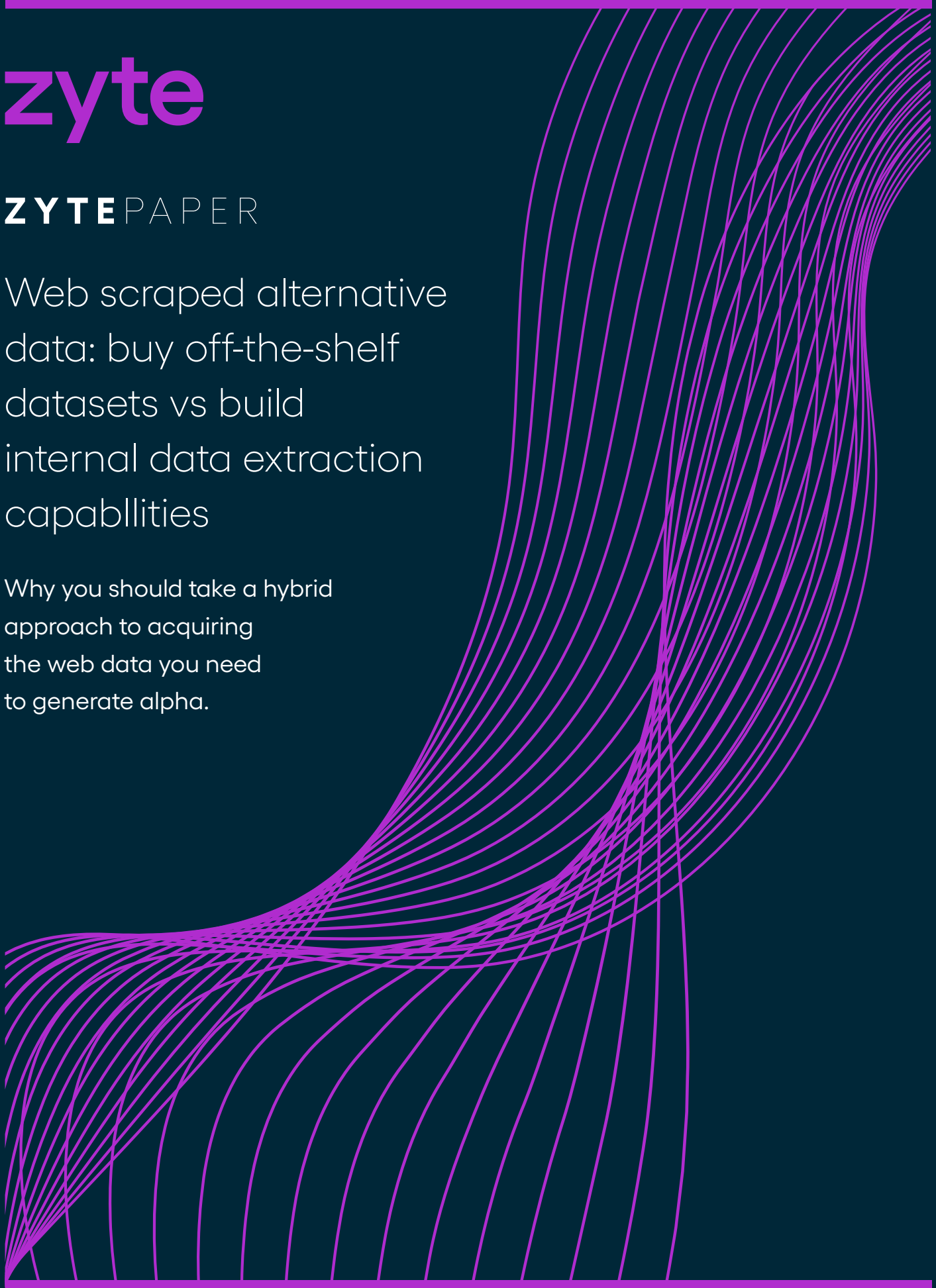
zyte

ZYTEPAPER

Web scraped alternative data: buy off-the-shelf datasets vs build internal data extraction capabllities

Why you should take a hybrid approach to acquiring the web data you need to generate alpha.



