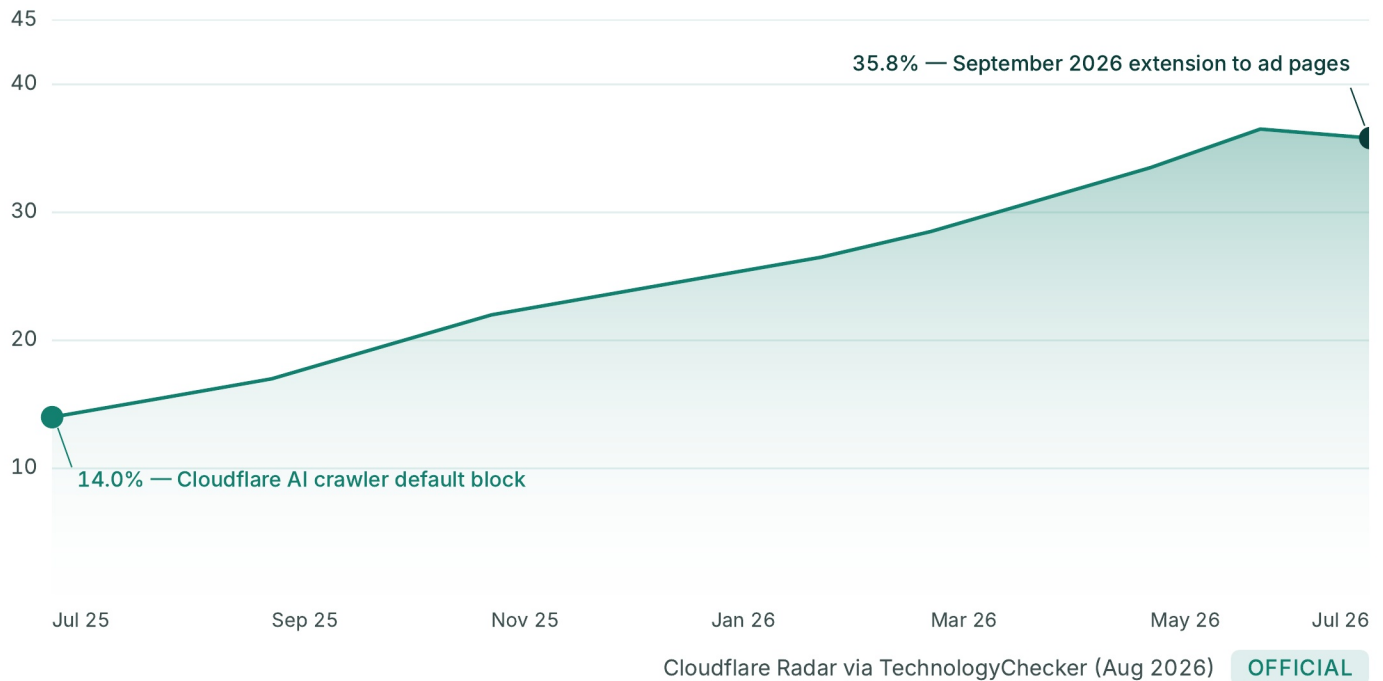




The web is pushing back.

What it costs data-driven companies to keep web data flowing · September 2026

53%

of all web traffic automated in 2025

OFFICIAL

2.5×

crawler block rate doubled in 12 months

OFFICIAL

\$126K

median total pay, US scraping engineer, before infrastructure

EST.

40%

of engineers' time firefighting pipelines

EST.

\$2.8B

alt-data spend 2025, +17% YoY

EST.

Data companies don't have a collection problem. They have a maintenance problem.

ACCESS PROBLEM

The web is pushing back

In July 2025 roughly 14% of crawler requests on Cloudflare's network were refused. By July 2026 that figure was 36%. The shift in twelve months is the clearest signal in the data: the web is moving from open to permission-required.

53% **416B** **12.5×**

of web traffic automated in 2025; bad bots alone 40%

OFFICIAL

AI-bot requests blocked in 5 months after default blocking

OFFICIAL

growth in AI-enabled bot attacks; 2M to 25M daily

OFFICIAL

1 Jul 2025

Cloudflare blocks AI crawlers by default (~20% of the web)

2025-2026

HTTP 402 "pay to crawl" responses appear

15 Sep 2026

Cloudflare extends default blocking to advertising-monetised pages

Today

~25% of top 1,000 sites block GPTBot; 36% of all requests refused

What breaks a scraper in 2026 is rarely the HTML. It is the access layer — challenges, rate limits, fingerprinting and 403s. That layer changes every week.

COST PROBLEM

One seat costs more than most teams budget

TOTAL PAY, UNITED STATES

EST.

Web scraping eng.

