ZYTEPAPER

E-commerce case study

Product price, promotion and positioning monitoring at scale

How Zyte built a data extraction system to monitor over 1 billion products each day.

Overview

Web scraping in e-commerce is big business these days. E-commerce companies are increasingly using web data to fuel their competitor research, dynamic pricing and new product research.

As a result, there is a huge demand for service providers who can help them monitor their pricing, placement and promotion of their products in comparison to that of their competitors.

However, extracting product data at a large scale can be an extremely challenging specialised problem to solve due to the volume of data being extracted and the need for up to date high-quality data. Consequently, a lot of e-commerce and consumer brands turn to product monitoring providers instead of developing a competitor monitoring solution in-house.

These companies monitor millions of products online and provide e-commerce sites and consumer brands real-time competitive intelligence into how their products are positioned and performing relative to their competitor's offerings.

In this case study, we will show you how Zyte helps one of these leading providers of price, promotion and positioning intelligence extract the product data they need to generate competitive insights for their customers. We will cover:

The company's web scraping objectives & challenges

The scope of the web scraping project

How Zyte executed this project

Why the company choose to grow with Zyte

Next phases of the project

Note: Due to the sensitive nature of the information in this case study we won't name the customer or give any identifying information for confidentiality and competitive advantage reasons.

Our customer’s objective

The company in question is one of the leading providers of product monitoring analytics to some of the world’s largest consumer brands. Their products allow brands to:

Track their products prices and compare them to their competitors multiple times per day

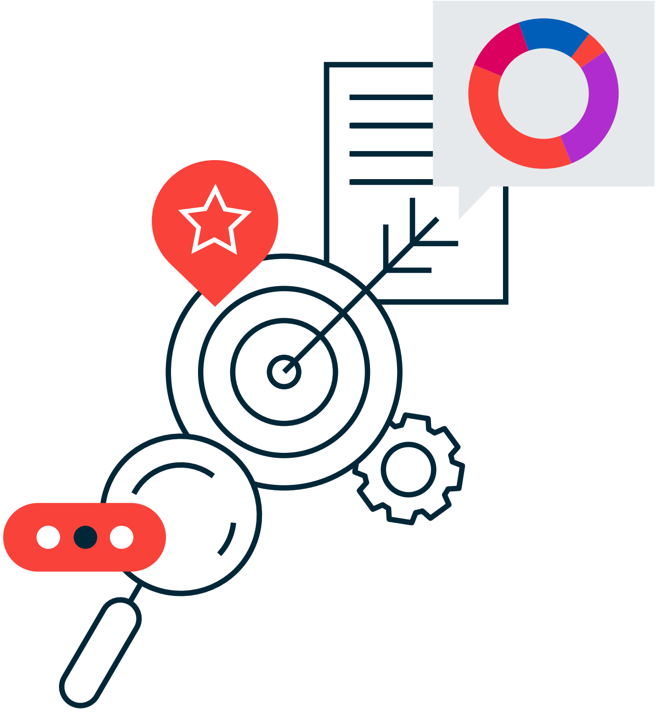
Monitor and track reviews, social media sentiment, product search ranking, online content accuracy and performance.

Track levels of promotion taking place in the market and build promotional calendars for their own products.

Ensure all resellers are correctly marketing their products (brand images, paid advertising, etc.).

To provide these services to their clients, the company needed a robust and reliable source of daily product data from hundreds of online stores.

They had already built an internal web scraping team, however, scraping at such a large scale became a huge burden on their team and was turning into a bottleneck for their growth.



At this point, they decided to outsource certain key areas of their web scraping and data extraction to enable their team to focus on analytics and insights for their clients.

In 2016 the client approached Zyte looking to engage us on an initial trial project. This project proved very successful so today Zyte manages 95% of all their data extraction activities.

Project scope

This project presented a series of very complex web scraping problems due to the scale and scope of the project. Zyte had to design, build and maintain a scalable data extraction infrastructure that satisfied all of these requirements.

Something that really excites our team of engineers, as one of our mottos is “We will always find a way, no matter how complex a problem is”. The following were the main challenges the project posed:

Large scale of data

The sheer size and scale of the data needed required a highly scalable web scraping infrastructure. The infrastructure needed to be able to easily scale from hundreds of spiders to thousands in a matter of months, enabling it to scape 1 billion products from 700 online stores every day.

