zyte | 2026 Web Scraping Industry Report

The age of fast-forward web data

Accelerants and roadblocks

Table of contents

The future I dreamed of is dawning

Shane Evans, CEO, Zyte

I started Zyte 15 years ago with a single, driving wish: to make it easier to gather data from the web.

Back then, and for so many years since, the reality of our industry was defined by fragility. Web data acquisition was often a painful, manual process. You built a scraper, it broke, you fixed it, and you hoped it held together long enough to get the data you needed. It was a constant battle against friction.

Today, however, the mood is different. For the first time, we now have a clear line of sight to the complete solution I dreamed of all those years ago. The path from "fragile code" to "effortless data" is a reality being actively built today.

I have never been more excited or optimistic about the future of our industry - and the reason is artificial intelligence.

The era of AI agents

We are moving past the era where AI is just a buzzword or a helper tool. We are entering a phase in which AI agents will automate all aspects of web data extraction to a meaningful degree.

While we won’t solve every single edge case this year, the landscape is going to look vastly different 12 months from now. AI is fundamentally changing the mechanics of our work - from writing the initial code to handling quality assurance.

It is taking what was once a manual, error-prone craft and turning it into an automated, resilient engineering discipline.

Access on auto-pilot

For years, we have talked about the "cat-and-mouse" between data-gatherers and data owners. But 2025 marked a significant transition in this space. The pace of updates from bot management vendors has accelerated, and the old methods of maximizing data success are becoming less effective.

This is a *good* thing.

The complexity of modern anti-bot measures has gone past the point where human developers can, or should, manage it manually. This is forcing a necessary evolution. Data people want to work with data; they don't want to spend their days developing responsive strategies to unblock a website.

So the industry is moving toward automated systems that conduct stringent website analysis to generate crawling recipes automatically - something we are heavily investing in at Zyte. This shift will free developers to focus on value, not on friction.

The science of compliance

The relationship between web data and AI has become symbiotic. AI is not only driving the methods of web scraping; web data is becoming the fuel behind a massive new wave of AI businesses.

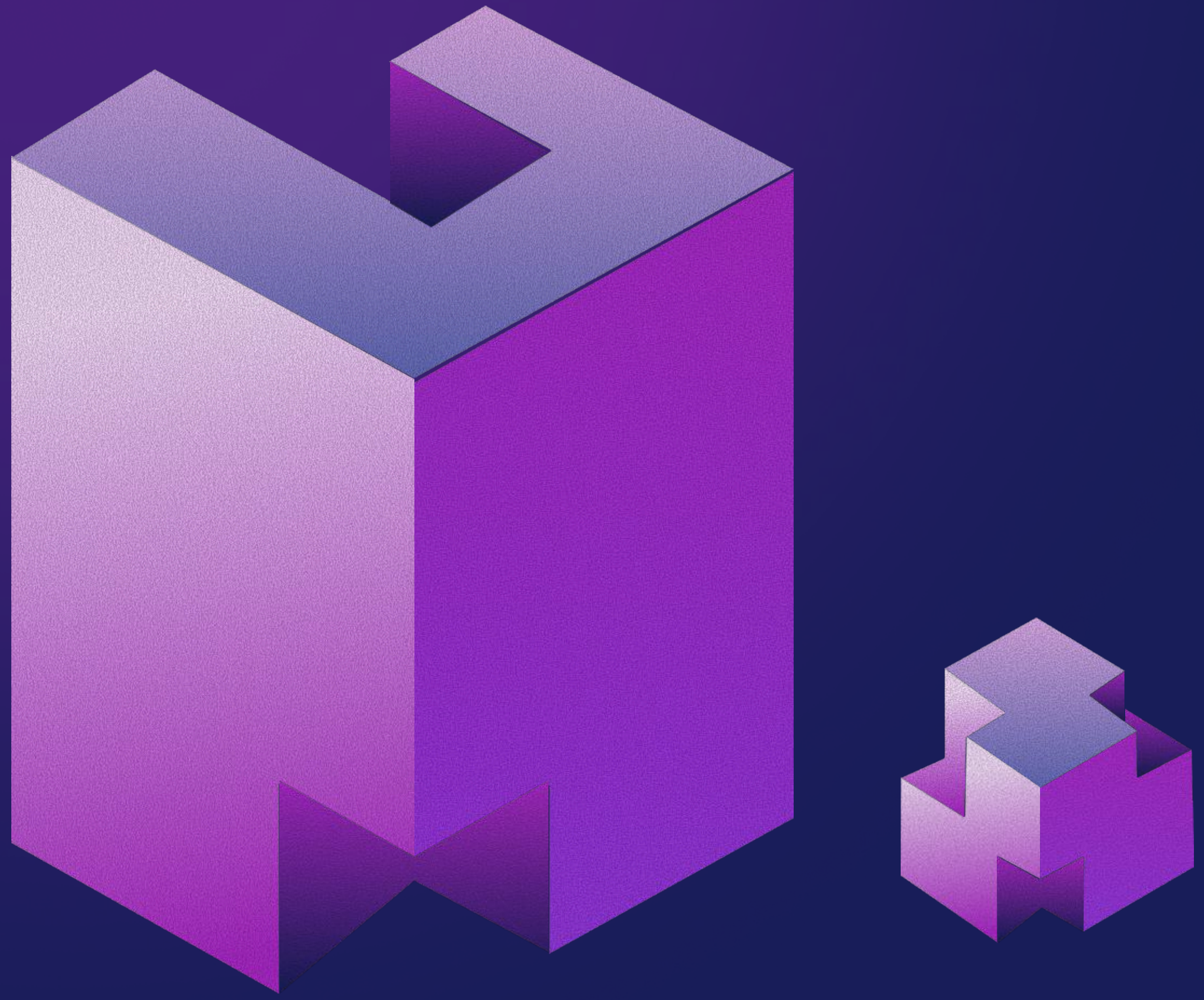
Naturally, this brings new questions regarding copyright, compliance, and ethics. The legal landscape governing data extraction is shifting significantly, and a central goal for industry leaders is to advocate for and adhere to good practices.

Rather than viewing these as roadblocks, I see them as signs of a maturing industry. We are beginning to see answers to these legal questions emerge. The "wild west" days are steadily being replaced by a landscape where compliance is built into the infrastructure, enabling sustainable growth for AI-driven companies.

Back to the future

We are witnessing the transition from a dream to a defined path. The friction of the past 15 years is giving way to a future where data extraction is seamless, compliant, and powered by intelligent agents.
It is incredibly exhilarating for me to see that the technology is finally catching up to the vision.

Overview



For much of its history, web data extraction existed in the margins of enterprise technology. It was something developers did, often manually and invisibly.

The industry that grew around proxy vendors, anti-bot providers and scraping tools - those many, individual links in the data chain - also seemed to operate in a regulatory gray zone.

