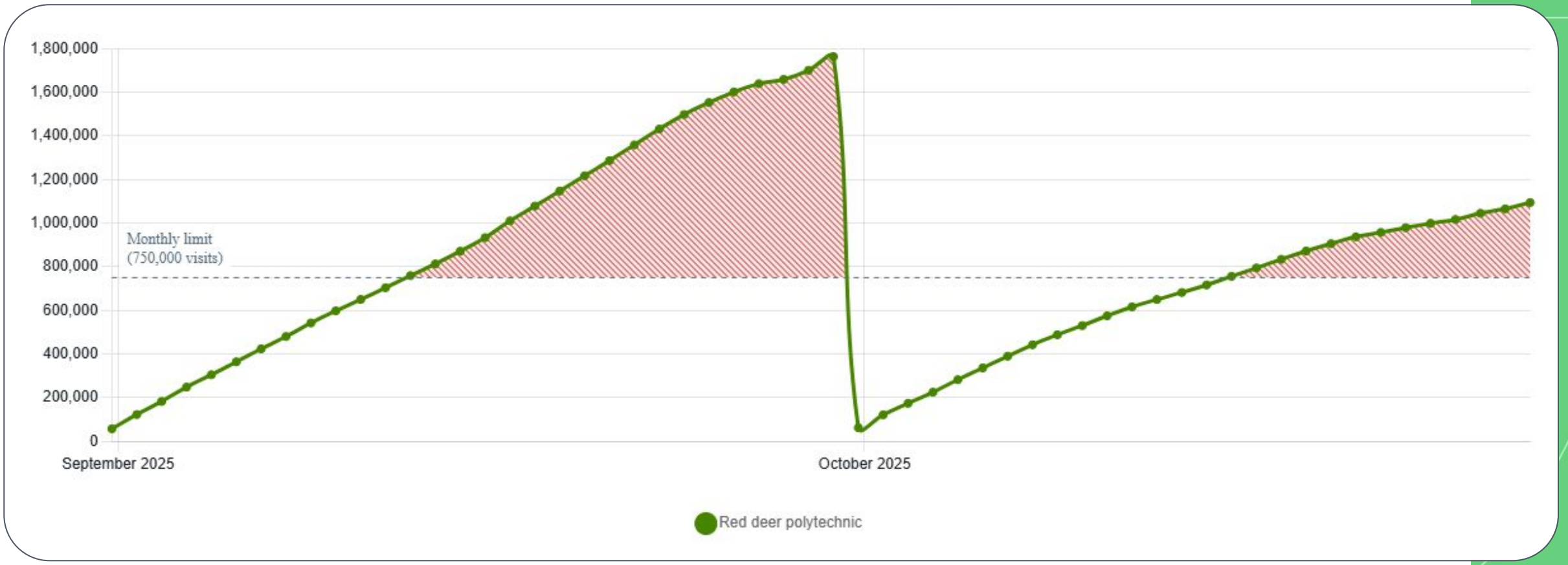
Page loading slows to a crawl, until finally the pages won’t load at all

Analytics show a traffic spike of 800%

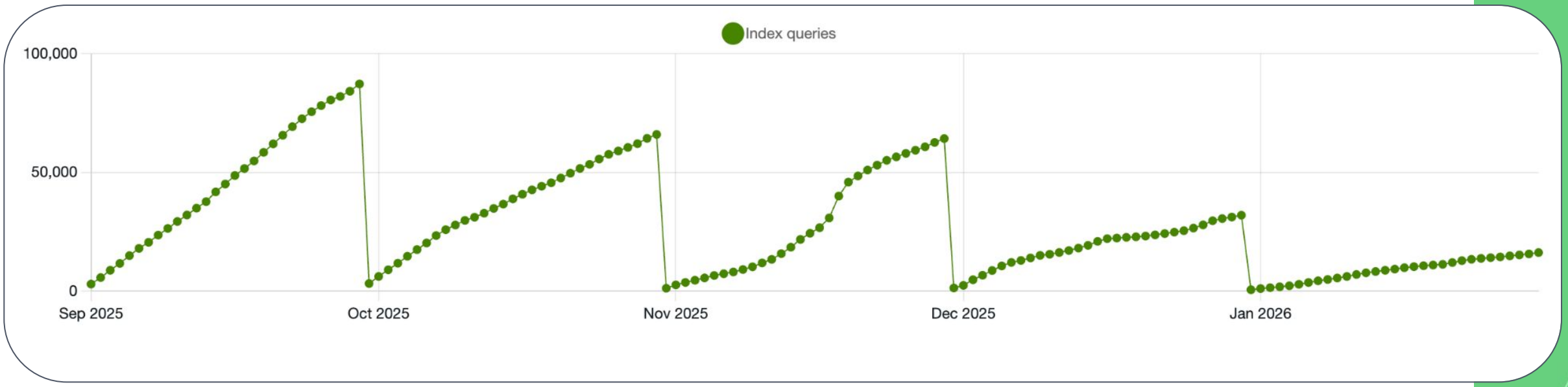
Some of our affected clients

Client	Brief description	Timeline
Sollio	Large agri-food cooperative	Started in April, resolved in December
INSPQ	Government public health institute	Started in October, ongoing
Canada Lands Company	Canadian government Crown corporation	Started in September, resolved in November
Red Deer Polytechnic	Public educational institute	Started in July, resolved in September...or was it?