Frequency requirements

As this company is providing product intelligence services to consumer brands the frequency and freshness of this data is critical. While there is value in being able to analyze historical pricing trends, etc, the real value of this data is in its freshness. This means a web scraping infrastructure able to scrape hundreds of thousands of products daily.

Quality of data

The product data that Zyte would provide is the foundation of this company’s services. A drop in data quality would immediately filter through to their customer-facing application compromising their value proposition and damage their relationship with the customer.

Reliability

As the company’s clients rely on the product data extracted from the web on a daily basis, they need a 100% reliable supply of web data. If a data feed dropped for even a day it would have huge consequences for their clients as they rely on this daily data to make product pricing and placement decisions that could cost their brands millions if made incorrectly.

Project execution

As can be seen from the project scope, this project was a very challenging and complex web scraping project.

It required Zyte to build a robust and reliable data extraction system that was capable of scaling to millions of records per day, whilst still being flexible enough to accommodate an ever-growing and changing list of products and data types to extract.

Working with our Solution Architecture team, the customer was able to communicate their business requirements and we were able to develop a data extraction architecture that extracted the data the customer required at scale. Whilst ensuring the highest possible data quality and reliability.

Web scraping stack

To accomplish this Zyte built a robust web scraping system for the client. A web scraping stack is all the individual components you need to scrape the web at scale: data extraction spiders, infrastructure configuration, proxy management, data quality assurance, etc. The web scraping stack developed for this customer includes:

- **Scrapy** - the open source Python web scraping framework developed by Zyte's founders.
- **Scrapy Cloud** - Zyte's dedicated hosting service specifically designed for deploying and managing web crawlers in the cloud.
- **Frontera** - the open source framework Zyte developed to facilitate building a crawl frontier, helping manage your crawling logic and sharing it between your web scraping spiders.
- **Crawlera** - Zyte's single endpoint proxy solution that automates proxy management and ensures our customers can reliably scrape the web a scale.
- **Splash** - the lightweight scriptable browser developed by Zyte that enables fast rendering of web pages that use JavaScript.
- **Spidermon** - the spider monitoring system developed by Zyte to enable the realtime monitoring crawl status and detection of spider failures, bans, errors, etc.
- **GATF** - the automated data quality assurance framework developed by Zyte to assist QA engineers validate and monitor the accuracy and coverage of datasets.

This web scraping system comprises over 1,200+ active spiders. Enabling the client to scrape 1 billion products from 700 stores every single day. On average this system will scrape 10,000 records per minute (RPM) during the day, and increasing to 120,000 RPM during the night when there is less load on the target websites.

A key focus for all our data extraction projects is to minimise the impact our spiders have on the target websites. As a result, when configuring spiders we always ensure that it doesn't harm the site. Once scraped, the quality of the data is automatically verified by the automated QA systems and manually checked by our QA team before being sent directly to the client's S3 bucket.

This team can call on deep organisation knowledge of the technologies, websites, anti-bot countermeasures, etc. if a problem ever arises.

As the data extraction process was so critical to our customer, having frontline access to the team was extremely important. As a result, the customer has full unrestricted access to the team working on their project.

The client's engineers can communicate directly with the Zyte team on Slack and can raise JIRA tickets with them if there is ever an engineering task that needs completion.

Dedicated team

To deliver this web scraping stack to the client Zyte created a dedicated internal team to work exclusively on this project. Whilst maintaining the option of bringing in more resources from our pool of over 100 crawl engineers if the need for engineering resources surges.

Currently, this project comprises a team of 3 engineers, a shared project manager and account manager, and 1 QA engineer. However, we often spin up separate time-bounded teams to deal with new projects or problems that arise and need immediate attention (anti-bot issue, etc.).

