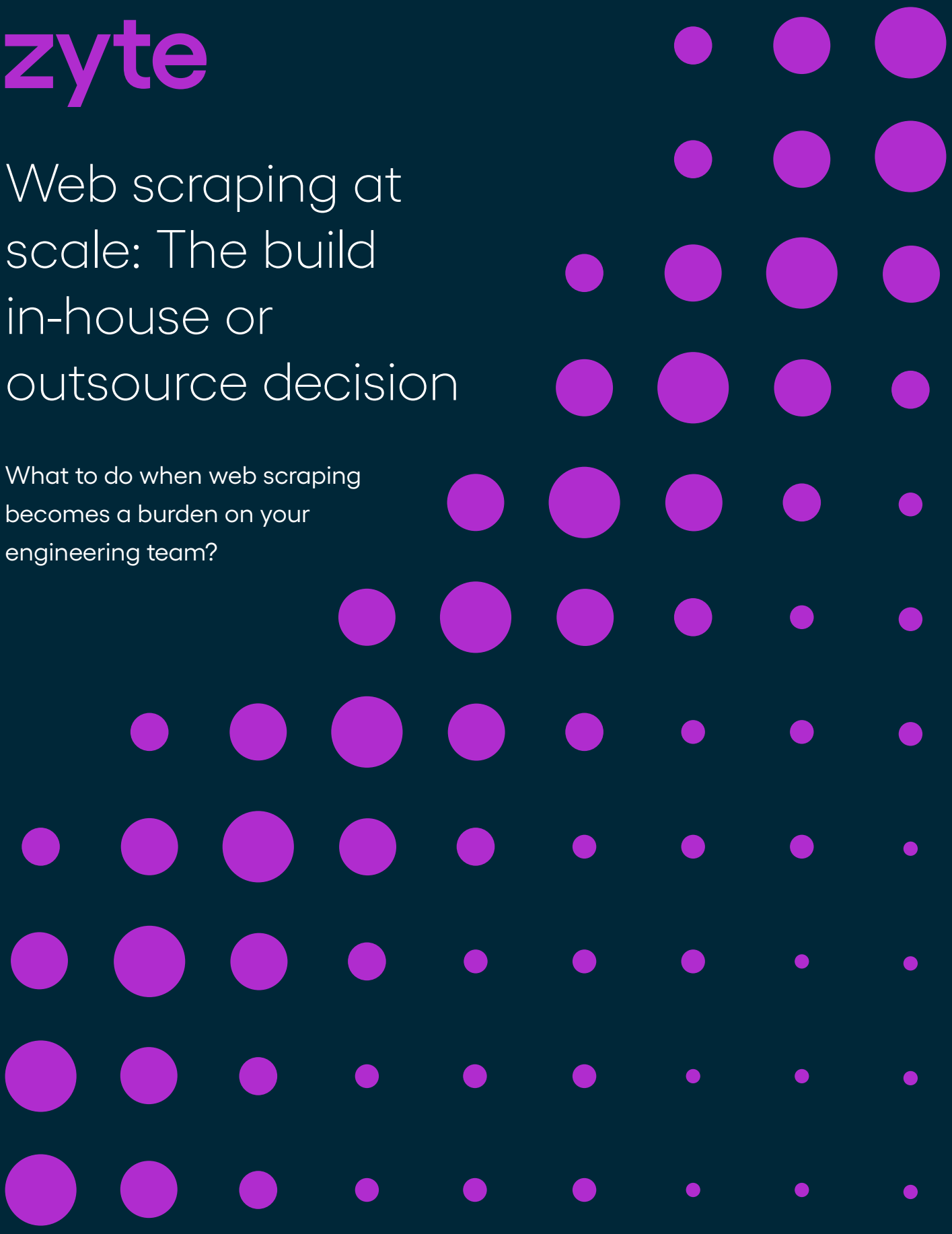
zyte

Web scraping at scale: The build in-house or outsource decision

What to do when web scraping becomes a burden on your engineering team?



Introduction

As you probably know, web scraping can quickly become a burden on your company's engineering team. What started out as a simple data extraction project became a time consuming task that sucks engineering resources away from your core projects.

The question most companies face at this point is whether or not they should double down and hire dedicated web scraping engineers or should they outsource their web scraping to a firm that specialises in web scraping.

This is the question this white paper is intended to answer.

Over the last 10 years, countless companies have come to Zyte with this question. Most had already developed some form of web scraping solution in-house (or externally) but are now struggling to maintain or scale it to meet their business needs.