Traffic growth

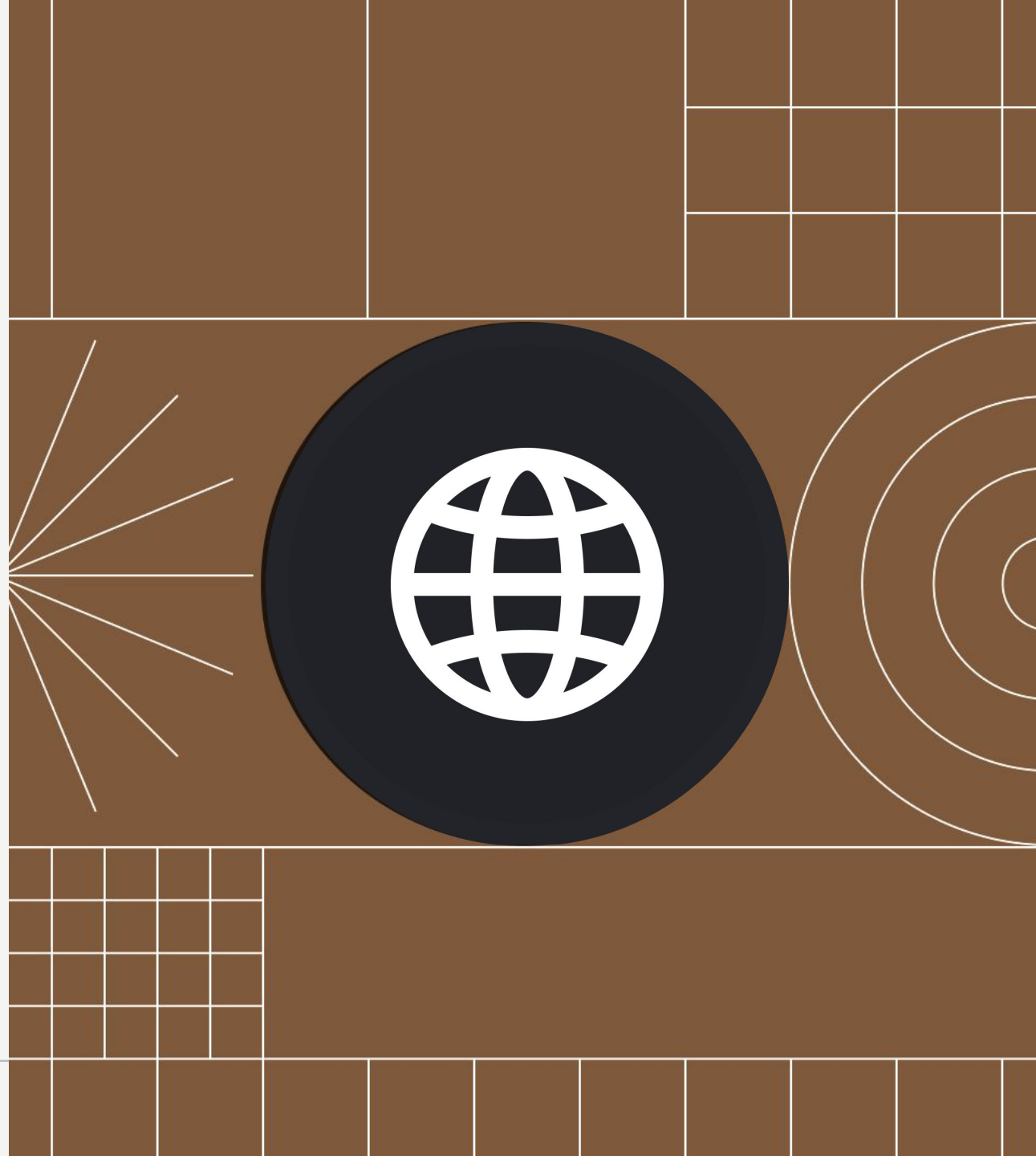


Search index mitigation

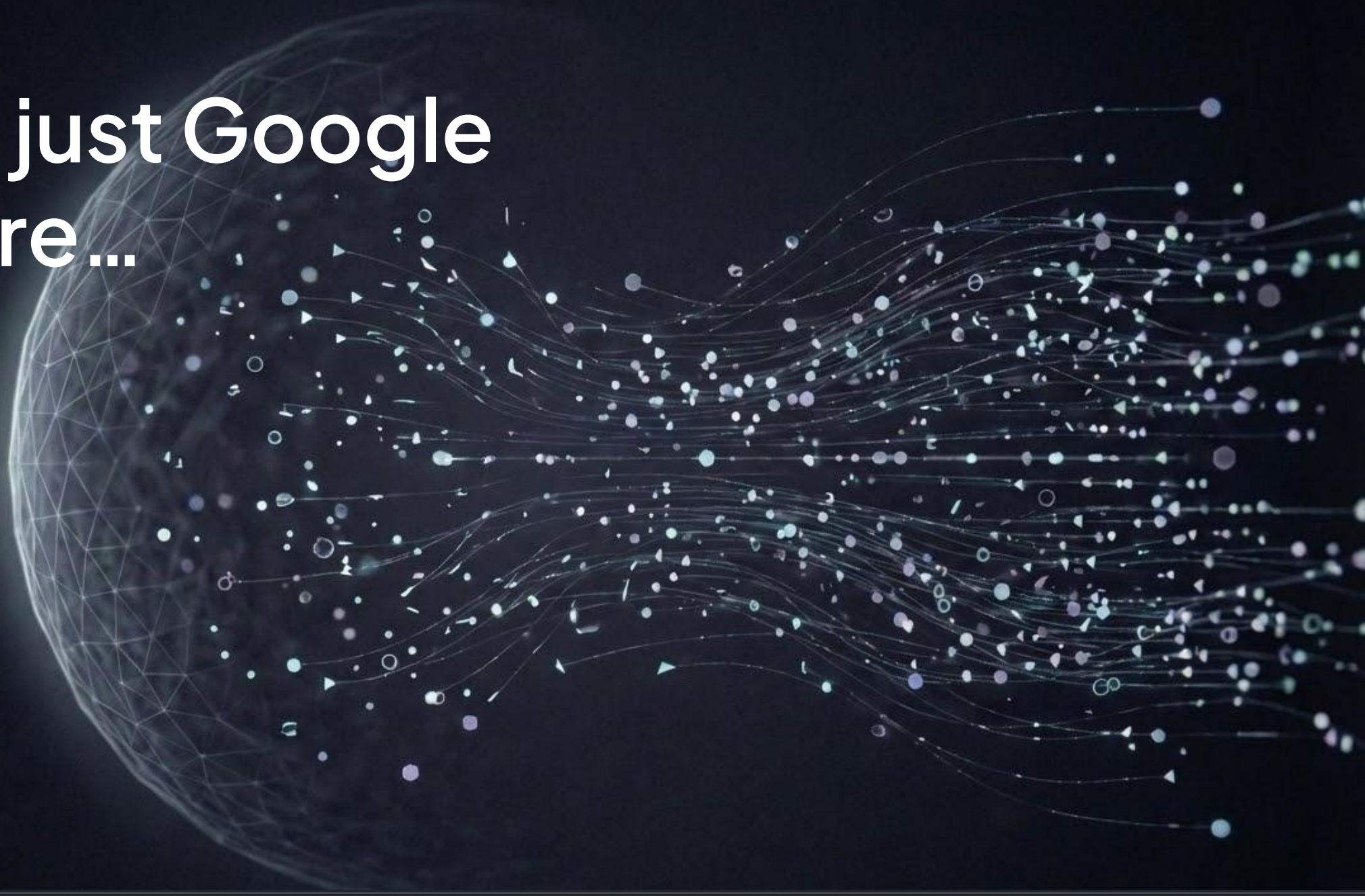


02

The current bot landscape



It’s not just Google anymore…



Types of bots

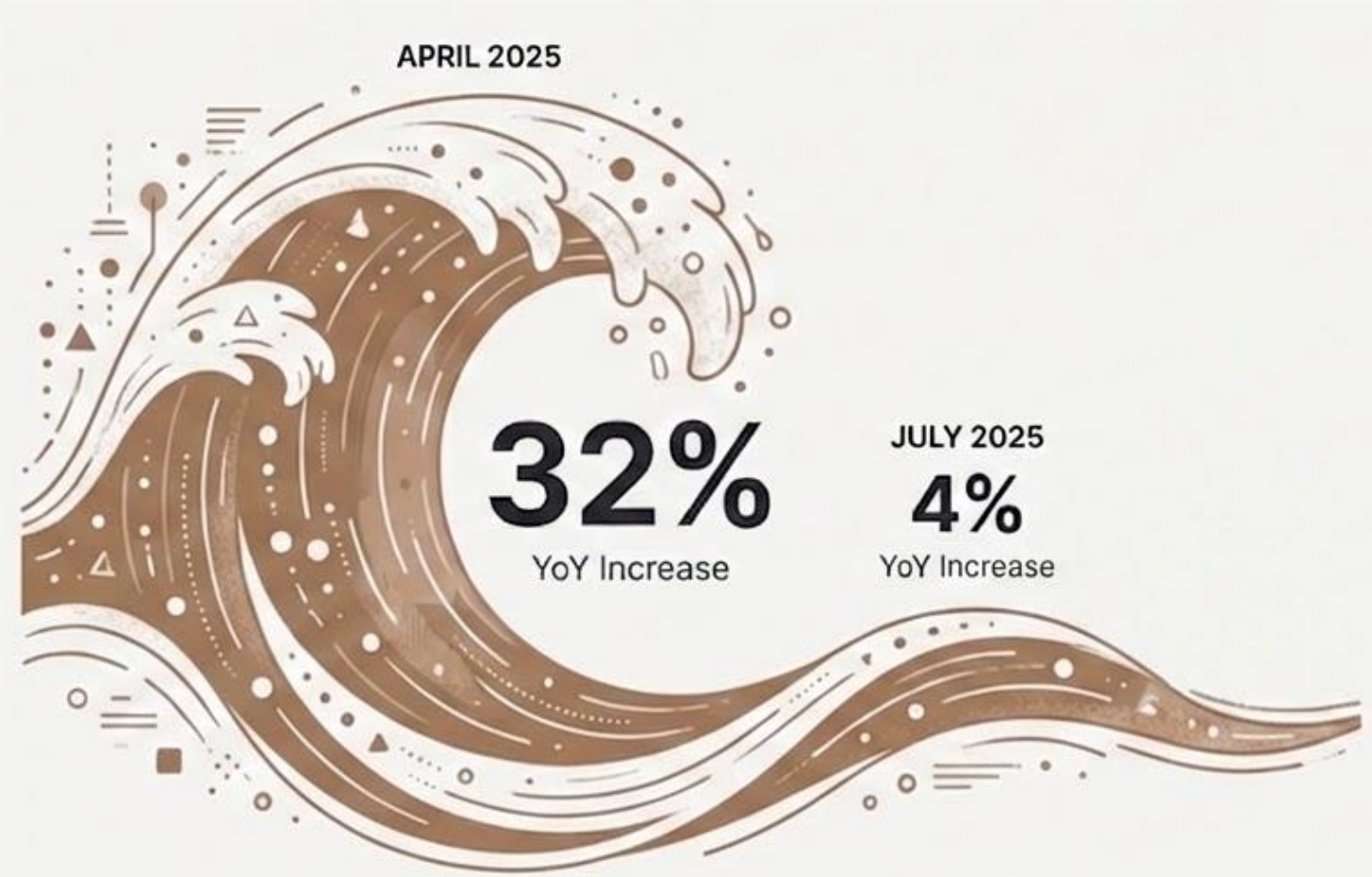
Type	Description
Web crawlers	Scan web pages, typically for indexing search engines (e.g., Googlebot, Bing)
AI crawlers	Ingest, index, and extract large volumes of data for LLMs (e.g., OpenAI, Anthropic, Perplexity, etc.)
Malicious bots & mystery scrapers	Disregard policies (e.g., robots.txt) and perform negative behavior: ■ Web/content scraping ■ Denial of Service (DoS) ■ Distributed Denial of Service (DDoS) attacks

Majority of traffic from bots

In 2025, 51% of traffic was automated

Source: https://www.imperva.com/resources/resource-library/reports/2025-bad-bot-report/

Scope of AI bots

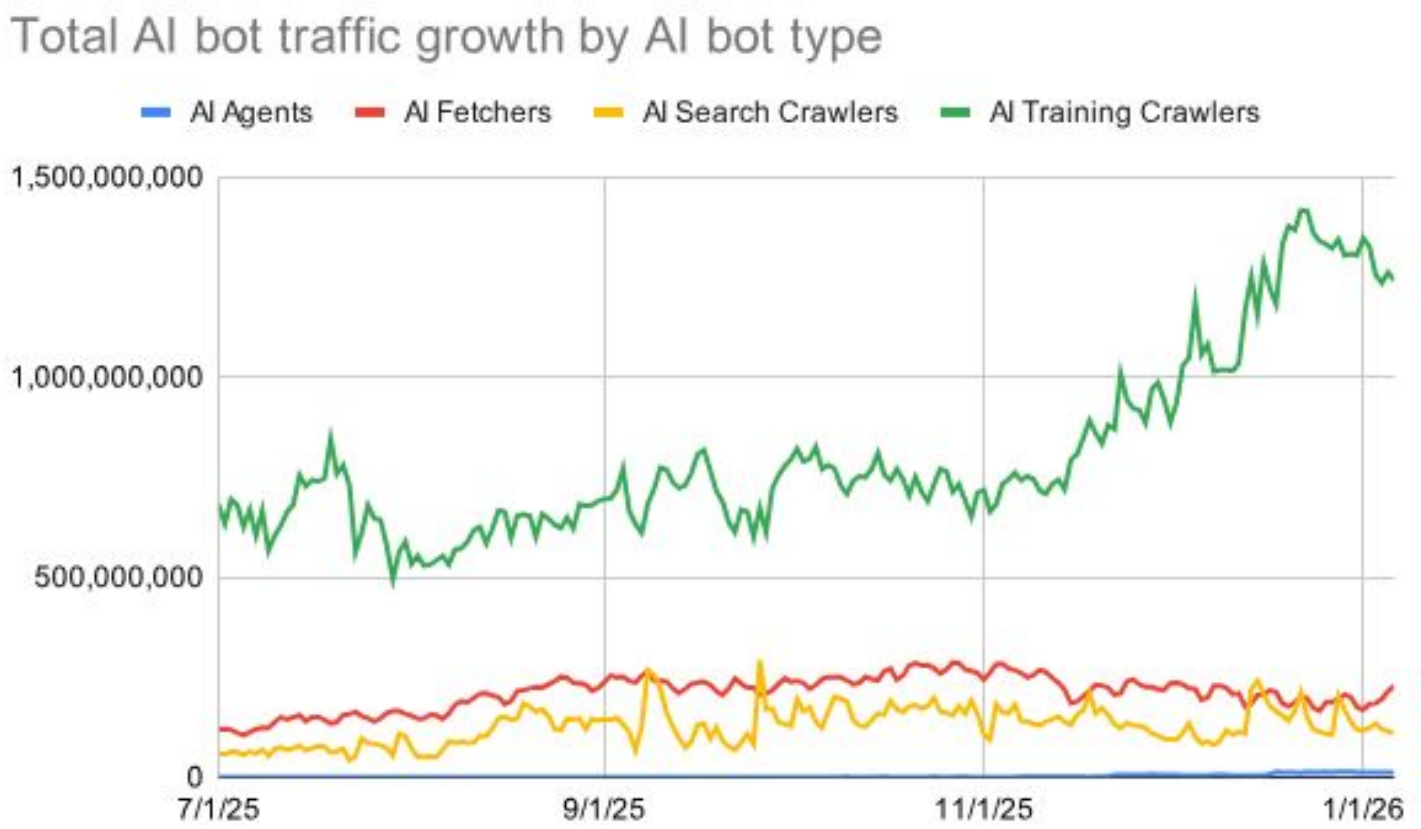


Data source: https://blog.cloudflare.com/crawlers-click-ai-bots-training/

Huge surge in traffic in April 2025

Continue to see large growth

Scope of AI bots



Source: https://www.akamai.com/blog/security/ai-pulse-how-ai-bots-agents-will-shape-2026

Why should we care about bots?