Data engineer

Senior data eng.

\$100K

\$150K

Bar spans P25-P75; dot marks the median. Glassdoor / ZipRecruiter · Jul-Aug 2026 · EST.

RESIDENTIAL PROXIES, PER GB

EST.

Decodo

\$2.75 – \$3.75

Bright Data

\$5.00 – \$8.00

Oxylabs

~\$6.00 – \$10.00

Published list prices · Aug 2026 · EST.

FULLY LOADED SEAT, ONE YEAR

MODELED

Salary + on-costs \$179K

\$50K

Proxies + infra \$50K

Cover + tooling \$17.5K

Low \$177K · High \$316K per year

before the 40% maintenance tax

Glassdoor; ZipRecruiter; Proxyfacts Aug 2026; Proxidize Jul 2026; Monte Carlo survey via DataHen; DataDwip model.

MARKET

\$1.56B and growing at 17% a year

2026 VALUE AND FORECAST, FOUR ADJACENT MARKETS

EST.

Web scraping software + services

17.4%

\$1.56B → \$3.49B
2031 · Mordor Intelligence

Alternative data (investment mgmt)

~17%

\$2.8B → ~\$23.1B
2030 · Neudata

AI training datasets

22.6%

\$3.9B → \$16.3B
2033 · Grand View Research

Price-intelligence software

16.8%

\$3.7B → \$13.5B
2034 · Market Intelo

\$0 \$5B \$10B \$15B \$20B \$25B

Market-size figures vary by analyst definition. Mordor Intelligence retained as the reference for web scraping. Others independently sourced.

Competitive intelligence

AI and LLM training

Investment research

Every one of these markets is a continuous data requirement. The buyer does not pay once; they pay every month the data keeps flowing.

SEGMENT ATTRACTIVENESS

DATADWIP ASSESSMENT

Where the recurring data need is strongest

SEGMENT	DAILY FRESHNESS	BUDGET	DATADWIP FIT
Hotel rate intelligence	● ● ● ● ●	● ● ● ● ● ○	● ● ● ● ●
Pricing & digital-shelf	● ● ● ● ● ○	● ● ● ● ● ○	● ● ● ● ●
STR, real estate & auto listings	● ● ● ● ●	● ● ● ● ○ ○	● ● ● ● ● ○
Job-posting & company data	● ● ● ● ● ○	● ● ● ● ● ○	● ● ● ● ● ○
AI products built on web data	● ● ● ● ● ○	● ● ● ● ● ●	● ● ● ● ● ○
Competitive & brand monitoring	● ● ● ● ○ ○	● ● ● ● ○ ○	● ● ● ● ○ ○

Scores reflect recurring managed-collection attractiveness, not market size. DataDwip has live case studies in hotel rate intelligence and manufacturer pricing. — DataDwip assessment.

THE DATADWIP OPPORTUNITY

From fragile scrapers to a data acquisition operation

14%→36%

Crawler block rate, Cloudflare network, Jul 2025–Jul 2026

Protected sources require proxies, headless browsers and constant tuning. That layer changes every week.

\$126K

Median total pay, US scraping engineer, before infrastructure

The hire takes months. The maintenance tail — 40% of time firefighting — never ends.

+17%

Alt-data spend growth 2025; AI dataset market +22.6%/yr

Buyers expect daily or faster refresh. Coverage claims are collection promises.

What we promise, in writing.

Freshness	95% of deliveries inside the agreed window
Coverage	≥97% of expected records per run
Data quality	≥99% pass validation; duplicates <0.5%
Repair	Detected in 4h; fixed in 24 business hours
Schema	30 days' notice on breaking changes

What your team inherits.

Collection

Marketplaces · OTAs · listing portals · career pages · company sites

Maintenance

Change detection · repair · proxies · headless browsers · monitoring

Quality

Validation · de-duplication · schema stability · warehouse / API delivery

Outcome

Coverage claims · freshness SLAs · products that stay current

Start with a 4-week pilot.

2–3 of your hardest sources, delivered daily, SLA tracked every week. Pilot fee credited against month one.

[Book a pilot](#)[Talk to us](#)

SOURCES & METHODOLOGY

Tier 1 — Official	Cloudflare Radar; Imperva/Thales Bad Bot Report 2026
Tier 2 — Est.	Mordor Intelligence; Grand View Research; Neudata; Monte Carlo
Tier 3 — Est.	Glassdoor; ZipRecruiter; Proxyfacts Aug 2026; Proxidize Jul 2026
Tier 4 — Vendor	Coverage figures reproduced as stated, not verified
Modeled	Fully loaded seat cost; segment scores — DataDwip Research

Market-size figures vary by analyst definition; Mordor Intelligence is the reference for web scraping. Legal summaries are for general information, not legal advice.