In 2026, that era ends. The point solutions have consolidated. Web scraping has become critical economic infrastructure.

Market explosion

The web scraping market reached $1.03 billion in 2025 and is forecast to expand to $2 billion by 2030, according to Mordor Intelligence. Other estimates put that figure at double.

At this point, the majority of mid-to-large enterprises use web scraping for competitive intelligence, while most e-commerce companies monitor rival prices using scraped data.

In 2026, then, it is clear that web data extraction is no longer in the shadows. Rather, it is the steady heartbeat of the data-driven economy.

The data paradox

This year, though, it is also possible to see that, inside the heart of scraping itself, there is a growing dichotomy.

Scraping is becoming easier

We already know that full-stack web scraping APIs have been abstracting away the tedious, responsive infrastructure work that goes with manually orchestrating all those individual components of the traditional scraping cycle.

Now, just as it is disrupting software engineering generally, AI-powered code generation is ready to revolutionize scraping, drastically cutting development time. LLM-based extraction is now handling layout changes that would break traditional scraping systems. Headless browsers built for AI can reason about what they see, leading to smarter, responsive scraping.

For developers who have access to the right tools and platforms, then, scraping has never been simpler. The barrier to entry is dropping and, with it, time-to-data is shrinking.

Access to data is becoming harder

Simultaneously, however - and perhaps in response to this growing capability - a new wave of limits is being applied on many data sources.

Anti-scraping measures have reached a sophistication that renders the old manual workarounds impotent. Meanwhile, new gatekeeper technology, giving website owners control over which visitors are permitted, means the web could be fragmenting into distinct islands, each with different rules and economics. Regulatory frameworks are arriving with real teeth, too. Some think the web may no longer be a single, undifferentiated resource with one set of rules but, rather, a collection of discrete, governed ecosystems.

In this environment, organizations with the right infrastructure, compliance systems, and strategic partnerships will thrive. Those without will find themselves locked out of valuable data sources.

Year of the machines

The tension created by this contradiction is surfacing thanks to the arrival of a new kind of organism on the web - artificial intelligence.

AI automation is fuelling increasingly strong anti-bot defences, while agentic tooling has proliferated so quickly that the era of machine-originated traffic is no longer a mere theory.

According to Cloudflare Radar, in 2025, AI bots generated 4.2% of all HTML requests, non-AI bots generated 47.9% and humans generated 43.5%. Bots and humans, then, are about neck-and-neck. AI bots are now the most frequently disallowed user agents found in robots.txt files.

But, with visibility in new AI services becoming paramount for publishers and some data owners seeing a revenue opportunity in the shape of data-hungry crawlers, the web can no longer afford to treat all bots as adversaries - and crucially, not all website owners want to.

In 2026, these shifting sands will transform not only how data is collected, but how websites perceive, price, and govern machine access. The anti-bot arms race may continue – but, across much of the ecosystem, a new equilibrium is forming, one built on identity, policy, and managed, selective access for permitted partners.

Web data has never been more vital, and accessing it has never been more fascinating than right now.

The six trends that follow trace this transformation in detail. They show how technology, economics, and regulation are converging to reshape web data access in 2026.

#1
Trend

Data outcomes are top of the scraping stack

By 2026, the question won't be "which proxy should I use?" It will be "which API should I call?"

Managing your own proxy infrastructure is starting to feel economically irrational.

That's not because proxies have stopped working. Rather, the entire scraping stack - proxies, browsers, unblocking, parsing, retry logic - is increasingly being consumed through unified APIs that handle all of this automatically. Developers will spend less time configuring IP rotation or tuning ban-handling logic, and more time working at a higher level, interacting with websites themselves.

In 2024, Zyte began migrating customers of its dedicated Smart Proxy Manager to Zyte API. Such APIs go further than proxies alone, pairing proxies with browser automation, rendering, unblocking, and extraction to deliver reliable data outcomes that proxies alone can no longer guarantee.

Other vendors have followed suit, introducing scraping APIs that wrap proxy infrastructure inside broader data collection and delivery workflows. The pattern is clear: API-first platforms abstract component-level complexity into unified services that use smart algorithms under the hood, so users can get straight to data.

Key developments

The proxy industry has entered a state of controlled commoditization. Over 250 vendors now crowd the marketplace, according to Proxyway research. Price wars have gutted margins across the proxy market. Since mid-2023, sustained price competition among major providers has eroded differentiation and pushed proxy access toward commodity economics. Every proxy vendor offers the same baseline: rotating residential IPs, global coverage, mobile pools, sticky sessions. The arms race for network footprint has reached a point where more IPs no longer translate to better outcomes.

Meanwhile, websites have evolved their defenses far beyond simple IP blocking. Modern anti-bot systems use TLS fingerprinting to analyze cryptographic handshakes, behavioral analysis to detect unusual patterns, canvas fingerprinting to create device signatures, and JavaScript traps to catch automated visitors. Some systems report 99.9% accuracy in distinguishing humans from bots through behavioral biometrics alone.

The proxy, once the master key to web scraping, has become just one piece of an increasingly complex puzzle.

This impact is material: every year, billions of HTTP requests are wasted on retries, bans, and failed requests. In our experience, as many as one in 10 requests never deliver usable data. But outcome-based web scraping APIs eliminate this wastage by using smarter strategies, by absorbing failures internally and by returning only successful results.

This is where the traditional web scraping software stack collapses. Proxy management, browser automation, unblocking, parsing, and retry logic are no longer separable concerns. They're interdependent layers that must work together or fail together. They must be selectively and dynamically combined based on site behavior, because over-stacking quickly increases cost without improving outcomes. The cost of orchestrating them has become higher than the cost of the components themselves.

The market response confirms this shift. In 2025, volume of requests through Zyte API grew by over 130% year-over-year, a clear signal that the industry is migrating from DIY component management to unified data outcomes.

Shift	What changes
Proxy abstraction	Proxies become embedded inside scraping APIs rather than configured directly
Vendor strategy	Proxy providers expand into platforms or compete as commodity infrastructure
Value concentration	Differentiation moves to orchestration and cost-per-success optimization
Developer behavior	Manual tuning declines in favor of outcome-based APIs

Implications

Component commoditization enables API-level abstraction. As proxies, browsers, unblocking services, parsing tools, and retry logic gets standardized and lower-cost, each is increasingly bundled behind inside scraping APIs that hide component-level complexity and optimize for data outcomes rather than infrastructure control. The future stack looks remarkably simple: call an API, receive a clean JSON, build a data pipeline.

Proxy vendors move up the stack – or down the value chain. As proxies become embedded features rather than standalone products, vendors will either expand into higher-level platforms or compete on price as infrastructure suppliers. In 2026, consolidation will accelerate and many point-solution proxy providers will exit the market.