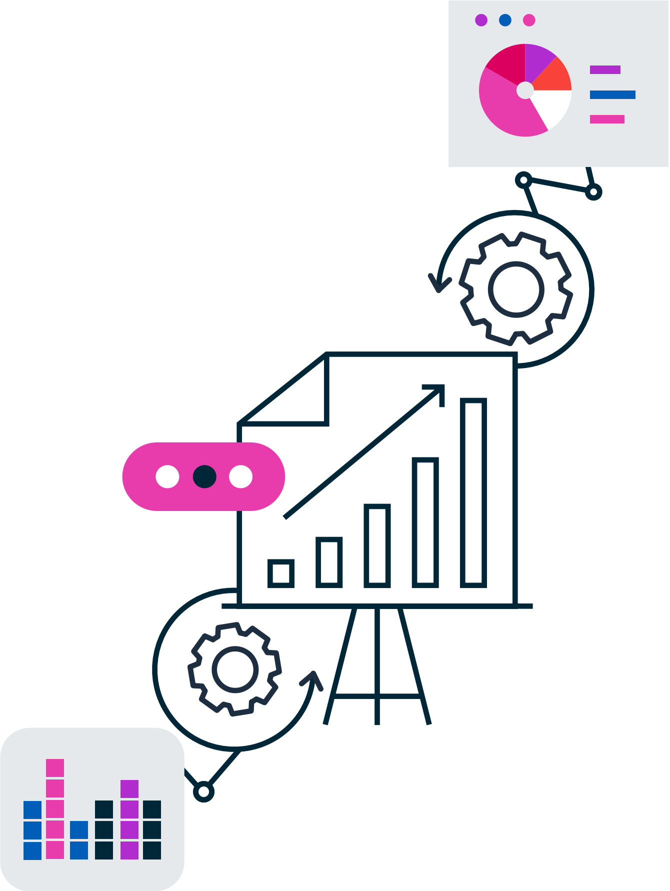
Reliability and data quality

A key feature of this project is the contractual commitment to reliability and data quality. In our Service Level Agreement (SLA) with the customer, Zyte explicitly guarantees:

- **Data quality** - Our SLAs guarantee a 99% accuracy rate as measured by the total number of correct items over the total number of items extracted. Zyte ensures this level of data quality through a combination of robust spider QA during the development stage and continuous automatic/manual QA of the data by our dedicated Quality Assurance team.

- **Response times** - Our SLA with the client also guarantees that any reported problem, JIRA ticket or spider break will be picked up within 12 hours. However, in reality, 90% of tickets get picked up within 30 minutes, fixed within 4-5 hours, passed QA within 3 hours and back in production within 24 hours.

  Ensuring 90% of issues (spider breaks, etc.) are fixed within a day. This is a huge commitment for Zyte given the number of active spiders in this project. Due to the inevitable constant changes to the structure and function of a website, spiders often break every 2-3 months on average. As a result, in a project with over 1,200 spiders, our team has to update 10-15 spiders per day.



Technology ownership

Despite Zyte building and maintaining the entirety of this web scraping system, the client retains 100% ownership of the code and any IP. At Zyte, we believe in no vendor lock-ins so the client is able to walk away from the arrangement at any time with the full code base.

This is a huge peace of mind and risk-reducing feature for the client as they know they have the ability to bring their web scraping operation in-house or transfer it to another 3rd party web scraping provider if the need ever arose in the future.

Planning & communication

As noted above, this client has a dedicated product and account manager to ensure their web scraping system runs smoothly.

The client has weekly calls with the project manager and account manager to discuss KPIs, new projects and roadblocks, ensuring their web scraping system is always perfectly aligned with their business goals. The project manager and account manager also visit the client once per quarter.

Why Zyte?

For this client, Zyte was more than just a web scraping provider. The web scraping system we built and maintain for them is the backbone of their business, without it, they wouldn't have been able to grow as quickly as they have done.

Entrusting Zyte with their entire data extraction process enabled them to focus 100% on building the best products and services for their clients. Here are some of the main reasons, the client choose to partner with Zyte:

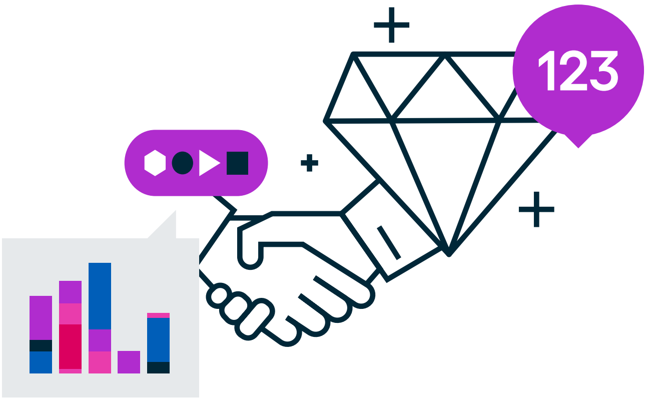
Reason #1 - Industry experts

When working with Zyte you are partnering with the industry's leading data extraction experts. Data extraction and web scraping is in our company DNA, all we have done for the past 8 years is find better ways to extract data from the web at scale. As a result, we've built a deep institutional knowledge that we are able to call upon when tackling a new project or problem.