Introduction

In recent years, there has been an explosion in the use of alternative data sources in investment decision making amongst hedge funds, investment banks and private equity firms all with the aim of generating alpha.

Although investors are using countless sources of alternative data, satellite, weather, employment, trade data, etc. the leading type of alternative data is web data.

Unlike a lot of other types of alternative data, data professionals in hedge funds can choose either to purchase web data from alternative data vendors in the form of off-the-shelf datasets or develop their own internal data extraction capabilities to extract the data themselves.

The path they ultimately choose can have a large impact on the compliance risk the data poses, its ability to compliment other forms of traditonal and alternative data sources in the decision making process, and most importantly it's ability to generate alpha.

In this whitepaper we're going to look at four of the main factors firms should take into account when evaluating which option to choose:

Exclusivity

Compliance risk

Historical data

Customisation

Firms don't need to exclusively choose one option or the other, and often don't. They can take a hybrid approach and build internal web scraping capabilities themselves or through partnering with a dedicated web data extraction provider like Zyte and supplement these datafeed with off-the-shelf datasets. Which is what we recommend.

This guide is designed to help you evaluate your web data acquisition options, be it offthe-shelf datasets or building your own data extraction capabilities.

Exclusivity

The issue of exclusivity is of huge significance for the users of alternative data and often falls into a catch twenty-two. Alpha generating data is by definition material. To be able to generate alpha the dataset must give investors an informational edge over the market that better informs their decision making.

However, if everyone has access to the same data the ability to generate alpha from that data will be arbitraged away very quickly. Effectively commoditizing the data.

In an ideal world, asset managers would like to have access to as much exclusive material data as possible when making their investment decisions. However, the significant downside (and risk) to having exclusive access to material non-public data is insider trading prosecutions. There is the possibility that prosecutors could deem paying for exclusive access to material data would give a firm an unfair market advantage over their competition which could potentially be deemed insider trading.

Off-the-shelf vs data extraction capabilities

As web data is largely publicly available on the web and free for everyone to access and use, it is difficult for prosecutors to argue that web data could be considered material non-public information (MNPI).

Question marks do arise if the data was extracted from behind a login, so these project should be investigated on a case by case basis.

As a result, even if an investor was to pay for exclusive access to an off-the-shelf dataset the associated insider trading risks would be much lower than if they were to pay for exclusive access to a non-public data source such as credit card transactions.

That being said, when it comes to exclusivity developing your own data extraction capabilities do offer asset managers a chance to develop an informational edge over the market without falling foul to exclusivity issues.

If an off-the-shelf web scraped dataset was proven to be valuable then it would be quickly commoditized as a result of others purchasing the same dataset or extracting the data themselves. However, if a hedge funds were to discover a material web data source and extracted the data themselves they could maintain their informational edge for a much longer period of time as no one would know they were extracting the data in the first place. Giving them proprietary access to a highly valuable data source.

Using this approach, the user of the data wouldn't fall victim to insider trading violations as the data is freely available on the internet. Just they were the only one to develop the data extraction capabilities to exploit this informational opportunity.

Compliance risks

Regulatory compliance is of huge concern for asset managers when considering the incorporation of a new data sources into their investment decision making processes, including web scraped alternative data. Any lapse in compliance standards could pose significant headline and financial risk for the firm.

In this article, we outline the compliance best practices for web scraped alternative data: Navigating Compliance When Extracting Alternative Financial Data From the Web.