Control becomes an illusion. Many developers still cling to manual proxy management like a security blanket. But this attachment now represents nostalgia more than necessity. In practice, hand-rolling proxy rotation, managing ban lists, and tweaking headers now looks like costly indulgence. Smart developers will focus on what matters: data quality, schema validation, and business logic. Winning teams delegate, not dabble.

Recommendations

Stop optimizing proxy configurations.

If you're spending engineering time tuning proxy rotation, managing subnet diversity, or debating IP types, you're fighting a war that's already lost. Redirect that effort toward data quality, schema validation, and integration logic. APIs handle proxy optimization better than humans ever could.

Evaluate API-first platforms over component-based infrastructure.

By 2026, the ROI calculation will increasingly favor platforms that bundle proxies, browsers, unblocking, and parsing into a single interface. These platforms amortize the cost of anti-bot adaptation across thousands of customers, making them cheaper and more reliable than in-house solutions. Our study shows that moving the bulk of traffic to a full-stack web scraping API can drive engineering effort spent on access management down to around 10% – freeing the engineers to work on higher-value work.

Plan for the transition now.

If you're currently managing your own proxy pools, begin evaluating managed alternatives. The transition period is where most organizations struggle - don't wait until your proxy infrastructure becomes unmaintainable to start exploring options.

> As a web scraping professional, the thing I prided myself on was my ability to build scrapers myself.
>
> But customers do not care how good my scraping skills are. They care about getting the data that they need reliably, consistently, and on time.
>
> Despite my initial skepticism, migrating to a web scraping API is the smart way savvy engineers move the needle for their business.

John Rooney

Developer Engagement Manager, Zyte

The proxy, once scrapers' great master key, has become just one piece of an increasingly complex puzzle. Suddenly, proxy management becomes someone else's problem.

Developers who once spent hours fine-tuning proxy pools can make HTTP requests and receive structured JSON.

Senior Editor (Content & Editorial), Zyte

Robert Andrews

Read more

[Death of the proxy? There's an API for that]()

DIY scraping has hidden costs, including developer time, broken scripts, and proxy maintenance. Scraping APIs eliminate those.

With web scraping APIs, what you pay for is successful data - no more time or dollars wasted on blocked requests or proxy subscriptions that go nowhere.

Product Marketing Manager, Zyte

Daniel Cave

Read more

[7 myths about web scraping APIs – busted]()

Zyte API

Zyte API turns the chaos of web scraping into a single API call. Send any URL and get back clean, structured JSON, quickly.

Automatically handles smart proxy rotation, blocking, and headless browser rendering, with AI that parses content from tricky pages.

Named the fastest, most effective and cheapest web scraping API on the market by Proxyway's Web Scraping API Report 2025.

[Discover more]()

#2
Trend

AI is the new engine for web scraping

The experiments are over. In 2026, AI is the foundation for a reinvented scraping toolset.

Surveys by Google's DORA and StackOverflow confirm it - up to nine in 10 software developers now use AI in the development process. LLM-powered code editors have quickly revolutionized the job.

But web data gathering has never quite been as self-contained as software development at large - data scraping is less a product, more a pipeline, a process. Those of us who embraced AI early for scraping often found it ill-suited to our particular workflow.

In 2026, that is changing. Now we are seeing AI that better caters to all the discrete individual cogs of the data machine. That includes equipping IDEs with specialist scraping know-how - this year, scraping engineers' code gets to play a full part in the copilot party.

Covering all these bases paves the way for the ultimate prize - agentic AI that runs the full gamut of data-gathering workflows, end to end; a data-gathering machine that thinks for itself.

Key developments

According to a November 2025 analysis by Technavio, AI-based web scraping is projected to reach $3.16 billion by 2029, growing at 39.4% annually - a sign that the market has moved beyond experimentation.

That is because, piece by piece, we are seeing AI enhance all the links in the chain:

- Auto-classifying page content for schema-specific field extraction.
- LLM-powered extraction of unstructured page data.
- Automated identification of on-page selectors and field mappings.
- Change detection and revision for page markup alterations.
- Generating crawler and scraping code.
- Natural language, not code, is becoming a viable primitive for browser interaction.
- Data cleaning through smart normalization.
- Leaps forward in data quality validation and anomaly detection.
- Smart real-time unblocking strategies for maximum success.

Data users will gravitate toward “easy”. This trend is visible in our own data at Zyte: for example, usage of AI-driven discovery for job listings content - a critical building block for these pipelines - surged over 50x in 2025.

The result: AI components now sit across the entire scraping lifecycle. Planning and orchestration, crawling, unblocking, extraction and validation - each stage has AI tools that reduce manual work.

Implications

LLM-powered extraction finds its sweet spot

This year, more teams will be ready to adopt LLM-based extraction as a reliable approach to handle less-structured webpages. Some teams are sending HTML content directly into LLMs with their own prompts and data models. Some teams adopt one of the growing set of extractor APIs now available in the market. Research shows that AI extraction methods now provide resilience to layout changes, natural language extraction without parser logic, and the ability to handle unstructured content - capabilities that traditional CSS selectors simply don't have. But these benefits still come at a cost: LLM-powered extraction burns tokens on every page and, as a probabilistic system, can introduce inconsistency over long crawls. Engineers will need to understand when to use this approach and when not to.

Enhanced LLMs automate gnarly spider code generation

Fortunately, code generation tools have arrived in scraping, bringing new AI efficiency to spider production. Tools like Web Scraping Copilot, launched by Zyte in November last year, add specialist scraping expertise that generic LLMs lack - allowing developers to input sample URLs and natural-language extraction instructions to receive working scraper code in return, complete with all the appropriate selectors and logic. This approach preserves the advantages of traditional scraping - deterministic behavior, schema stability, and low per-page cost - while dramatically reducing development and maintenance effort.

Ours won’t be the only tool giving teams the code they used to write and rewrite by hand. As more code generation comes on-stream, in 2026, writing scraper code from scratch will feel outdated. We expect data engineers and

Headless browsers get brains and eyes

AI-native browser engines and browser interaction frameworks are now helping tackle more of the decision-making needed to complete a task. Here's a common scraping browser challenge: should the browser wait for certain elements, scroll, click a button? Instead of a human conductor instructing every click, browsers like Lightpanda and frameworks like Stagehand are built to reason about these questions, determining the best course of action for themselves.

Some scraping projects require navigating multi-step, stateful flows - like applying filters, entering form inputs, and progressing through gated screens - before data is ever visible. Vision-based computer-use models are improving rapidly to address these scenarios. By interpreting forms, buttons, dialogs, and dynamic UI states, these models enable automation of long-tail workflows that previously relied on brittle browser automation scripts.