Situation

- Typically crawl full sites at large/massive scale
- Disregard crawl-delays and may not acknowledge other measures
- Highly repetitive crawling

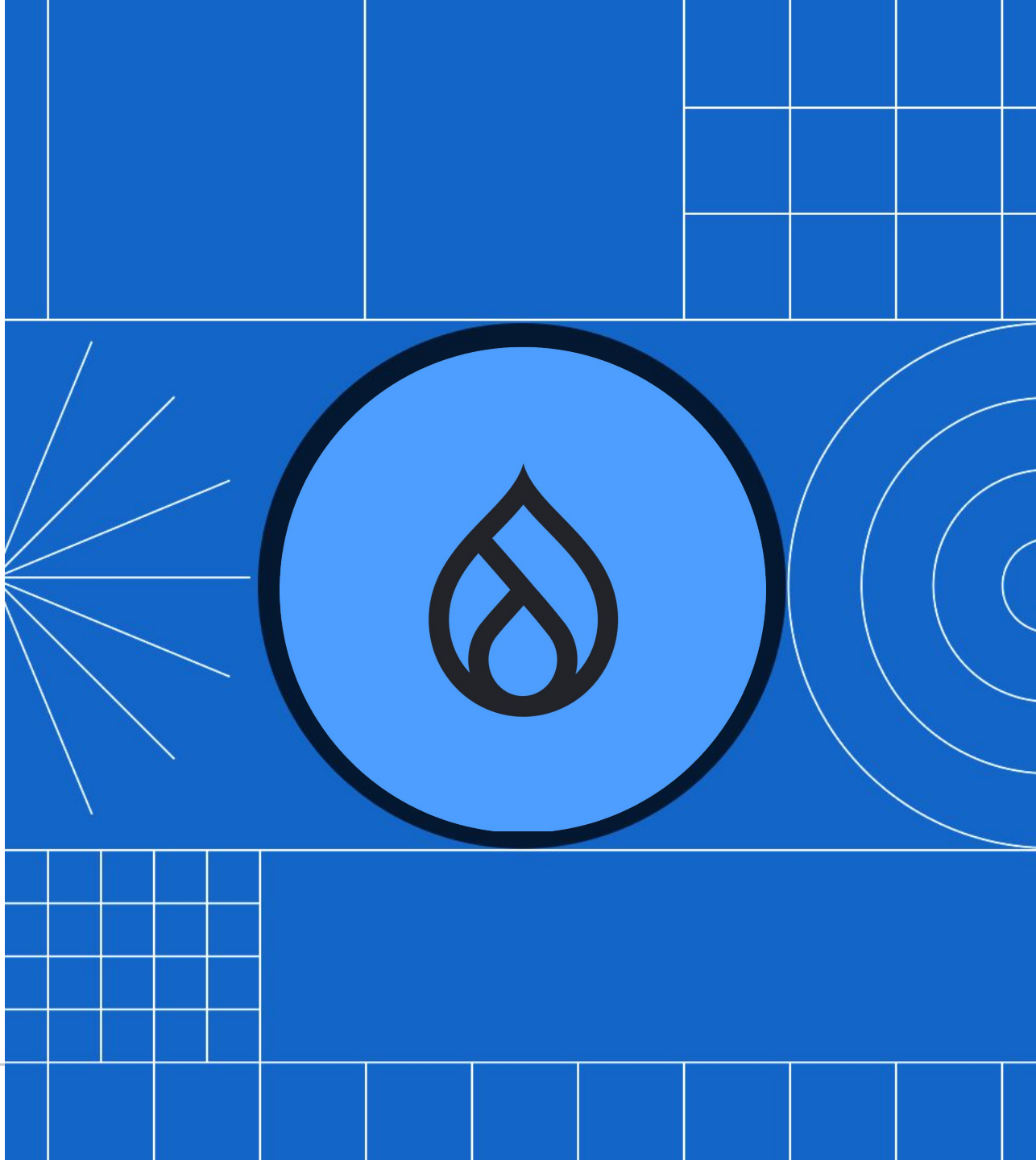
Impact

- Increased server load
- Bandwidth and other costs
- Slower performance for real users
- Content scraping at scale

```
# Main robots.txt file for example.com
User-agent: *
Disallow: /private/
Disallow: /temp/
Crawl-delay: 10

# Googlebot specific rules
User-agent: Googlebot
Disallow: /no-google/
Allow: /public/

# Bingbot rules
User-agent: Bingbot
Disallow: /no-bing/
Crawl-delay: 5

Sitemap: https://www.example.com/sitemap.xml
```

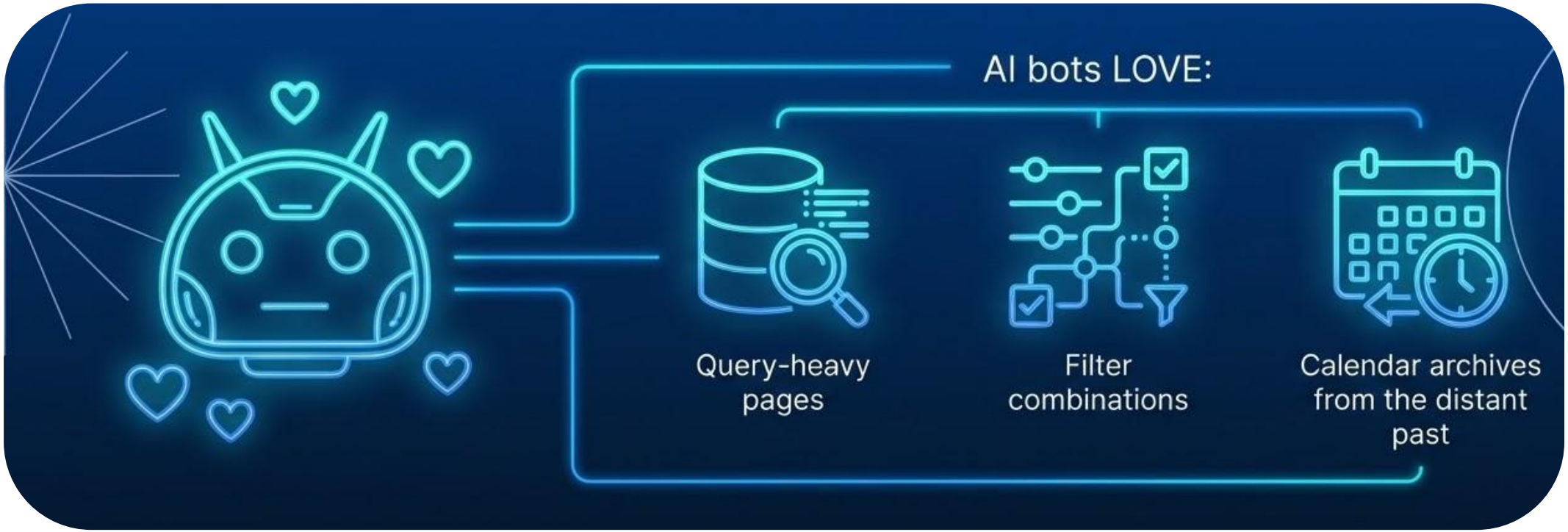
03

What happens to Drupal?



Drupal: There's a lot (for bots) to love

Drupal sites have search, filters, Views, and more.



What’s the impact on your Drupal site?

A lot depends on any mitigations that you have in place and your hosting infrastructure.

- Without caching, your site may go down.
- Even with caching, the bots may target pages that hit the Drupal backend and negatively affect performance, even for unauthenticated traffic

Hosting on Pantheon?

- Monthly visits cap

Hosting on Acquia?

- Views and Visits usage limits and overages
- Solr search on Acquia Search and SearchStax limits

Measure analytics and performance

Review your analytics and site performance.

If you're under active attack, you may want to focus on making some of the adjustments that we'll talk about soon, particularly at the edge.

1. Measure bot vs. human ("real") traffic to provide a baseline
2. Monitor over time and review after making adjustments or as you're working to mitigate an attack

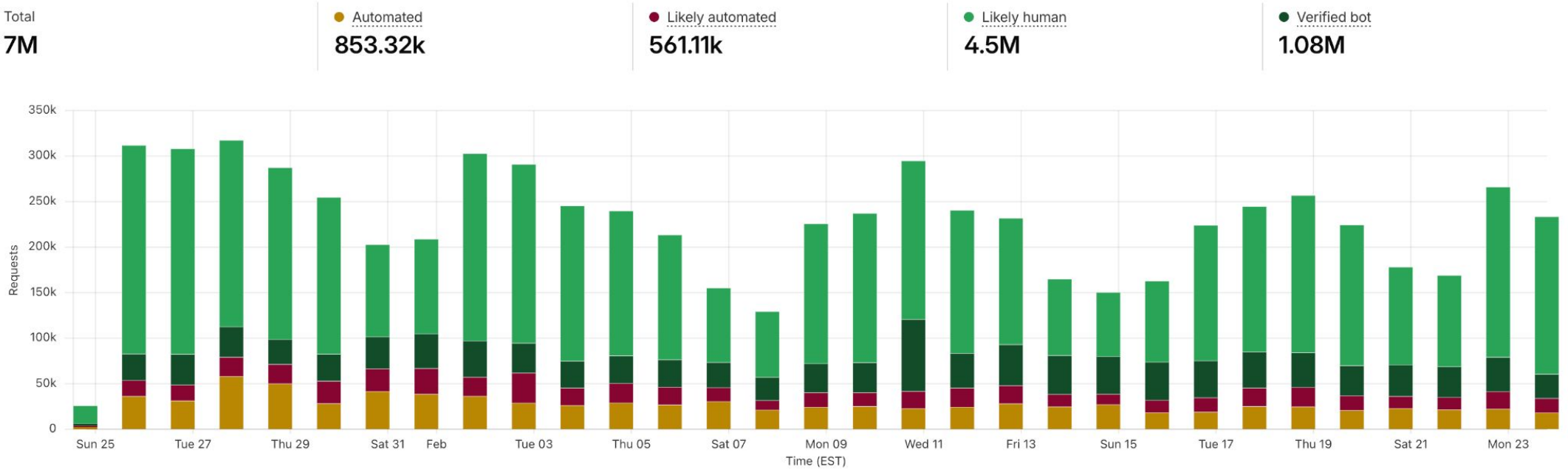
Most AI crawled paths

Path	Host	Content type	↓ Allowed requests
/programs/find-program	rdpolytech.ca	HTML	12,262
/find-program	rdpolytech.ca	HTML	11,403
/sitemap.xml	rdpolytech.ca	XML	2,829
/	rdpolytech.ca	HTML	1,732
/log	rdpolytech.ca	Empty	490
/programs/practical-nurse-diploma	rdpolytech.ca	HTML	248
/	media.rdpolytech.ca	HTML	233
/programs/licensed-practical-nurse-lpn-bsc-nursing-bscn-transition-program	rdpolytech.ca	HTML	233
/programs/health-care-aide-certificate	rdpolytech.ca	HTML	233
/about-us/services-departments/office-registrar/course-outlines	rdpolytech.ca	HTML	215