It might have been that:

01. They developed a successful proof of concept, but they ran into roadblocks when trying to scale up the scope of the project.

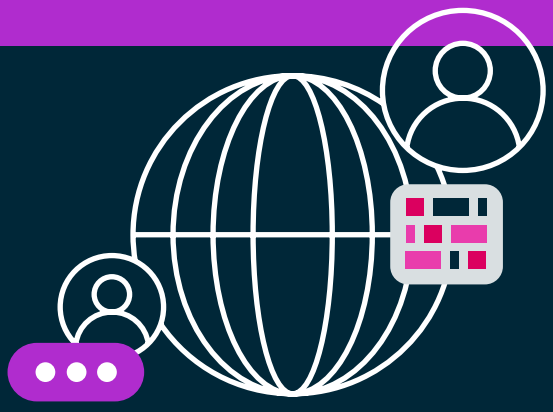
02. Their web scraping started as a side project, but their engineers now can't cope with the maintenance workload.

03. Their web scraping infrastructure is a legacy project, but now nobody on the team knows or has the time to maintain it.

04. Their data science team started extracting web data for business intelligence, but now they are spending more time maintaining their spiders than analysing the resulting data.

05. Their need for web scraping is very elastic, their customers regularly come to them looking for new data feeds but they don't have the resources to satisfy this demand immediately.

In this white paper we will discuss the economics and challenges of web scraping, when you should build your web scraping team in-house and when it is best to outsource your web scraping activities.



What you need to know when planning your web scraping roadmap

When planning your web scraping roadmap, there are two facts you need to keep in mind:

- Web scraping frameworks like Scrapy can make web scraping seem deceptively easy initially.
- Unlike most other software development projects, web scraping projects are a continued big commitment that usually require more time to maintain spiders than the time you spent building them initially.

From our experience of working on thousands of web scraping projects, the complexities and resource requirements of a web scraping project typically follow a S-curve.

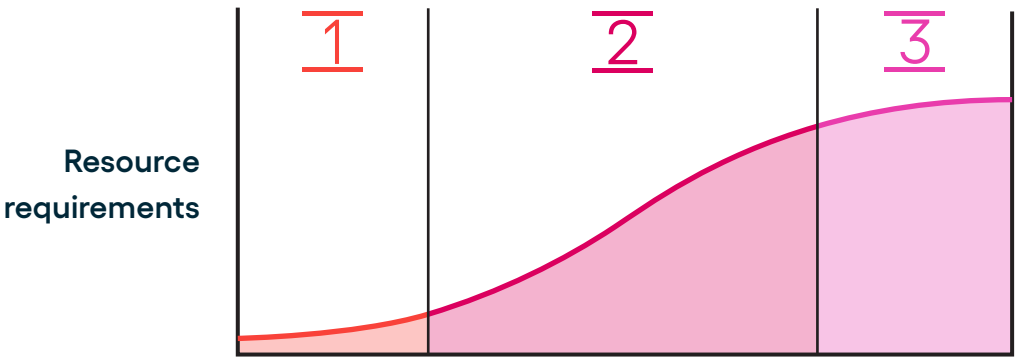
1

If you have a team of experienced software developers, extracting data from a small number of websites can initially be quite straightforward.

2

However, as the scope of the project and the need for ongoing maintenance grows, at a certain point the resources you will need to commit to your web scraping projects can increase significantly. This precise inflection point will depend on the individual characteristics of your project:

- The number of websites being scraped.
- Which websites are being targeted (certain websites are more complex and have stronger anti-bot defenses).



Scale & complexity of web scraping

- The complexity of the crawling logic (i.e. does it require logins, specific regions, nested data, etc).
- The frequency of scraping (i.e. hourly, daily, weekly, monthly, ad-hoc).

Regular changes to the target websites structure, anti-bot countermeasures and the ongoing need for data quality assurance can see spider maintenance consuming 10X the time it takes to develop your spiders in the first place.

What's more, this resource load is only set to increase.

Over the last years, as more and more websites are changing their website formats more frequently (due to special promotions, A/B split testing, etc.) and are using ever more sophisticated anti-bot counter measures, the workload of maintaining spiders has increased considerably and we expect it to continue to increase over the coming years.