Data quality goes through the roof

In 2026, teams will more widely adopt AI to validate extracted data, detect anomalies, and enforce schemas. This catches errors that would have slipped through manual QA and reduces the need for human review. The AI In Data Quality market is projected to grow at 22.9% annually through 2029, according to Technavio, indicating that organizations are already investing heavily in AI-powered validation and quality assurance.

End-to-end agentic data acquisition made possible

With these key components - data interpretation, code generation, intelligent browsing and quality control - having gone from specs to working systems, the stage is set for an AI data step-change: agents that will connect them together, in concert, into a radically different whole. Increasingly, engineers won't need to write, re-write and re-deploy ad infinitum; they will supervise autonomous agents to do that on their behalf.

Recommendations

Use LLM-powered extraction for low-volume projects with complex and volatile sites.

If a site changes layout frequently or relies on loosely structured content, LLM-based extraction can be an effective option. While the cost per request is higher - as this approach involves sending webpage content to LLMs to process - it can be a strong fit for low-volume, lower-stakes tasks such as sales lead research, where speed and flexibility matter more than perfect consistency. This approach reduces reliance on engineering time, but costs and output quality should be monitored closely. Always evaluate ROI for your specific use case.

Adopt AI-powered code generation for high-scale, mission-critical projects.

For large crawls and low-tolerance use cases, use tools like Web Scraping Copilot to generate spider code rather than to extract data at runtime. Generated code can be tested, versioned, and run cheaply at scale, delivering consistent results while still accelerating engineering work.

Use computer-use, vision-based extraction for projects with complicated browser workflow.

If you're struggling with sites that require multi-step browser actions and render content dynamically, try a screenshot-based computer-use framework. It's not always cheaper, but the reliability could be worth the trade-off versus DOM-based approaches.

Invest in AI-powered validation.

As extraction methods diversify, AI will be useful in validating extracted data and detecting anomalies. You can only scale development as fast as you can scale QA.

The opportunity lies in finding the right balance between the intelligence of AI and the control required by engineers.

We're using Web Scraping Copilot in our own professional services team to automatically generate and test extractor code - it's boosting productivity 2.5x. Support for navigation code is being added.

The next major milestone is automated spider maintenance. With this, we see productivity increasing from 2.5x to 3x, to 4x and beyond.

Iain Lennon
Chief Product Officer, Zyte

LLMs can help by reading page content to guide crawler discovery, while fuzzy extraction makes scraping work more like human-reading than machine-powered parsing.

Now you can automate spider generation, replacing manual work that scales linearly with the number of sites, while tools like spaCy, fastText and ydata-profiling help classify, clean and detect anomalies in natural-language text output.

Theresia Tanzil
Web Scraping Strategist, Zyte

Read more

Four sweet spots for AI in web scraping

Web Scraping Copilot

A free Visual Studio Code extension that transforms generic AI assistants and LLMs into web scraping specialists, generating production-ready spiders up to three times faster.

Full code transparency with scrapy-poet PageObjects and automatic test fixtures, plus integrations with Zyte API anti-ban handling and Scrapy Cloud deployment, right in your code editor.

Discover Web Scraping Copilot

Zyte API - AI Extraction

Use natural language input to extract any content with a single parameter.

Includes patented machine learning to return structured JSON from any web page, without writing selectors - delivering product, article, job posting, and SERP data that is cheaper and more accurate than LLMs.

See more

#3
Trend

Birth of the autonomous data pipeline

In 2026, web scraping is evolving from AI-assisted efficiencies for individual parts of the process, to autonomy for entire web data pipelines.

End-to-end automation will become the default trajectory for web data pipelines, as agentic scraping shows its potential as an autonomous loop that keeps data deliveries healthy, while humans specify goals, design technical constraints, and define acceptable risks.

Deloitte's 2025 Emerging Technology Trends study found that, while 30% of organizations are exploring agentic approaches and 38% are piloting, only 11% have production deployments. This gap will narrow substantially through 2026. The autonomous agents market is forecast to grow from $4.35 billion in 2025 to $103.28 billion by 2034, with agentic AI expanding at a 44.6% compound annual growth rate.

Picture the new scraping workflow:

- A data team specifies an outcome - a dataset with a schema, coverage targets, freshness, and failure tolerance.
- An AI agent explores the site, discovers what actions are necessary to locate the data, and chooses the cheapest reliable method to fetch it: direct requests where possible, browser interaction where necessary.

In scraping, agents will discover convenient existing website APIs, isolate the relevant endpoints, and propose an efficient extraction plan. When the site changes, the agent won't simply fail; it will diagnose breakage, regenerate code, re-validate outputs, and escalate only when confidence drops below a threshold.

Key developments

Over the past year, tool-building for AI agents has accelerated across the software landscape, and web scraping is following the same trajectory. Agents can now treat the entire scraping stack as a toolbox - browser execution, Document Object Models (DOM) analysis, and the team's own data validation codebase.

In practice, agentic scraping will be more robust as a multi-agent system than a monolith - not a single scraping agent, but a team of specialist agents under an orchestrator. As specialized agents proliferate, teams will combine them into a coordinated architecture where each agent does one job well, while a reasoning supervisor agent routes work, maintains state, and enforces guardrails across the workflow.

API discovery agents – Agentic "API self-discovery" is rising as a general development paradigm, and scraping benefits disproportionately: once an agent identifies the right endpoints, it can swap brittle UI automation for stable API pulls. We are seeing tools built to capture network traffic and catalog API calls automatically – exactly the substrate an agent needs to move from "browse" to "extract."

Schema-first extraction agents – As research like PARSE (EMNLP Industry 2025) makes clear, LLM-driven schema optimization can make entity extraction more dependable.

Self-healing testing agents – The testing world demonstrates the pattern through self-healing browser automation tests that adapt to UI changes by using model reasoning over current application state, preventing unpredictable breakage from stopping data collection.

Vision-based computer-use agents – Models like Google's Gemini "computer use" can "see", click, type, scroll, and navigate independently, which makes them effective for unfamiliar, UI-heavy flows and long-tail interfaces with no usable API.

DOM-native browser agents – Instead of "pixels first," browser agents operate on browser primitives like DOM, network events, and local storages. This approach is typically cheaper and offers more consistent results than computer-use agents.

Coding agents – Because code is the connective tissue of any data pipeline, coding agents are poised to become the backbone of agentic scraping. Early signals are already visible in the emergence of scraping-specific assistants built around scraping-specific patterns and pulling different pieces of the workflow together. What's making this possible is enhancements to general-purpose language models, which now reign across different major coding benchmarks.

Implications

Workflows become modular and context-aware. Rather than monolithic scrapers, systems will comprise specialized agents such as CAPTCHA handlers, behavioral intelligence, DOM analyzers, and session managers.

Pipelines will adapt dynamically - for example, falling back from rendering to simple fetches, switching extraction methods based on page structure, or retrying with alternative access patterns. Context will flow between components, enabling intelligent decision-making across the workflow.