1 - 10 of 4916 items

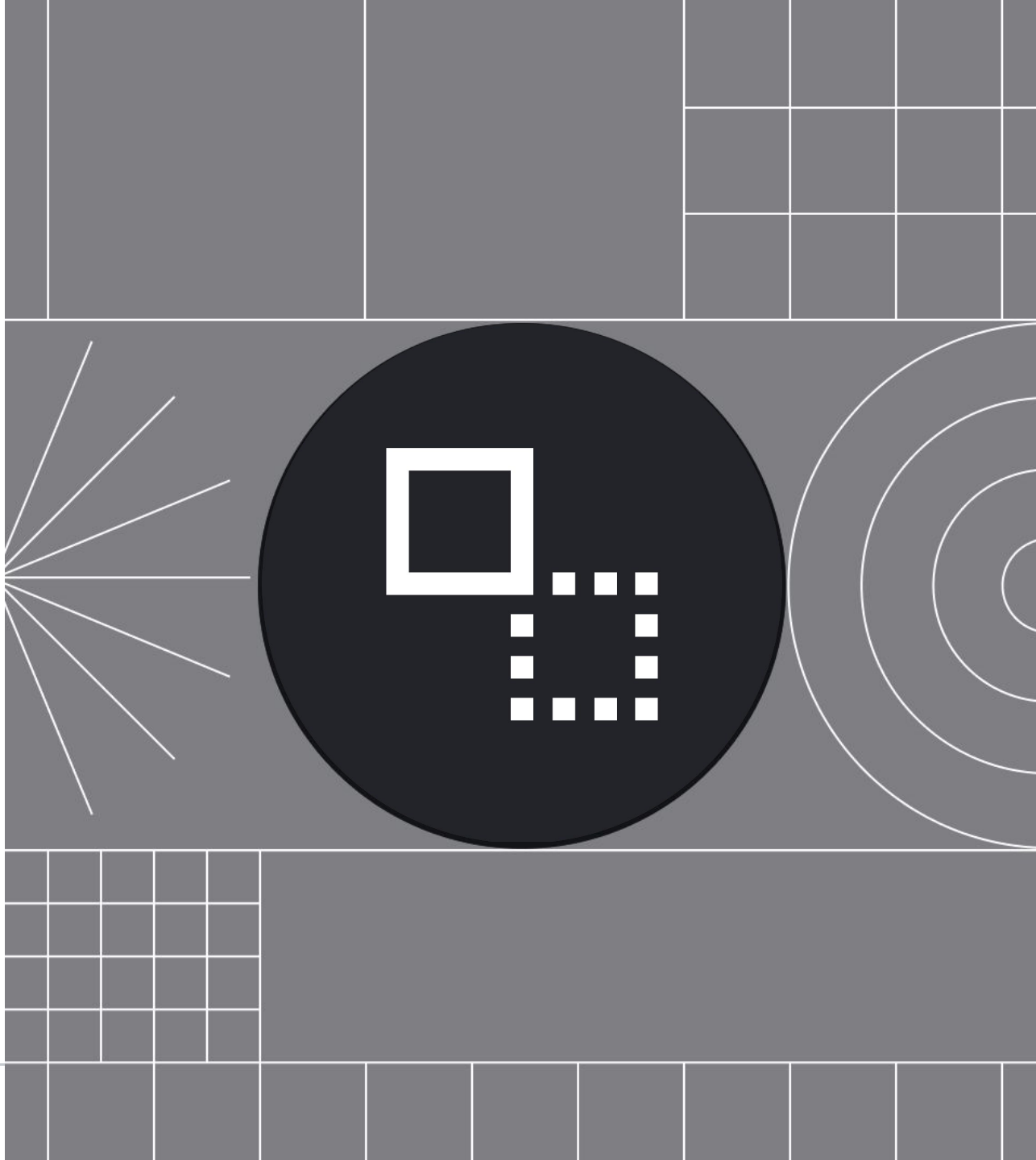
Page 1 ▾ of 492

Number of requests made by bots and humans

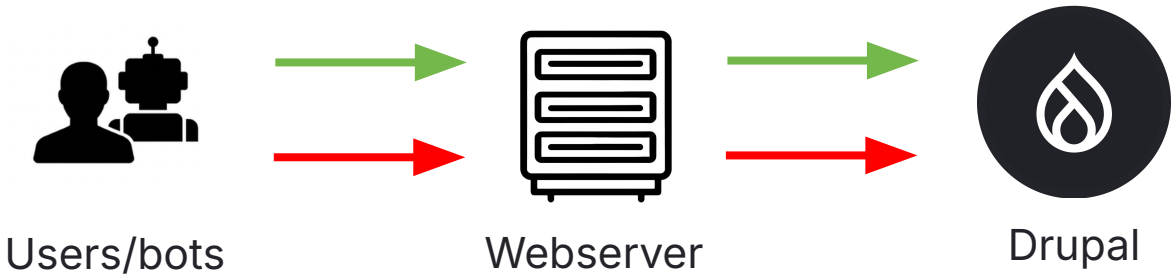


04

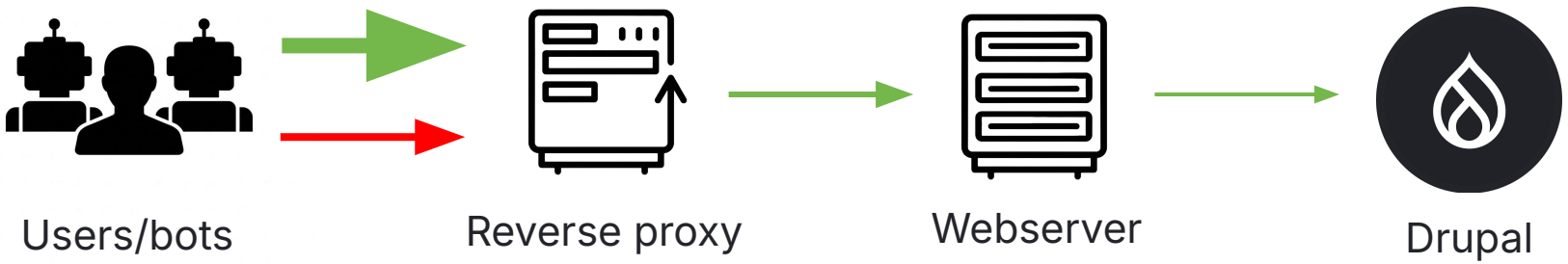
Layered defense strategy



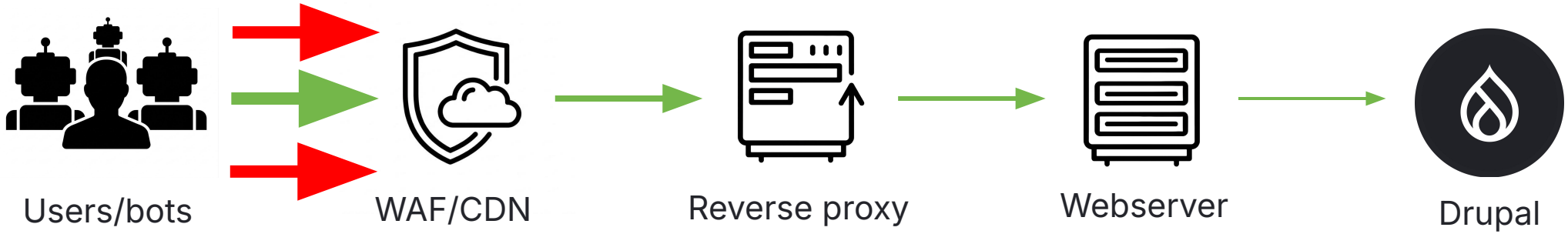
The traffic flow: Minimal infrastructure



The traffic flow: Reverse proxy



The traffic flow: WAF/CDN



WAFs and CDNs to the rescue!

A Web Application Firewall (WAF) and Content Delivery Network (CDN) protect your Drupal site at the edge:

- Cloudflare
- Fastly
- Sucuri

Drupal-specific examples:

- Acquia Edge
- Pantheon Advanced Global CDN

WAF/CDN benefits

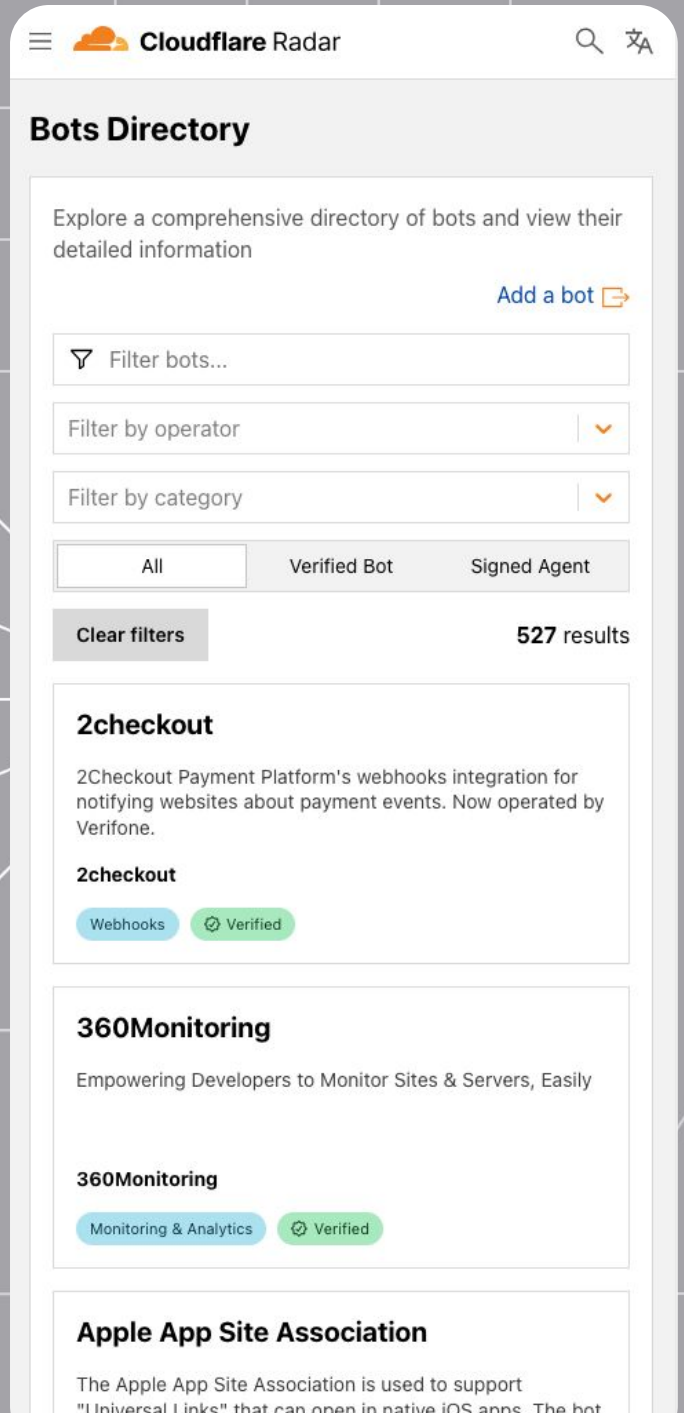
- Serve or stop traffic before it hits Drupal or the origin infrastructure
- Global Distributed Denial of Services (DDoS) absorption
- Edge-based rate limiting
- Bot detection scoring
- Country-based rules
- Challenge process
- Latency reduction for legitimate users
- Behavioral/fingerprint-based detection
- Request anomaly detection
- Credential stuffing protections
- IP reputation databases (limited benefit)