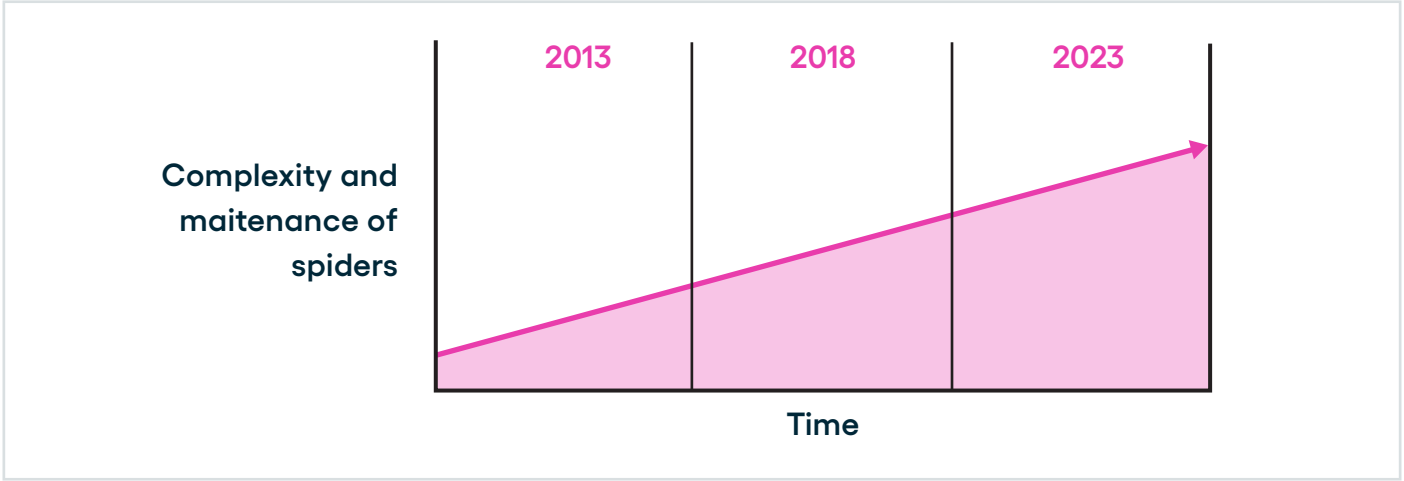
3

However, going back to Figure 1, once you've hit a certain scale threshold the marginal resources requirements often begin to drop. As the scale of the project warrants your engineering team to develop custom tools to tackle recurring problems and automate proxy management and quality assurance.

Zyte has reached this point where scraping more websites leads to only a marginal increase in resource requirements, however, most companies that come to us wrestling with the build in-house vs outsource decision are at the first inflection point.

They have already built some form of web scraping infrastructure in-house but they recognise that the growing workload of maintaining their spiders has the potential to impact their ability to work on their core business processes.

Next, we will discuss when you should build up a web crawling team in-house and when it makes more sense to outsource this work to a dedicated web scraping solution provider.



Build in-house vs outsource questions

As we've seen, building your web scraping infrastructure in house can quickly turn into a big resource commitment.

Sometimes it makes sense for companies to take this step, however, in other cases companies would be better off outsourcing the development and maintenance of their web scraping infrastructure.

Thereby allowing them to focus on their core business.

When deciding whether or not you want to build out an internal web scraping team or partner with a web scraping firm you need to be asking yourself these questions:

01. How integral is web scraping to your business?

Is extracting web data at the core of your business or is web scraping just a small component of the value you offer to your customers?

02. Do you have the technical manpower, or have a plan to hire it?

Do you plan on building out a full web crawling team or will it always be a side project?

03. How much web scraping experience do you have in-house?

Is your teams experience confined to building small scale simple web crawlers? Or does your team have experience building and maintaining large scale complex spiders?

04. How would web scraping downtime affect your business?

Is extracting web data at the core of your business or is web scraping just a small component of the value you offer to your customers?

05. Is your need for web scraping elastic?

Is your need for web scraping dependent on the elastic demand from customers?