Off-the-shelf vs data extraction capabilities

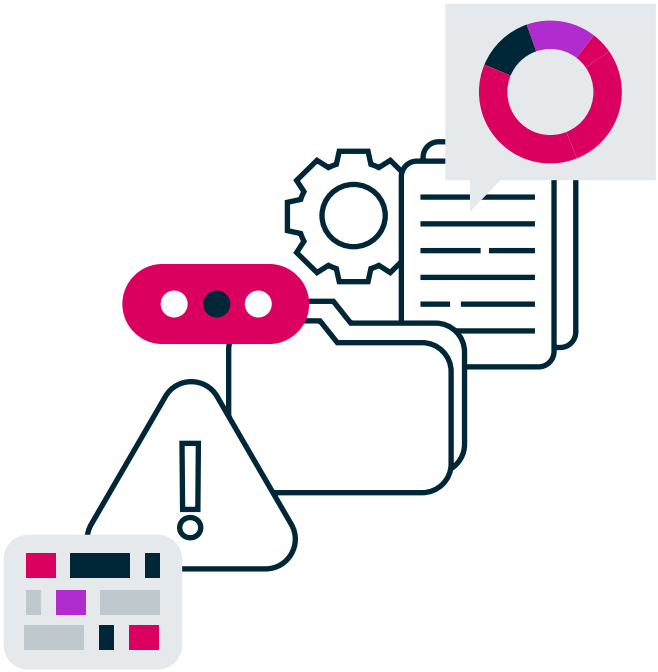
When deciding whether to go with an off-theshelf web data solution or the development of your own data extraction capabilities the compliance decision really comes down to the level of control and oversight your firm wants to have over the mitigation of compliance risks.

Web data can pose significantcompliance risks if the underlying risk factors aren't managed and mitigated correctly. Everything from what data is scraped, how it is extracted, how friendly the crawlers are, and how it is processed can greatly increase the compliance risks of the data.

When you are in full control of the data extraction process, you can take steps to ensure these issues never arise as they all fall under general web scraping best practices.

However, if you are considering using an off-the-shelf dataset from a 3rd party provider it can often be much harder to ascertain if the data was extracted safely.

In a lot of cases, the data provide operates more so as a data marketplace. Collecting and organising a wide variety of alternative data types. Oftentimes, outsourcing part or all of their web data extraction to another 3rd party. As a result, the level of oversight of historical web scraped dataset could be quite patchy.



Although some data providers may fully or partially indemnify you against these risks, the only true way to fully mitigate them is by directly controlling the data extraction process, either by moving the data extraction process in-house or partnering closely with a data extraction provider who has experience extracting alternative data for financial use cases.

Historical data

When it comes to using data in multi-million dollar (or billion dollar) investment decisions, the ability to validate your investment hypothesis via benchmarking and backtesting is crucial.

An asset managers investment thesis must be rigorously stress tested to evaluate the soundness of their underlying assumptions, the predicted risk and return from the investment, and then benchmarked versus other competing investment theses being for the same pool of investment money.

The most effective way of evaluating how this investment thesis would have faired in past situations is by stress testing it with historical data.

What this means for web scraped data is that web data doesn't start to become truly valuable until you have a complete historical dataset.

The key here is the word "complete".

Not only is it vital to have historical data, it is crucially important that the data is of high quality with no data gaps or corruptions.

Off-the-shelf vs data extraction capabilities

This is the area where off-the-shelf data sources do have an edge over the development of your own data extraction capabilities. At least in the immediate term.

Alternative data vendors know the importance of historical data so they offer historical datasets as off-the-shelf solutions. The completeness and value of these datasets can be validated with some internal modelling and benchmarking, making them an ideal solution if you have an immediate data requirement.

Often the development of your own data extraction capabilities is more of a medium to longterm initiative as it takes time to produce a usable historical dataset from scratch. This approach is akin to planting seeds that have the potential to turn into highly valuable proprietary resources for your firm.

A data extraction team can act as an internal skunkworks, sowing the seeds for new alpha generating alternative data sources that could give your firm exclusive access to material data in the future. Enabling them to develop investment theses that give them a unique edge over the market.

Customisation

The ability to tailor data sources to the specific needs of your asset managers is of great value to any firm.

Different hedge funds and asset managers often have their own individual data requirements as a result of focusing on different investment opportunities and taking different approaches to evaluating those opportunities.