Our team of engineers know the major websites inside-out, know how to squeeze the highest performance out of Scrapy, the open source python web scraping framework developed by Zyte, and have developed industry-leading technologies in the areas of proxy management, QA, etc. They knew that when they were working with Zyte, they were working with the best in the business.

Reason #2 - Reliable partner

Not only is Zyte the leading web scraping expert in the market, it is the most established business in the market. Whereas most web scraping consultancy firms or providers are either small operations with 2-10 staff (mainly engineers) or risky venture-backed startups, Zyte is a robust and established business (founded in 2010) that has moved past the risky "startup" phase where companies often close overnight with little warning for their customers.



Zyte has a diverse client portfolio, a strong leadership team, a rapidly growing team of engineers and support staff. This was a huge reason why the client (and many other companies) decided to outsource such a mission-critical component of their business to Zyte. They knew that if they outsourced their web scraping to Zyte they would have a partnership that would remain with them for the long term.

Reason #3 - Guaranteed data quality & uptime

By working with Zyte the client was guaranteed reliable and a high-quality data feed. Every customer contract comes with an SLA that explicitly guarantees near perfect data quality from their data feeds.

They just needed to specify the data they wanted and Zyte would do the rest. Not only that but when a spider eventually breaks due to website format changes, they knew that it would be fixed in a matter of hours. Enabling them to completely outsource the headache of monitoring and maintaining thousands of spiders.

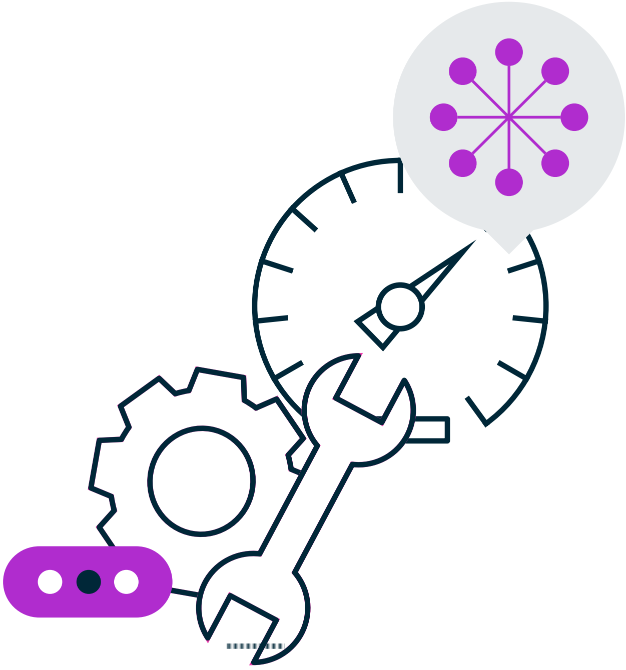
Reason #4 - Access to the best talent

The huge advantage to working with Zyte is the fact that our clients get access to the best web scraping experts in the world. The fact that Zyte is 100% remote and hires globally, means we are able to hire only the best web scraping experts.

Whereas if a company was based in Boston, or Edinburgh, for example, their talent pool would be limited by their geographical location. By working with Zyte, companies get access to the world's best web scraping engineers.

Reason #5 - Elastic pool of engineers

The beauty of working with Scrapinghub versus managing web scraping in-house is the ability to cope with surges in demand. If a big customer requires a new data feed, then Scrapinghub can quickly pull in more experienced crawl engineers from other parts of the business to work on the project. Whereas if it was an internal project, the company would often only be able to bring in inexperienced engineers or hire external help.



This flexibility was a huge advantage to this client as they were confident that no matter what new consumer brand client they signed as a customer, they would have a fully functioning data feed live within a matter of days for that client. Establishing them as the best in the business for product monitoring services.