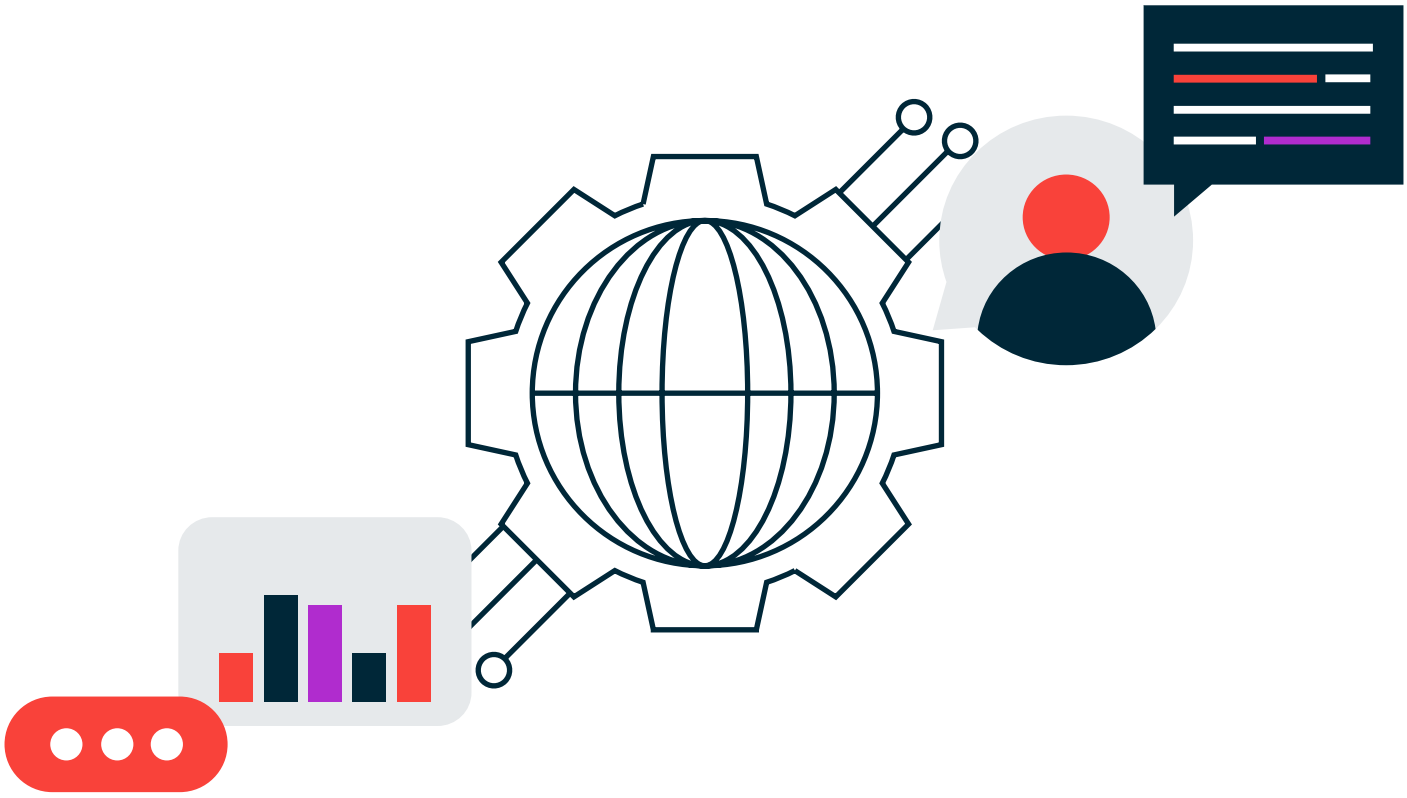
06. What is the future scope and scale of your web scraping?

Will your web scraping always be confined to scraping a small number of websites or do you plan to increase the scale of your web scraping overtime?

07. What is the future scope and scale of your web scraping?

Do you have budget and technical resources/expertise to build your own custom automated proxy management, data quality assurance, complex spider logic if you plan to scrape at scale? Would those resources generate a higher ROI if focused on your core business?

Your answers to these questions will have a **major impact on what web scraping option best suits the needs and priorities of your company**.



When should you build an in-house web scraping team?

Building a web scraping team in-house does have its advantages. Initially, it can be cheaper and ensures that the technical web scraping expertise you develop remains in-house.

However, unless you are fully committed to building out a dedicated web scraping team it can often lead to a sub-par web scraping infrastructure that can become a burden on your engineering team and doesn't fully meet your business requirements.

In general, building out a internal web scraping team makes sense in only two situations:

When web scraping is core to who you are

If extracting large amounts of web data is at the core of who you are as a business, the demand is constant and you have (or plan to build out) a dedicated web scraping team, then it can sometimes make more sense for companies to build out their web scraping team in house.

Examples of companies like this would be search engine marketing tools or raw web data providers. However, even in scenarios like this a lot companies choose to outsource the development and maintenance of their web scraping infrastructure.

Outsourcing their web scraping allows them to have a more sophisticated web scraping operation, whilst enabling them to focus on analyzing the resulting web data and refining their core products.