Specialized agents take on larger, end-to-end roles in production scraping. Scraping-specific agents will handle larger portions of the increasingly modular workflow, combining tasks such as discovery, code generation, validation, and iteration into more autonomous units of work. While general-purpose models are rapidly improving, production scraping continues to benefit from domain-specific tooling, context, and guardrails rather than raw model capability alone.

The human role in agentic scraping shifts from implementation to supervision and accountability. Rather than writing selectors and retry logic, engineers will instruct, evaluate, and monitor the agentic systems. This shift reflects a broader change in how technical work is organized around AI systems. Ownership shifts from "who wrote the scraper" to "who owns the data product," with clearer SLAs, auditability, and decision logs for what the system did and why.

More automation increases website pressure and fragmentation. As autonomous agents proliferate, sites have stronger incentive to harden interfaces, gate access, and formalize automation lanes, reinforcing the macro forces behind escalating anti-bot dynamics and giving rise to the fragmentation of the web.

Recommendations

Apply agents selectively.

Not all scraping tasks will warrant agentic approaches. Agentic approaches are not a default for every scrape. For sources that are straightforward and stable, a conventional scraping setup will remain the most cost-effective option.

Match agent types to the job.

Combine approaches based on best fit and tool maturity. For example, computer-use agents are best suited for site exploration and complex interactions where probabilistic behavior is acceptable. Reserve code-generation systems for high-volume, repeatable extraction where deterministic output and cost predictability matter.

Build supervision capabilities.

Establish tools for evaluating agent performance, enforcing schema constraints, and implementing feedback loops. Human oversight should shift from code generation to output validation and system guidance.

Pilot agents in bounded roles with explicit success criteria.

Deploy agent capabilities first where they are easiest to evaluate and safest to contain: site exploration, endpoint discovery, schema mapping, and test generation. Treat "self-healing" as a phased rollout. Start with agent-proposed fixes that require approval, then move to limited autonomous fixes in low-risk segments once the evaluation harness consistently catches regressions.

Iterate toward agent-native workflows.

Organizations succeeding with agentic scraping will typically start by integrating agents into existing workflows, then progressively evolve toward more agent-native designs as confidence, tooling, and reliability improve. This requires rethinking task structure, context provision, and output validation. The technology matters, but process redesign is equally critical.

2026 is the year web data becomes truly automated. Not just faster scrapers, but AI creating, fixing and scaling them - from a site name to working production code - and then keeping it running as the web changes.

The next evolution is already here: agents that transact. When you give an agent a wallet, you give it economic identity. By supporting x402, Zyte will empower the coming generation of agents to carry out and pay for on-the-fly web scraping operations.

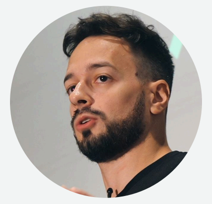
Jan Seidler

Chief Technology Officer, Zyte

Read more

[Why your agent deserves a wallet]

Web scraping presents a set of challenges that push current AI agents to their limits. However, delegating these hard tasks to specialized sub-agents or tools is a very effective strategy. Our goal is to apply them to fully automate customizable web scraping. We are adapting our tools to be more usable by LLMs, working on a system that can create a plan for crawling complex websites based on a simple user query, and exploring ways to use tools and MCPs to automatically find the best browser configuration for a given website.



Iván Sánchez

Senior Data Scientist, Zyte

Read more

Why AI agents struggle with web scraping (and how to help them)

In its first wave, agentic web scraping faced several challenges. But, for scraping at scale, agentic effectiveness may be just around the corner. Getting there will require equipping agents with robust, domain-specific, scraping MCP tools; tuning them to orchestrate with minimal human intervention; re-engineering with context limits in mind and building interfaces that go beyond chat.

Konstantin Lopukhin

Head of R&D, Zyte

Read more

Agentic web scraping: Hype, reality and what happens next

Zyte API and x402

X402, Coinbase's open standard for payments, allows APIs, apps, and AI agents to transact seamlessly over HTTP.

Zyte integrates with x402, empowering agents to dynamically purchase structured web data, such as market insights and product listings, via micropayments.

See it in action

Scrape via MCP

Empower your LLM or agent to extract live web data using natural language instructions, by calling Zyte API through a Model Context Protocol (MCP) server using FastMCP, Zapier or more.

Read more

#4

Trend

Automation drives power in the data arms race

Anti-bot systems now change faster than human teams can respond. Scrapers must counter with their own automation, or fail at scale.

The relationship between scrapers and bot mitigation systems has entered a new phase. For years, it operated at human timescales. Countering a ban by a website owner was manageable, and the restored access lasted long enough to justify the engineering cost - until a reciprocal counter-ban broke the pipeline again.

By 2026, this cadence no longer holds. Anti-bot systems are reconfiguring their detection mechanisms continuously, driven by machine learning models that adapt in as little as a few minutes. Proxyway's 2025 report captured the new reality perfectly, half-jokingly noting: "Two days of unblocking efforts used to give two weeks of access... now, it's become the other way around."

Our internal data confirms it. We observed one major bot management vendor deploy more than 25 version changes over a 10-month period, often releasing updates multiple times a week. Manual configurations that once lasted weeks now fail on a daily basis, especially on high-value, frequently targeted sites. Cloudflare is known to deploy a near-real-time system that adapts bot detection strategy every few minutes. Azure's Web Application Firewall updates IP rulesets multiple times daily. This creates a critical vulnerability: manual scraping strategies fail more frequently and become expensive to maintain.

But the stark reality is that, while defenders have fully automated, scrapers, to a large degree, have not.

The gap between defense automation and scraper automation is the defining constraint of 2026

Key developments

Three factors amplify the speed mismatch.

Machine learning-driven detection is proliferating. We observe protections now increasingly incorporate polymorphic JavaScript, WASM obfuscation, Runtime Application Self-Protection (RASP), and passive fingerprinting at scale. These techniques change constantly, making them difficult to hone and update manually.

Detection mechanisms are expanding. Beyond traditional IP blocking and CAPTCHAs, defenders now monitor subtle technical signals, like timing patterns, network-level anomalies, device fingerprint consistency, pointer curves and scroll variance. The tiniest mismatch across multiple dimensions might trigger a block.

Bot traffic volume is creating urgency. AI bots are taking up a larger share of overall internet traffic. Sites' traffic management systems are responding with continuous, automated tuning.

Implications

Manual access strategies will be unsustainable at scale. Organizations relying on static fingerprints, fixed headers, or manual retry logic will face an escalating rate of failures. Breakages will increase, maintenance load will grow, and the shelf life of any configuration will shorten. For complex, high-value targets, manual approaches will require constant intervention - an operational burden that grows faster than resources can scale. Teams need systems that continuously monitor their own performance, detect degradation, test alternatives, and adapt without human intervention.