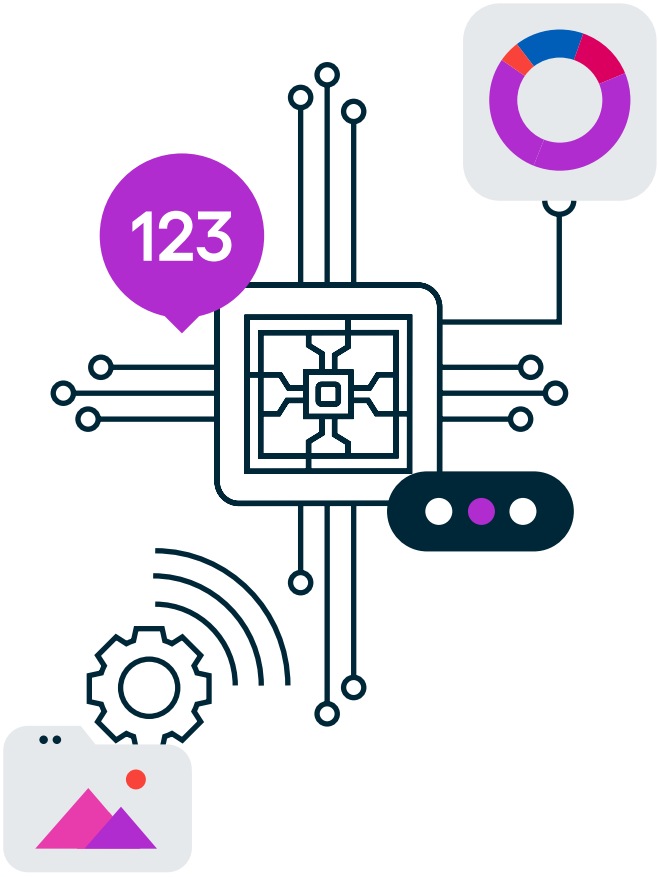
As a result, the ability to develop custom data feeds 100% optimised for your internal data needs and data infrastructure can result in huge efficiencies and give your firm an informational and operational edge over the market.

Off-the-shelf vs data extraction capabilities

Off-the-shelf datasets provided by many alternative data providers are by their very nature non-customisable. Oftentimes, alternative data providers have made an investment thesis decision years ago to develop the capability to gather very specific data types from the web.

Once the dataset has matured to a point where it would be of value to asset managers it is impossible to change the nature of the data contained in the datasets.

Making off-the-shelf datasets somewhat restrictive and less than fully optimised for your firm's individual data requirements.



On the other hand, through the development of your own data extraction capabilities, your firm can ensure your data feeds will be 100% optimised for your firm's individual data requirements.

The precise data, it's granularity and scope can be chosen by your asset managers and data analysis teams to ensure it will provide the highest value to your investment decision making process.

Conclusion

As we have seen, both off-the-shelf datasets and the development of your own data extraction capabilities have their own pros and cons making each of them better optimised to fulfill different types of data requirements for your investment decision making processes:

Off-the-shelf datasets

are better suited to keeping pace with the market or getting a short-term edge over your competitors. By using off-the-shelf datasets you can quickly fill an immediate web data requirement either as a result of your competitors developing an informational edge or the identification of a new investment thesis that requires web data.

Developing internal data extraction capabilities

either by building your own team or partnering with a data extraction service provider can give you the ability to build exclusive data feeds that yield an informational advantage over the longterm. Exponentially expanding the number and completeness of the investment theses your team can develop.

Whether you decide to build your data extraction capabilities in-house or partner with a data extraction service provider, Zyte provides the best of both worlds. With a talented team of scraping engineers and the best-in-class tools for scraping professionals we can help support your data extraction needs no matter which route you choose.

At Zyte we have extensive experience developing custom regulatory compliant data extraction solutions for the financial section.

Supporting financial institutions fulfil their web data requirements in a number of ways:

Data on demand & enterprise solutions

where Zyte develops and maintains custom and compliant data extraction infrastructure for financial institutions. Removing the hassle of having to build internal data extraction capabilities in-house.

Market leading web scraping tools

numerous internal data extraction teams rely on Zyte's comprehensive suite of web scraping tools to operate their crawlers efficiently at scale. *Scrapy* to build their web crawlers, *Scrapy Cloud* to run and schedule their crawlers, *Crawlera* to manage their proxy pools, and *Splash* to render JS heavy pages.

Web scraping team

where clients are assigned a dedicated mid and senior level Scrapy engineering resources to work with their internal teams to develop, expand and maintain their data extraction capabilities.

Whilst maintaining full oversight of the data extraction processes, and full ownership of the spiders and the extracted data.

If you have a need to start or scale web data acquisition for your alternative data needs then our Solution Architecture team are available for a free consultation, where we will evaluate and architect a data extraction solution to meet your data and compliance requirements.

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