Additional benefits

- Proper implementation has a positive impact on performance for all users
- Better user experience with global distribution

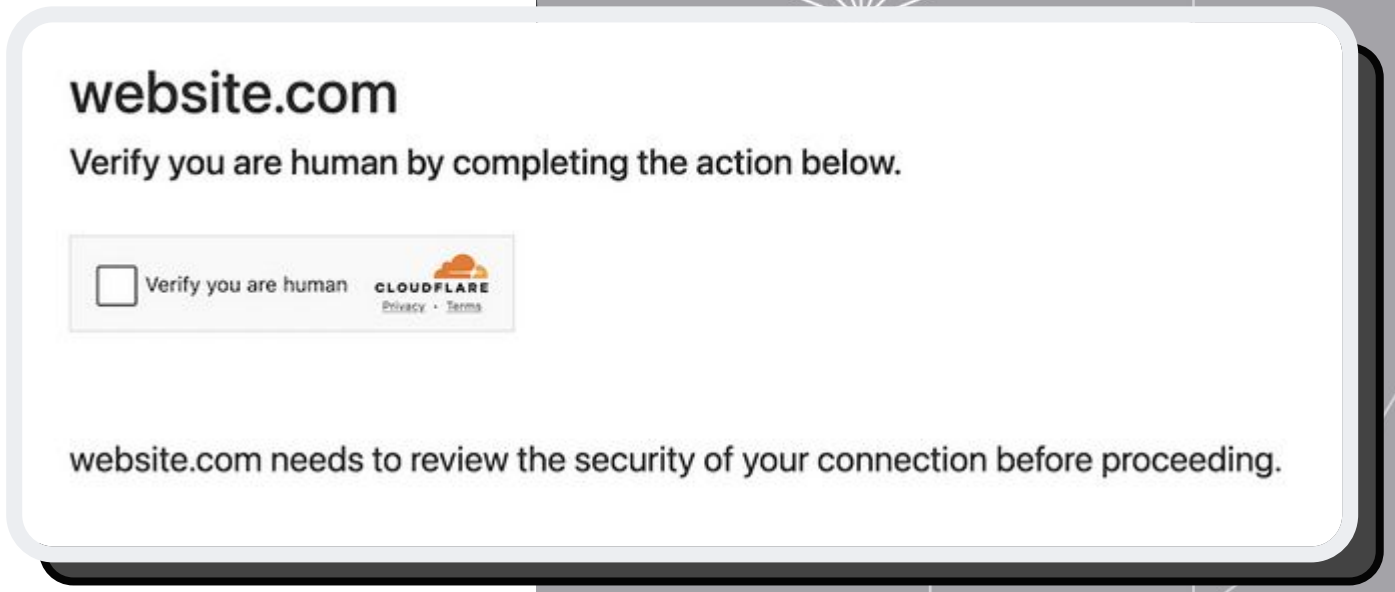
Bot management & bot scores

- Allow verified bots (e.g., Google)
- Challenge or block suspicious bots
- Create WAF rules based on bot scores
- Bots Directory to learn more information about individual bots



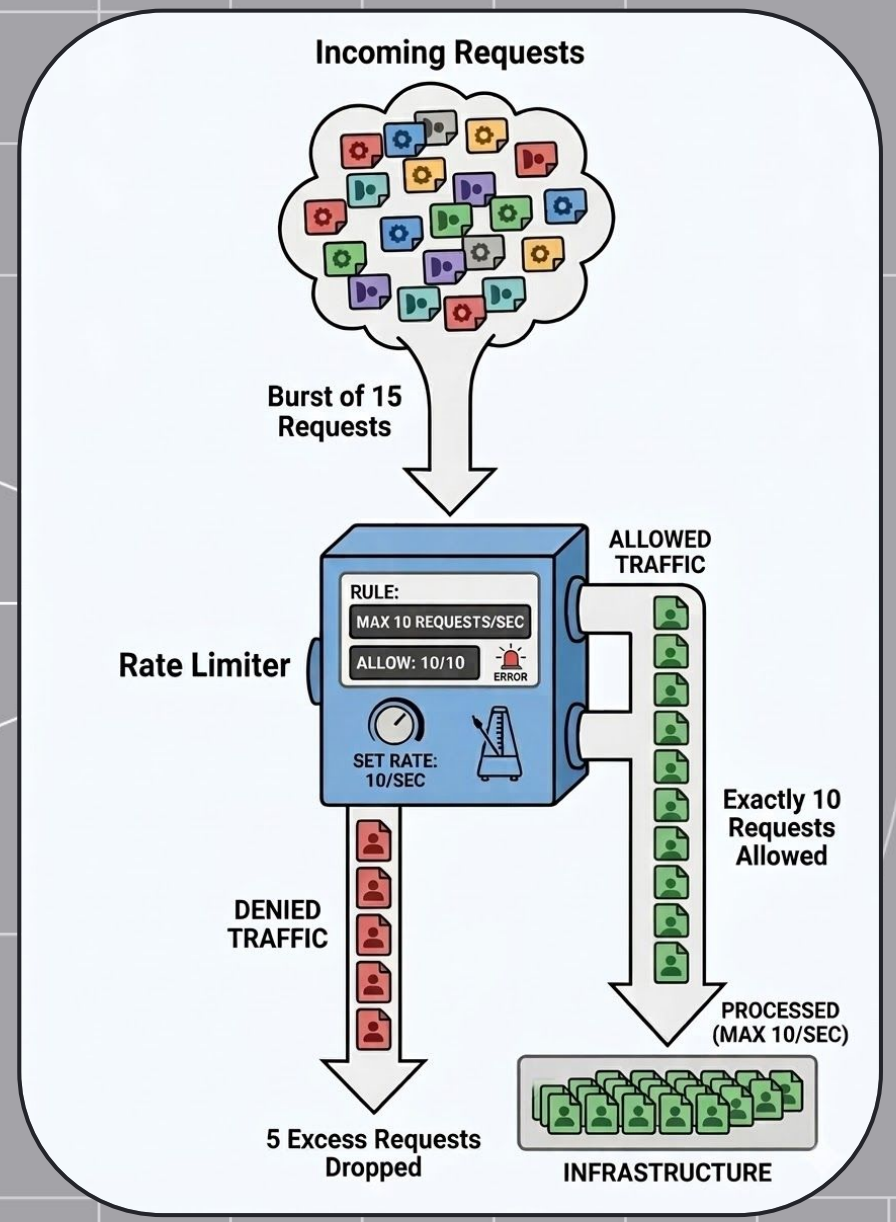
Targeting AI crawlers: Block or challenge

- Block specific AI crawlers
- Challenge AI crawlers (e.g., CAPTCHA, Managed Challenge)
- Rate limit them



Rate limiting

- AI bots tend to hit many pages very quickly
- X requests per Y seconds (can be per IP address, path, etc.)
- Strictly rate limit requests on key pages (e.g., `/search`)



Rules for high-risk paths

Create WAF rules for:

- `/search`
- Filter-heavy views
- Other areas that can be targeted with query strings

Challenge non-verified bots on dynamic endpoints

Block suspicious query string patterns

Strategically cache everything

The best requests are those that never reach Drupal

The CDN cache stores copies of frequently accessed content like images, videos, or entire pages, geographically distributing it across its data centers, which are closer to your end users.

This reduces server load and improves performance.

To block or not to block...

Should AI be blocked completely?

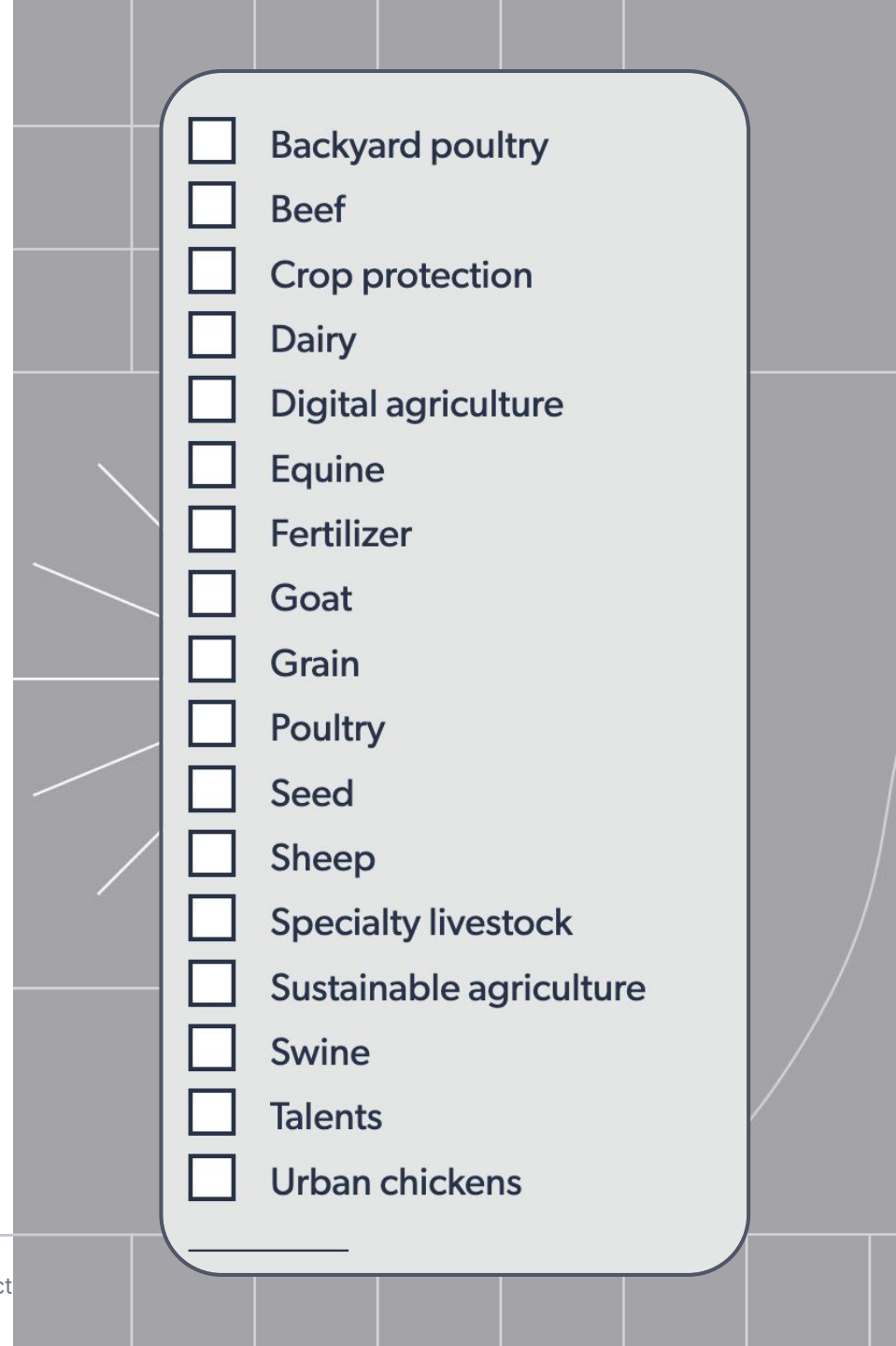
...it depends

Other infrastructure

- Firewall (e.g., iptables)
- Reverse proxy (e.g., Varnish)
- `.htaccess` for Apache, configuration for Nginx