We have a number of clients where the large-scale extraction of web data is the lifeblood of their business, but who've chosen to outsource their web scraping to Zyte because they feel that outsourcing their web scraping infrastructure allows them to focus all their energy on analyzing that data and building the best possible solutions for their customers.

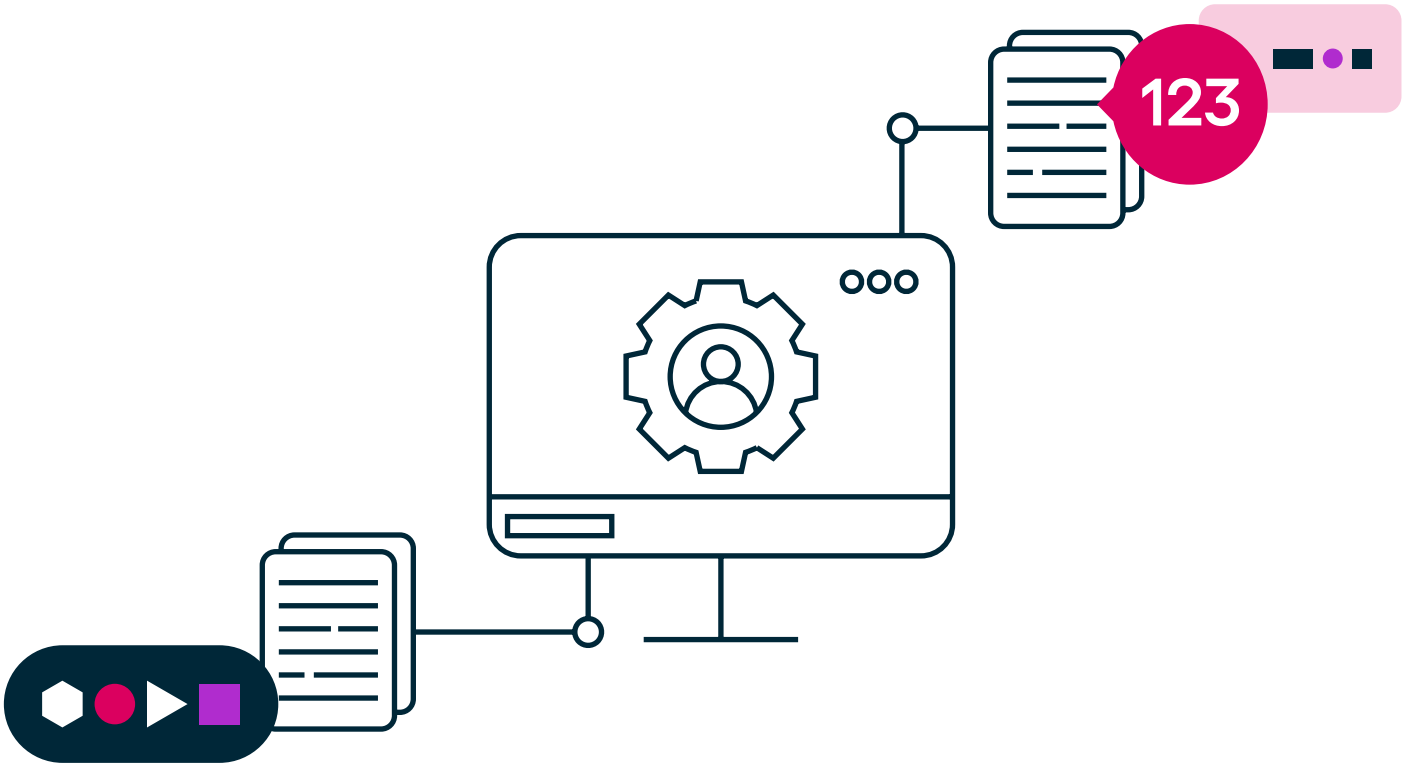
When web scraping is a very early stage side project & you don't have a budget to outsource

If you don't have any budget for web scraping, only need to extract very small amounts of non-mission critical web data and your technical team has enough bandwidth to develop and maintain this system then allocating web scraping as a side project in house makes sense.

However, if you have a small budget you are better off subscribing to a data feed for the target websites.

For a couple hundred dollars per month you can have clean structured data delivered to you each month, with no need to build or maintain web scrapers in house.

Zyte offers company data feeds like this for as little as $150 per site per month.



When should you outsource web scraping?