Access configuration is becoming a first-class technical building block. Success now depends on adaptive configuration. The "access layer" must be treated as its own module in the scraping stack, equipped with its own monitoring, testing, and self-repair capabilities. Rather than hard-coding access strategies, teams will need systems that detect when a configuration fails and automatically test alternatives: switching from browser rendering to simple fetch, rotating through different fingerprint sets, adjusting scraper signals, or escalating to human intervention when necessary.

Only automated, self-adjusting pipelines survive at scale. By 2026, scraping systems - not scraping teams - must tango with dynamic defenses. These systems will maintain healthy fingerprint pools, understand IP reputation at the ASN level, and score sessions for reliability. They'll incorporate the appropriate behavioral signal and variance across multiple modes of user interaction. And they will enforce cost guardrails by deciding dynamically whether to render pages or fetch directly, based on cost and success probability. The engineering effort has shifted from “solving the site once” to “continuously adapting the solution”.

Machine identity will become a decisive factor. Call it a passport for your pipeline. Unsigned or unverifiable agents will receive heightened scrutiny. Bot mitigation systems are increasingly distinguishing between verified bots (search engines, analytics), AI bots (training, search, user action), and unverified scrapers. By 2026, the ability to present a coherent, consistent machine identity will separate successful operations from blocked ones.

Recommendations

Invest in automated access configuration and orchestration.

Build or adopt systems that optimize session configuration for the leanest working setup, perform cost-aware switching between browser and non-browser strategies, and self-heal after failures.

Maintain a crawler identity management process.

Keep sets of valid fingerprints updated and regularly refresh IP reputation sources. Use free IP reputation feeds with ASN-level checks.

Incorporate detection signals directly into access strategies.

Understand which signals are monitored and incorporate them into your strategies. Prepare to integrate multi-dimensional scraper signal modeling. Align requests with timing, protocol, and device fingerprints expected by modern defenses.

Understand the cost curve of different access approaches.

Know when to escalate from HTTP to headless browsing, when to retry versus abandon, and when to switch strategies. The leanest working configuration is the most cost-efficient. Monitor success rates, not just bandwidth consumption.

> Agentic workflows will lead to evolving UX patterns, potentially introducing new challenges for scraping.
>
> As agents become more adept at handling CAPTCHAs, vendors that currently rely on proof-of-work may pivot toward alternative mechanisms that preserve user experience while distinguishing humans from bots.

Akshay Philar

Head of Engineering, Zyte

Zyte API - Ban-handling

Zyte API automatically manages access challenges using machine-learning models trained on historical success patterns.

It selects from more than 320,000 combinations of access strategies through smart proxy routing, a headless browser fleet, CAPTCHA handling and more.

For each request, the system applies only the leanest successful measures, maintaining 98% to 99% success rates while minimizing unnecessary costs.

Discover ban-handling

#5
Trend

Web traffic splinters into access lanes

A plethora of autonomous agents is set to claim unprecedented traffic share. In their wake, new judgements about the intent and economic merits of diverse new programmatic visitors will usher in new access regimes for different bot species. Welcome to the new walled web.

For decades, website owners and scrapers had a simple relationship: websites published content; “good” scrapers accessed it politely, “bad” scrapers abused it. By 2026, this framing is becoming less useful.

The rise of autonomous crawlers, LLM browsing agents, shopping agents, and MCP-connected tools has created a new reality: websites can no longer afford to treat "bots" as a homogenous category - either “good” or “bad”. For website owners, different types of automated traffic generate different economic value and pose different risks.

Website operators, then, are coming to acknowledge diversity in the bot population, and are re-drawing the rules for how they welcome programmatic traffic.

Key developments

A huge share of the web will continue operating as it always has but, as AI-driven data access scales, a growing portion of the sites is nevertheless reorganizing into three new regimes:

The hostile web escalates defenses against abusive automation. These sites deploy aggressive honeypot traps, AI-targeted challenge flows, and increasingly sophisticated fingerprinting. Some search services are sending clear adversarial signals toward automation – steadily redesigning their search experiences to raise the cost and friction of automated access.. Meanwhile, Cloudflare rolled out traps for AI crawlers to over 1 million websites, boasting to have blocked 416 billion AI bot requests in six months alone. The message is clear: for publishers bearish on becoming data providers, visitor friction can be enabled at the flip of a switch.

The negotiated web emerges from economic pressure. Publishers facing declining search traffic or rising costs from AI crawlers indexing their sites adopt licensing, attestation, pay-per-crawl, paywalls, and attribution mechanisms. Creative Commons recently announced tentative support for pay-to-crawl systems, and Adweek reports that 2026 will see LLM deals shift from one-time training payments to usage-based revenue shares. New standards like ai.txt, llms.txt, and Really Simple Licensing (RSL) are attempting to make permissions machine-readable, but walled-garden data ecosystems may restrict machine access except via licensing, API, or verified bot status.

The invited web turns agents into first-class distribution channels. Sites, actively inviting programmatic access to desirable actors, expose machine-first interfaces for approved actions and real-time data. E-commerce platforms are leading this shift. Shopify, Google, Visa and Stripe along with OpenAI all now either support Model Context Protocol (MCP) or have launched their own protocols for AI shopping agents - Stripe's Agentic Commerce Protocol (ACP), Google's Universal Commerce Protocol and Visa's Trusted Agent Protocols. E-commerce is the first tangible sphere in which these access lanes are set to become valuable off-platform product data sources in their own right. But the same "invitation" pattern is likely to spread to other content and service categories, as websites work towards gaining more visibility in the age of AI-mediated information discovery. Going forward, expect more site owners with valuable data to make themselves available to approved agents through these kinds of structured programs.

Implications

Identity becomes a first-class citizen. New identity and attestation layers emerge. Expect standards and products for verifying bots and signing agents - initiatives like "Know Your Agent" will certainly gain traction. Verified, authenticated, or attested bots will receive preferential routing while unsigned or unverifiable bots face heightened friction. For many, machine identity will no longer be optional; it's operational.

Intention becomes a bargaining chip. Agent utility, not just legitimacy, matters. A shopping agent bringing qualified buyers is treated differently from a training crawler. Websites evaluate whether an agent's purpose aligns with their business model and data strategy. This shifts the conversation from "can you access?" to "should you access, and on what basis?"

The web becomes economically differentiated. Websites no longer operate under a single access policy. This will pave different paths for different agents. Some content remains broadly scrapeable but more guarded, other content is locked behind licensing or partnership agreements. Still other content is designed specifically for agentic interfaces. For data gatherers, this fragmentation breaks the idea of a single web access strategy.