Facets

- Facets are "smart filters" used in search and catalog pages (common in Drupal via the Facets and Search API modules).
- They allow users to narrow down results by attributes like Category, Price, Color, or Date.
- In Drupal, each facet selection usually generates a unique URL.
  - example.com/shop
  - example.com/shop?f[0]=color:blue
  - example.com/shop?f[0]=color:blue&f[1]=size:large



Facets: Exponential explosion

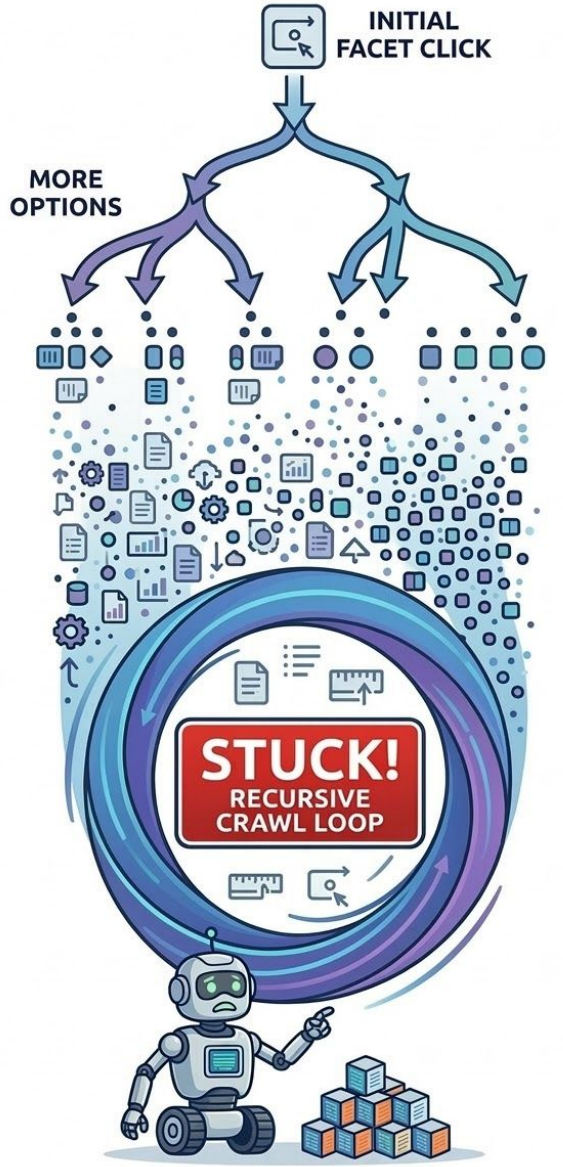
Total facets (n)	Possible URL combinations (2^n)	Real-world context
10	1,024	A very small blog with a few tags.
20	1,048,576	A standard e-commerce site category (e.g., shoes).
30	1,073,741,824	**Over 1 Billion URLs.**
40	1,099,511,627,776	More pages than the entire Google Web index.
58	288,230,376,151,711,744	Total unique search combinations on a client's site!

Why do AI bots get stuck in facets?

- Traditional bots (like Googlebot) have “crawl budgets” and some level of facet awareness.
- Newer AI training bots are more aggressive and want all the data.
- The bot enters the facet maze, and every time it follows a link, it finds more links, which is perceives as new content. It continues digging.
- The bot’s queue of “URLs to visit” grows faster than it can process them, trapping it in a deeply recursive loop.

The accidental DDoS

The AI bot hits thousands of these complex URLs per minute, spiking the CPU, locking up the database, and bringing the site down for real users.



Additional Drupal defenses

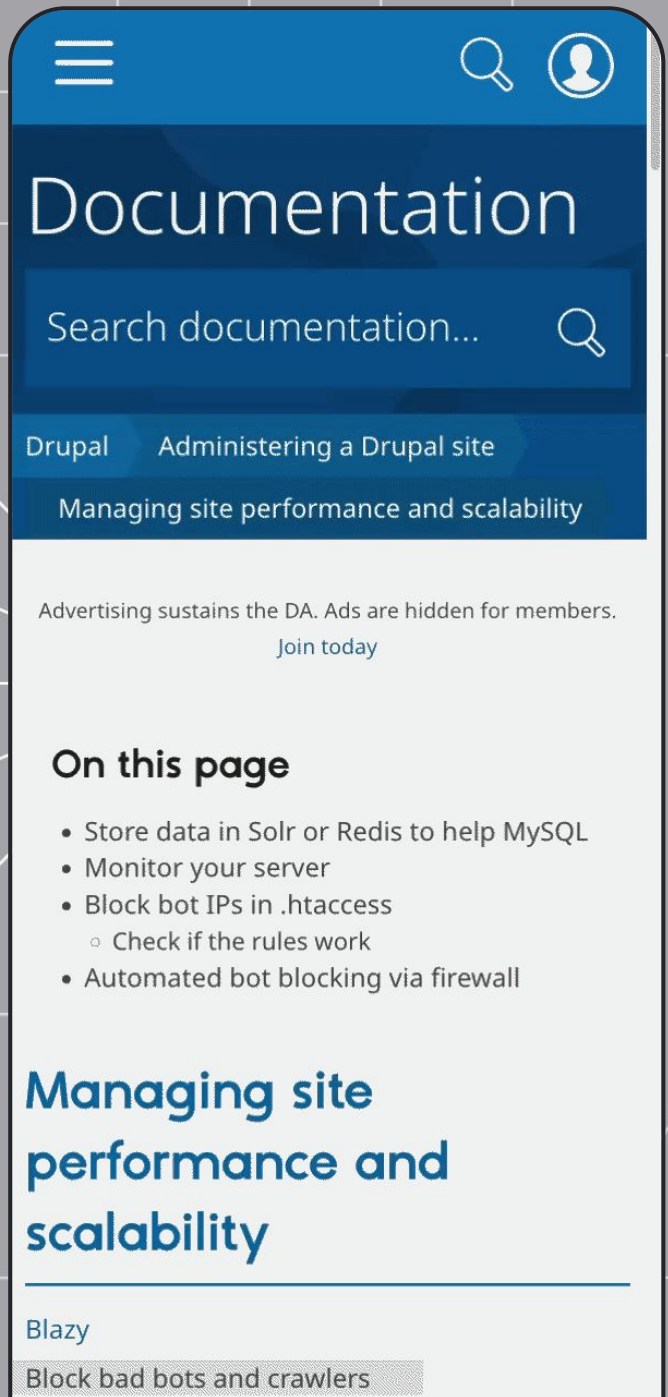
Drupal.org: Block bad bots and crawlers (documentation)

Modules

- Crawler Rate Limit
- Facet Bot Blocker
- Bot Blocker
- IP Limiter
- Protect Views Flood Control
- Cookie Bot Protection

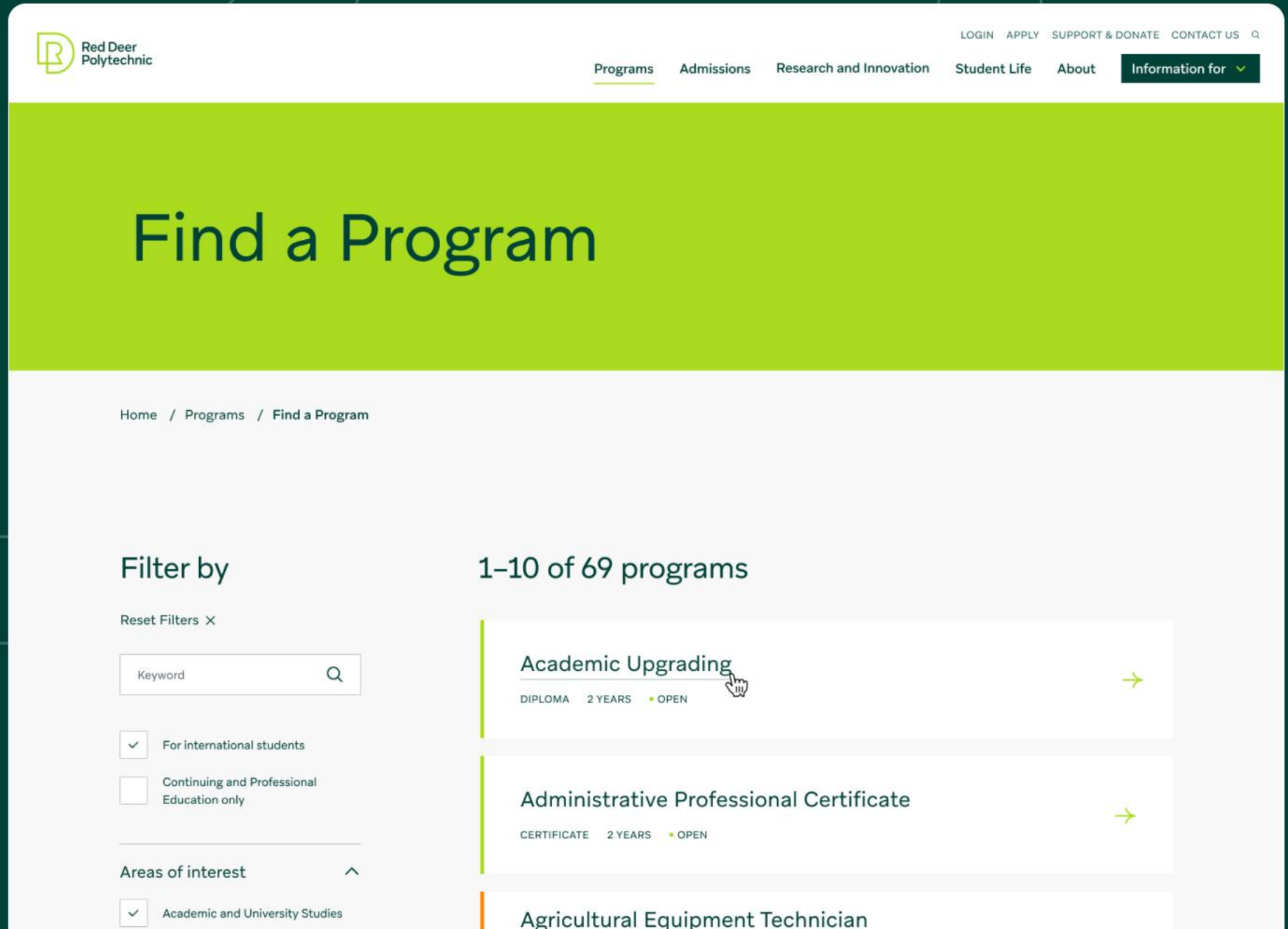
Other methods

- Use Solr for data storage to avoid database (e.g., MySQL)
- Redis caching
- Monitor your server with alerts
- Blocking bots with `.htaccess`
- Automated bot blocking via a firewall