In a lot of cases outsourcing your web scraping will initially cost you more than hiring a team inhouse, however, over the longterm outsourcing your web scraping infrastructure will allow your company to focus entirely on the activities that grow your business and outsource the headache that is maintaining web scraping spiders.

Outsourcing your web scraping to a firm with deep web scraping expertise also has the added benefit of allowing you to build much more powerful web scraping applications that most companies wouldn't attempt if they were building a team in-house.

For example, quite regularly companies contact Zyte looking to scrape real estate listings from a few specific websites.

However, most companies are blown away with the fact that Zyte can find and scrape every real estate listing for a given country using Broad Crawls that search a entire top-level domain (ex. ".co.uk", ".fr", etc.) or the entire web for real estate listing that match their defined criteria. The extensive data these more complex spiders can give companies often open huge new opportunities for them and completely change the way their business operates.

The following are some of the situations when you should seriously consider outsourcing your web scraping infrastructure to a experienced partner who can build you a more sophisticated web scraping application whilst guaranteeing data quality and uptime.

If you are scraping web data at scale

Scraping the web at scale is a complex and resource intensive process, posing difficult challenges for any web scraping team. If your business needs to extract web data from a large number of websites (especially large & complex websites), but your team has limited web scraping experience or web scraping is only a small input to your core business it is best to outsource your web scraping infrastructure to a dedicated web scraping team.

If you need guaranteed data quality & uptime

With websites increasingly changing their formats more frequently (A/B split testing) and the growing usage/sophistication of anti-bot countermeasures, ensuring web scraping uptime and data quality can be quite labor intensive. Often requiring a full team of experienced crawl engineers if you are scraping at scale.

If any disruption to your web scraping uptime or data quality would have a major impact on your business then you should consider outsourcing your development and maintenance of your web scraping infrastructure to an expert web scraping team. Especially if you are scraping at scale or complex websites.

If web scraping isn't a core part of your business

As we know, web scraping can be a very resource intensive activity that can take time away from other parts of your business. If extracting web data is only a small input to your overall business processes then you should consider outsourcing it.

Outsourcing the development and maintenance of your web scraping infrastructure will allow you to focus your engineering resources on building your core infrastructure and products, without the headache of maintaining a web scraping infrastructure.

If your need for web scraping is elastic

Quite often a company's need for web scraping is driven by the elastic demand for new products and data feeds from end customers making it uneconomical to build out a dedicated in-house web scraping team.

In these situations, it is best to outsource your web scraping infrastructure to a web scraping firm that can rapidly develop and deploy spiders to meet your web scraping needs.

If your technical team doesn't have the experience or resources

If your engineering team is already tight on resources or doesn't have a deep level of experience, you should consider outsourcing your web scraping activities.

Especially if having access to high quality data is mission critical for your business or you need to scrape large quantities of data from multiple websites.

Conclusion

When it comes to the challenges you face when scraping the web for valuable data, specifically the burden it can impose on your engineering team, there isn't a one-size fits all approach.

In some cases an in-house solution might make the most sense depending on your current resources and needs. However, the question you need to be asking yourself is how will these resources and business needs evolve over time?

Your goal isn't to have a solution that meets your needs just today, but one that can evolve and meet your needs in the future as your business requirements and goals change.

For a lot of companies this means outsourcing their web scraping to a dedicated web scraping firm is the best option.

It allows them to have a world class web scraping infrastructure without the hassle of maintaining it, and if their business needs change their web scraping partner can quickly upgrade the infrastructure to meets those new business requirements.

Enterprise solutions

Complete web scraping services for any size business, from startups to Fortune 100's

Web data, hassle-free, for real business needs

Whether outsourcing or aiding your in-house team, Zyte can assist your business with top notch expertise in web scraping.

Lead generation, competitor & sales intelligence

Monitoring of ratings and reviews, sentiment analysis & social network intelligence

Product aggregation & price monitoring for retail, e-commerce & manufacturers

Dark web, law enforcement & compliance

Alternative data for finance, equity and market research

Staffing, talent sourcing & job market research

Let's partner

Team up with the best web scraping engineers while you stay focused on your business goals.

Data on Demand

Any size scraping project. Data refreshed regularly, reliably and in the form you want.

Data Science

Team up with the best web scraping engineers while you stay focused on your business goals.

Training

Learn from the recognised experts in data crawling and scraping to grow your own in-house team.

Get a free consultation

From the world leading experts in web scraping

sales@zyte.com

+1-347-559-1901

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 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.
- **Smart Browser**
  Manage bans and get your data from sites using JavaScript and browser rendering with our single API.

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