Standards proliferate but enforcement remains uneven. ai.txt, llms.txt, RSL, MCP, and ACP all attempt to standardize machine-readable permissions. Adoption is growing but uneven; thus far, major AI providers have not universally honored these standards. However, the trajectory is clear: standardized, machine-readable access agreements will become increasingly common, particularly in commerce and publishing.

Recommendations

Map your data sources against the three new access regimes.

For each data source in your pipeline, determine whether it now belongs in the “hostile”, “negotiated”, or “invited” web buckets - or in none at all. Evaluate the long-term path based on technical difficulty, breakage risk, maintenance burden, and legal friction. The cost of acquiring web data must be compared against licensing costs, API fees, and partnership opportunities.

Build organizational capabilities.

Organizations must build capabilities across all three regimes. This means maintaining robust scraping infrastructure for hostile-web targets, developing identity and attestation capabilities for negotiated-web access, and integrating with agentic commerce protocols where applicable. The single-strategy approach no longer works.

Resolve the discoverability paradox for your own web assets.

Decide which automated systems you welcome and which you block. Design your interfaces, metadata, and feeds accordingly. If you want to be accessed, make it frictionless. If you want to negotiate, expose licensing endpoints. If you want agentic integration, implement the relevant protocols such as MCP and ACP.

Monitor standards evolution closely.

ai.txt, llms.txt, RSL, and emerging licensing frameworks will shape the negotiated web. Early adoption of supported standards positions you for better access terms and lower friction as these standards mature.

Zyte's polite products

Step one in Zyte's best practices for web scraping is: "Don't be a burden."

Zyte API limits the rate at which websites can be accessed, to safeguard against overwhelming publishers.

Zyte Data's in-house scraping experts carefully estimate sites' traffic tolerance to consider the impact of bots.

Read more

#6
Trend

Legal clarity arrives, with compliance demands

In 2026, major new regulations will take effect across multiple jurisdictions. Organizations using web data for AI will need to adapt to mandatory transparency, copyright respect, and provenance documentation.

2026 marks a turning point. California's Assembly Bill 2013 took effect January 1, 2026, while the EU AI Act's core obligations take effect August 2, 2026. These are binding legal requirements with enforcement mechanisms and significant penalties.

For organizations developing AI systems, compliance infrastructure is no longer optional. By mid-2026, operating without documented data provenance and compliance systems will create material legal risk.

Enterprises will not adopt AI systems without evidence of lawful data sourcing, and regulators will enforce actively. From this year, organizations that build compliance into their operations will now have competitive advantage.

Key developments

California AB 2013 mandates specific disclosures for generative AI data.
Developers of publicly available generative AI systems must publish detailed

documentation, including their data sources, dataset size, data types, whether data includes copyrighted material, whether datasets were purchased or licensed, whether personal information is included, and data processing methods used.

The EU AI Act imposes transparency and other obligations on AI service operators, based on risk to users' health, safety, and fundamental rights. All general-purpose AI model providers must publish "sufficiently detailed summaries" of training datasets and respect copyright holders' opt-outs. Providers cannot use copyrighted content if the rights holder has indicated non-consent. Penalties reach €35 million or 7% of global annual turnover.

Copyright litigation clarifies the boundaries. In the US, the 2025 Bartz v. Anthropic ruling established that training on legally obtained works is defensible, while training on pirated content is not. In Kadrey v. Meta, the court emphasized market harm as one of the decisive factors influencing the likelihood of a copyright breach ruling. However, fair use defenses will continue to be aggressively litigated. Organizations cannot assume blanket protection.

Regulators enforce actively, but make room for training with personal data. The French Commission Nationale de l'Informatique et des Libertés (CNIL) fined Kaspr, a B2B lead provider, €200,000 in 2025 for data scraping violations. This is not an isolated incident - regulators across jurisdictions are increasing enforcement activity. Organizations operating in Europe, California, or globally face real enforcement risk if they cannot demonstrate compliance

However, CNIL also made clear that training AI models on personal data sourced from public content can be lawful under the GDPR's legitimate interest basis, provided certain conditions are met. So there is a path forward for data scrapers and AI services to obtain public personal data lawfully under the GDPR.

Enterprise buyers demand provenance. Large organizations increasingly require evidence of lawful data sourcing before adopting AI systems. This is a procurement requirement. Enterprises face their own regulatory exposure and will not accept suppliers without documented compliance. This will create a new market signal: provenance is a competitive requirement, not a legal nicety.

Personal data handling remains strictly regulated. Despite some regulatory relaxation proposals, personal data handling remains tightly controlled in 2026. Identification, profiling, and biometric data trigger strict compliance obligations. Personal data handling will become a visible differentiator and enforcement priority. As noted above, a proper legitimate interest analysis (and in some cases a Data Protection Impact Assessment or DPIA) will be enough to satisfy the compliance burdens for public personal data.

Implications

Compliance becomes a core operational requirement. As it must be embedded across product, engineering, and data workflows, operational complexity and overhead grow. For many organizations, this will make partnering with specialized data providers more attractive than building and maintaining compliance infrastructure in-house.

Provenance tracking is a new foundation. Investors, auditors, and enterprise customers will demand evidence of lawful sourcing. By 2026, organizations without provenance tracking will face friction in capital raising and partnerships.

Global divergence forces standardization on strictest rules. Organizations operating globally must comply with the strictest overlapping standards. Building to EU AI Act standards will satisfy most jurisdictions. However, litigation is centered in the US, so tracking and following the results of the US litigation is critical as well.

Licensing markets accelerate. As legal risks of scraping increase, some organizations will increasingly seek formal data access agreements. By 2026, more standardized licensing frameworks and revenue-share models will emerge. The cost could be higher than scraping, but legal and operational risks are lower. This will bring about concerns around open access to public data, and compliant web scraping will emerge as a very important tool to keep open access alive.

Recommendations

Conduct a comprehensive audit of your training data.

Identify the source and legality of all data. If you cannot document lawful sourcing, stop using it. Pirated or stolen data is indefensible.

Implement provenance tracking systems.

Build infrastructure to document the origin, legality, and usage of all data. This must be auditable and transparent.

Respect copyright signals and opt-outs.

Monitor for machine-readable signals indicating "do not use for AI training." Implement systems to honor these signals where possible.

Ensure lawful basis when accessing personal data.

Identify all personal data in your datasets and conduct any required compliance analyses, such as an LIA or DPIA. Additionally, implement required data subject access, security, and data minimization protocols.

Design your systems for auditability.

Build your data infrastructure to support disclosure requirements. Maintain detailed records of sources and usage. Implement governance systems that can demonstrate compliance to regulators and customers.

Courts are increasingly ruling that scraping public web data is acceptable, even recognizing fair use in the context of training AI. But freedom comes with responsibility. Legal re-use of scraped data must materially transform it from its original shape and/or use it for non-competitive purposes to the source material. Do not use scraped data for AI products prohibited under emerging regulations like the EU AI Act.