05

Case study


Red Deer Polytechnic
LOGIN
APPLY
SUPPORT & DONATE
CONTACT US
Programs
Admissions
Research and Innovation
Student Life
About
Information for
Find a Program
Home / Programs / Find a Program
Filter by
Reset Filters
Keyword
For international students
Continuing and Professional Education only
Areas of interest
Academic and University Studies
1–10 of 69 programs
Academic Upgrading
DIPLOMA 2 YEARS OPEN
Administrative Professional Certificate
CERTIFICATE 2 YEARS OPEN
Agricultural Equipment Technician

Situation


1,800,000
1,600,000
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
Monthly limit
(750,000 visits)
September 2025
October 2025
Red deer polytechnic

What didn’t work

- Some of our WAF rules
- Adding special rules, such as `Crawl-delay`, to `robots.txt`
- Moving links from `href` anchor attribute to `data-link` (or similar) with JS redirect execution
- Masking a facet checkbox from regular users, since crawlers typically parse HTML structure to ban by IP

What did work

WAF/CDN

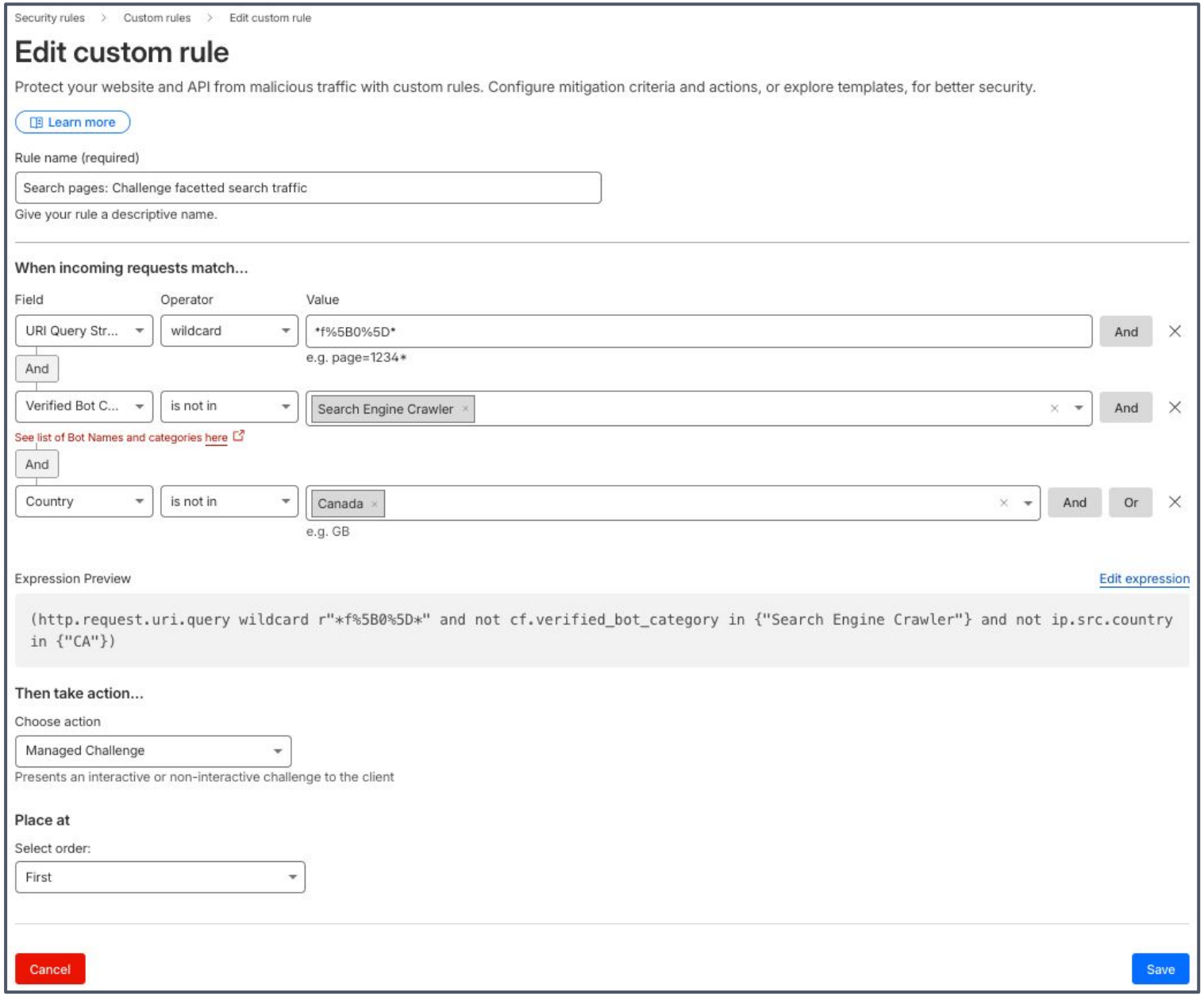
- WAF Rules
  - Rate limiting
  - Challenging faceted search traffic
  - Blocking
- Caching at the edge
- Redirects at the edge

Drupal

- Following best practices
- Upgrading Facets version 2 to 3
- Re-working facets so they are a part of search forms (instead of lists of links, for example)