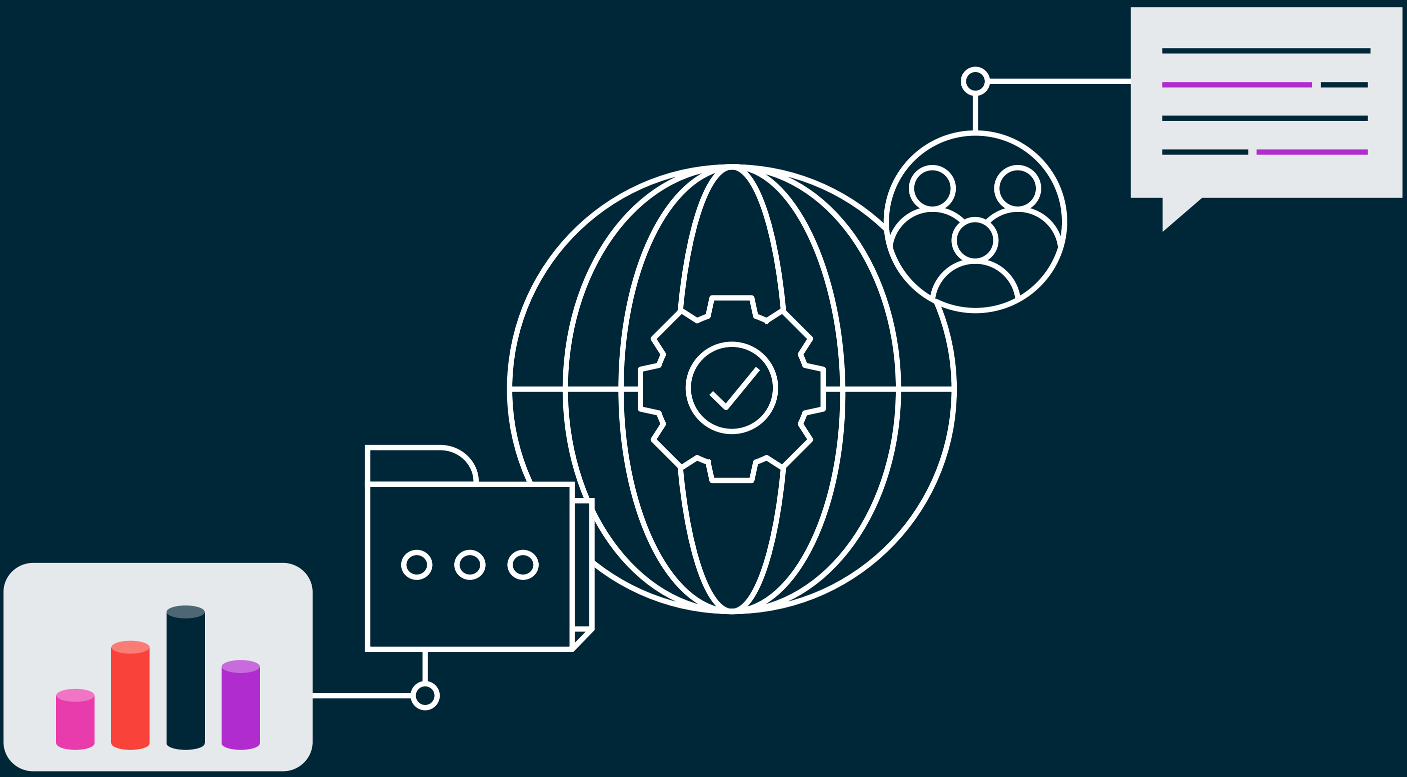
Conclusion

As we have seen, when it comes to building an enterprise web scraping system it can be quite an endeavor. You will face numerous challenges which can be a huge burden on any engineering team.

In this case, the company decided to outsource the development and maintenance of their web scraping system as it would enable them to focus on building better products and growing their business. This strategy proved very successful for them as they have gone on to become the leading provider of product monitoring analytics for consumer brands.

For those of you who are interested in scraping the web at scale but are wrestling with the decision of whether or not you should build up a dedicated web scraping team in-house or outsource it to a dedicated web scraping firm then be sure to check out our other guide, Enterprise Web Scraping: The Build In-House vs Outsource Decision.

If you would like to learn more about how you can use web scraped data in your business then feel free to contact our Solution Architecture team, who can help you design a web scraping system to get the data you need.



At Zyte we turn websites into data with industry leading technology and services.

Our solutions include:

- **Data Extraction Service**
  Let our web scraping experts build and manage the bespoke data extraction solution for your business needs.

- **Automatic Extraction powered by AI**
  Instantly access accurate web data through our user-friendly interface or various Extraction APIs and save time getting the data you need.

- **Smart Proxy Manager (formerly Crawlera)**
  Forget about proxy lists. We manage hundreds of thousands of proxies, so you don't have to.

- **Data extraction platform**
  Access developer tools, data extraction APIs and documentation, built and maintained by our world-leading team of over 100 extraction experts.

zyte

It's yours. The web data you need.

Access clean, valuable data with web scraping services that drive your business forward.

Talk to us

www.zyte.com

Cuil Greine House, Ballincollig Commercial
Park Link Road, Ballincollig / Co. Cork, Ireland