Sanaea Daruwalla

Chief Legal Officer, Zyte

Read more

Balancing innovation and regulation in data scraping

This is a pivotal moment for web data collectors. Never before has web scraping attracted such widespread global attention. European data protection authorities (DPAs) are concerned that web scraping in relation to AI may inadvertently collect personal data, and have published a wave of guidance. Companies must follow all applicable laws and regulations.

Victoria Vlahoyiannis

Senior Legal Counsel, Zyte

Read more

What Europe's privacy regulators say about scraping personal data

Zyte legal compliance

All Zyte API Enterprise customers receive compliance onboarding.

Zyte API automatically blocks login for many sites whose terms of service prohibit web scraping.

Zyte API's AI Extraction is configured not to extract personal data fields in most cases, while the most commonly copyrighted data - including image, video, PDFs and music - is also excluded.

Read more

Discover compliant scraping

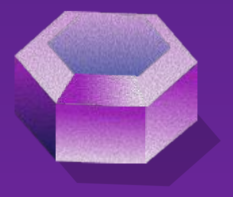
Summary

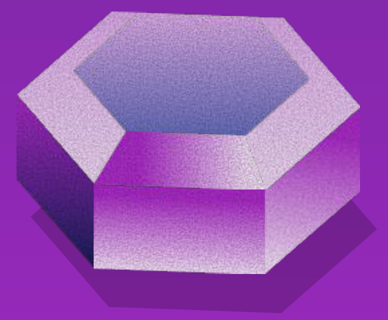
The six trends traced in this report - from the collapse of the component-based scraping stack to the emergence of regulatory frameworks - tell a single story: accessing web data is moving from a frontier practice to critical, governed infrastructure.

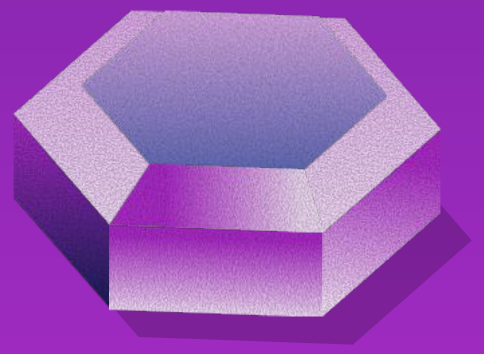
This maturation will create clear winners and losers. Organizations that invest in compliance infrastructure, adopt full-stack data acquisition platforms, and build strategic partnerships will enjoy access to data that others do not.

Building on present AI foundations, the dawn of the agentic extraction era is likely to see a blossoming of seemingly organic, semi-autonomous web scraping varieties in which whole ecologies of smart actors request, obtain and transact in the data that is the lifeblood of business.
2026 marks a turning point to a new frontier.

Takeaways







Scraping

Data outcomes replace the old scraping stack

- Stop optimizing proxy configurations.
- Evaluate API-first platforms over component-based infrastructure.
- Start planning the transition now.
- Focus on data outcomes, not infrastructure components.
- Prepare for compliance and governance assessment in the procurement process.

AI

AI is the new engine for web scraping

- Use LLM-powered extraction for low-volume projects with complex and volatile sites.
- Adopt AI-powered code generation for high-scale, mission-critical projects.
- Use computer-use tools for projects with complicated browser workflow.
- Invest in AI-powered validation.

Agents

The dawn of the autonomous data pipeline

- Apply agents selectively.
- Pilot agents in bounded roles with explicit success criteria.
- Match agent types to the job.
- Build supervision capabilities.
- Iterate toward agent-native workflows.

Access

Automation drives the balance of power in the data arms race

- Invest in automated access configuration and orchestration.
- Maintain a crawler identity management process.
- Incorporate detection signals directly into access strategies.
- Understand the cost curve of different access approaches.

Content

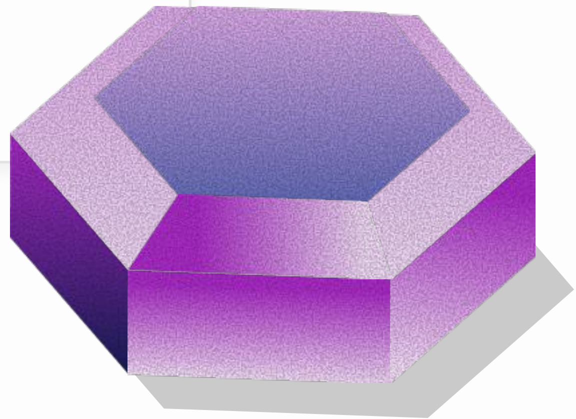
The web traffic splinters into access lanes

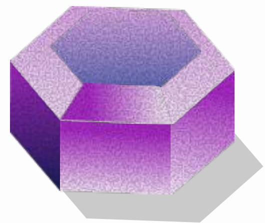
- Map your data sources against the three new access regimes.
- Build organizational capabilities across all three regimes.
- Resolve the discoverability paradox for your own web assets.
- Monitor standards evolution closely.

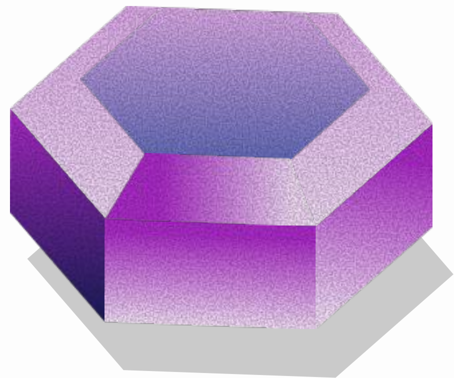
Legal

Regulatory clarity arrives - with new compliance demands

- Conduct a comprehensive audit of your training data.
- Implement provenance tracking systems.
- Respect copyright signals and opt-outs.
- Ensure lawful basis when accessing personal data.
- Design your systems for auditability.







Access the web's data

Zyte turns websites into data with industry-leading technology and services.
Over 15 years, Zyte has helped more than 4,500 businesses gather data from five trillion pages.

Products

Zyte API

Everything you need in one, AI-enabled toolkit. Automate discovery, unblocking and extraction. Hassle-free scraping, full control.

Get started

Web Scraping Copilot

A complete, production-ready spider workflow inside Visual Studio Code, from AI-generated code to cloud deployment.

Download now

Scrapy Cloud

Scalable cloud hosting for your spiders. Provides a simple way to run your crawls and browse results.

Deploy now

Services

Zyte Data

The feed you need for your dream data. When scale and speed matter, let Zyte's in-house expert engineers build, deliver and maintain your custom web data pipeline.

Contact us

Stay in touch

Zyte blog

Insights, updates and tips from the new data frontier.

Read it now

Subscribe for free

Follow us

Talk to us

How do *you* see the year ahead?

Tell us your project’s requirements.

Get in touch