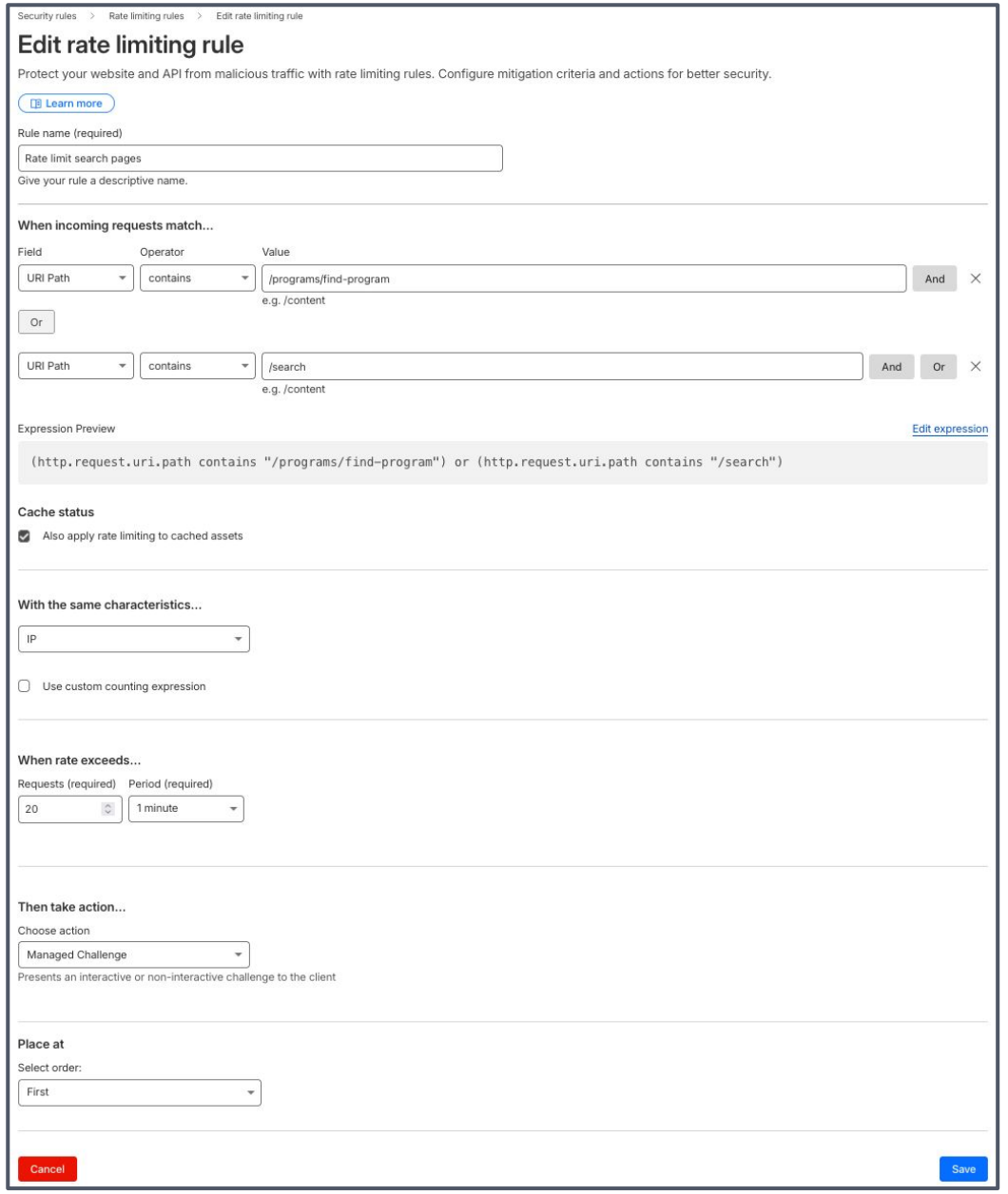
CASE STUDY

Challenge faceted search traffic



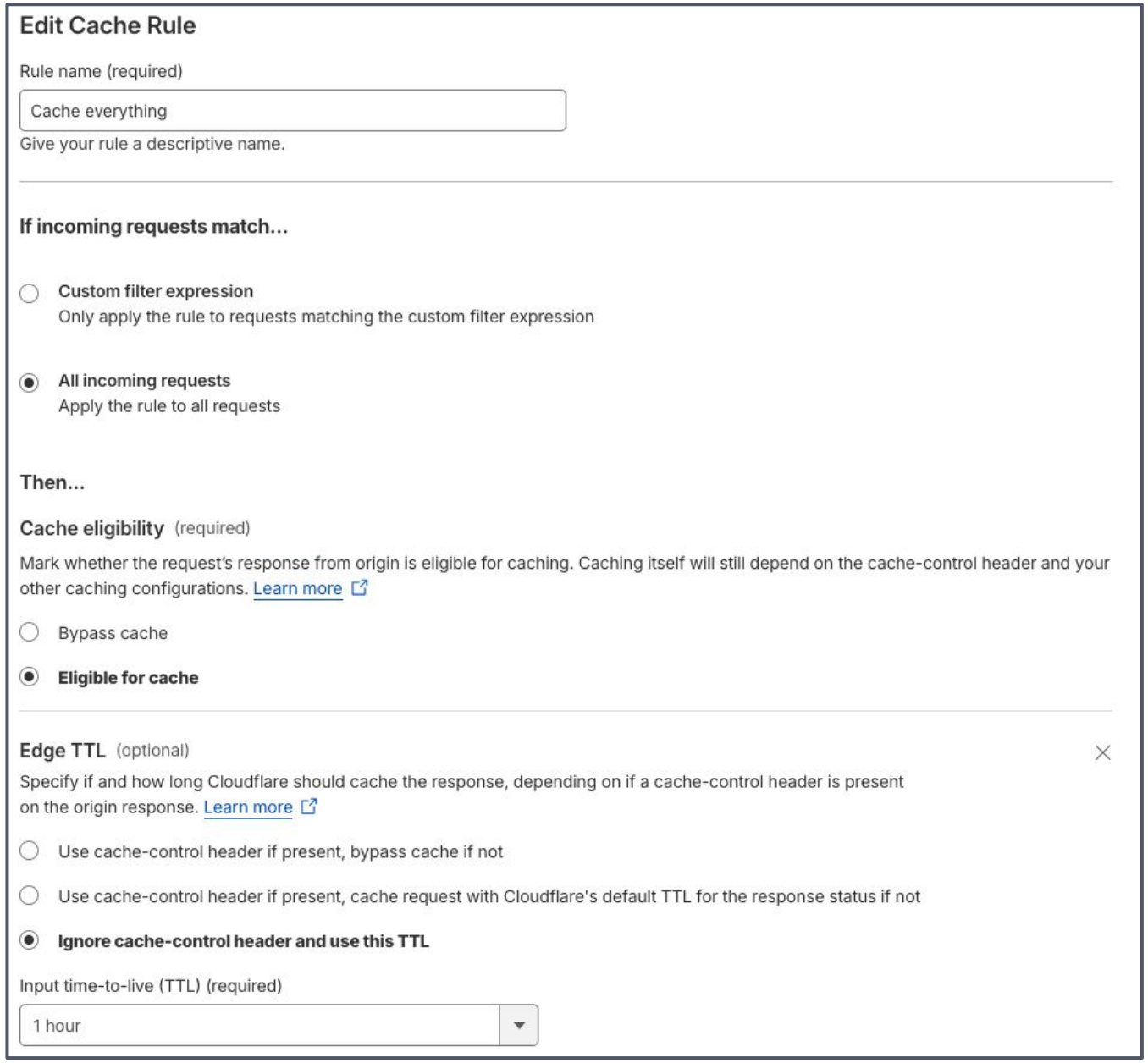
CASE STUDY

Rate limiting search pages



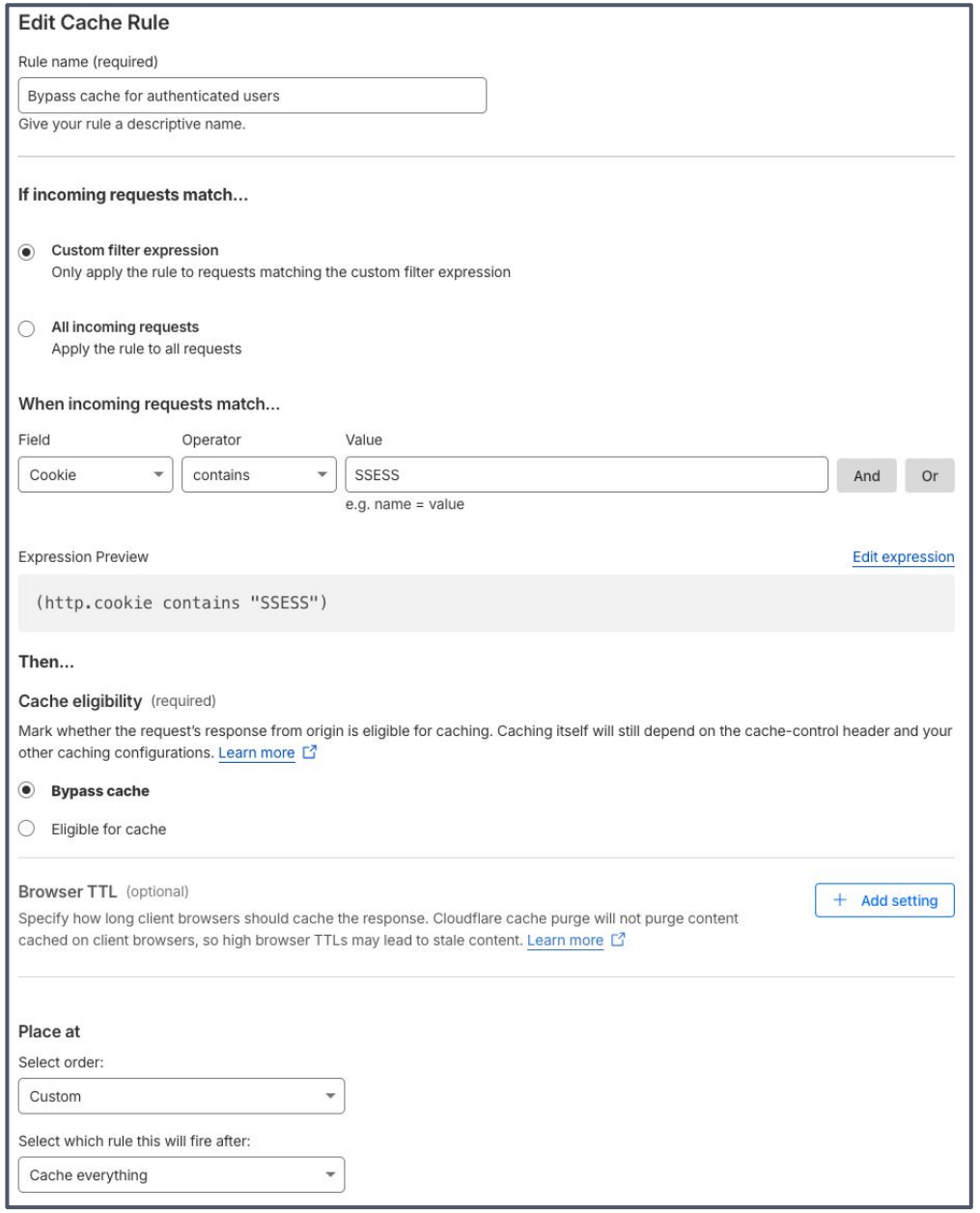
CASE STUDY

Incoming ‘cache everything’ rule



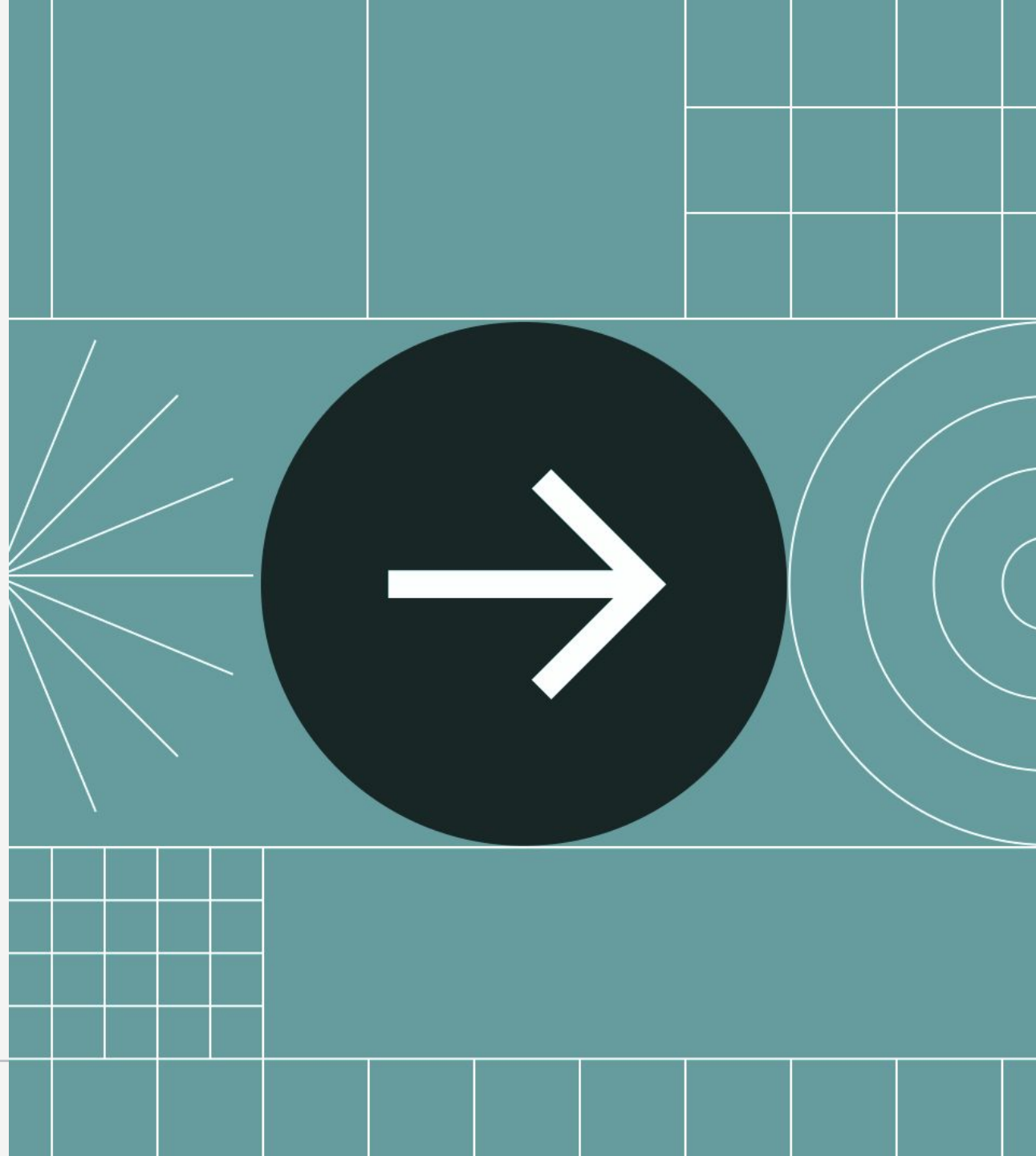
CASE STUDY

Bypass cache for authenticated users rule



06

Actionable playbook



Your 4-step bot defensive plan

1. Measure bot vs. human (“real”) traffic to provide a baseline, and then monitor/adjust
2. Add and configure a Web Application Firewall and Content Delivery Network (WAF/CDN) if at all possible
3. Review your facets
4. Follow Drupal best practices

07 Questions

Thank you