import.io

The Buyer's Guide to Web Data Integration

Learn more at import.io

Contents

The buyer's guide to web data integration

Step 1: Establish your needs

- What is your business objective and how will the business benefit?
- What data are you looking for? Do you know where it can be found?
- Is some of the data potentially hidden, for example data within the html source or available in an underlying API call but not visible?
- How will the data be used? For example, will it be ingested in an analytics platform? Will it be integrated into a business process or application?
- How often do you need the data? Hourly, daily, weekly, on a custom schedule?

Step 2: Considerations

Commercial

- What is your budget?
- How many internal resources and capabilities are available to support this effort?
- How much data needs to be collected, and how often?
- What level of data quality do you need?

Technical

- Run yourself or use a managed data integration service?
- Build internally or buy?
- Deploy on premises or SaaS?
- Visual interface or developer tool?
- What are the security, privacy, and data storage considerations?
- What integration points will be used? API, JSON, CSV, S3, or direct database load?

Step 3: Document requirements to help vendor selection process

- Service type: managed or self-service?
- SaaS or on premises?
- Does vendor offer service level agreements that guarantee data quality?
- Does vendor offer to indemnify against legal risk?
- How much experience does the vendor have?
- Can the vendor provide the services and support that you need?
- Can they get the hidden data?
- How much experience do they have?
- Can they provide the support that you need?

Step 4: Proof of concept

- Agree to pricing and commercial terms before embarking on PoC.
- Give the vendor samples of the sites that have the data you need.
- Evaluate the data to ensure it meets quality needs.
- Examine additional features and capabilities, such as built-in analytics.
- Given them a deadline to meet.

Step 5: Purchase & implementation

- Ensure a statement of work is included as part of the purchase documents.
- Ensure clear goals and milestones are agreed upon and documented.
- Establish a documented change management process.

Introduction

We have entered an era of data-driven competition. Through the combination of new and expanding sources of data and innovations in advanced analytics, artificial intelligence (AI), and machine learning, the very nature of business competition is being rewritten. Today, entire industries are being disrupted by companies that are leveraging data successfully.

To win in today's data-driven markets, organizations can't just rely on their own internal data, which can tell them where they are. Organizations must also leverage external, alternative datasets that can tell them where they should be going.

The web is the world's biggest repository of data and is a source of tremendous potential for data-driven insights. The problem is that the data on the web is not structured or organized to be read by machines; it is formatted to be read by humans. To fuel insights and innovation, this data needs to be made machine readable so that it can be consumed at scale.

In the past, teams have tried to use legacy web scraping tools to consume the web data they needed. However, invariably, these basic tools left organizations with data that's incomplete, inaccurate, unreliable, and out of date—while introducing high costs and business risk.

Web Data Integration is a new approach to acquiring and managing web data that focuses on data quality and control. Web Data Integration treats the entire web data lifecycle as a single, integrated process composed of the following steps:

- **Initial identification of data sources and requirements.**
- **Web data extraction.**
- **Data preparation and cleansing.**
- **Analysis and visualization.**
- **Data integration and consumption by downstream applications and business processes.**

To successfully capitalize on web data and gain a competitive advantage, organizations need a Web Data Integration solution that addresses all these requirements. With this solution, organizations can leverage web data in a fast, scalable, and cost-effective way that minimizes business risk.

This guide is designed to help you plan your Web Data Integration project.

In the following pages, we walk you through the process of finding a Web Data Integration software provider that is optimally aligned with your business goals.

Ensure you establish and agree upon needs

As you begin in a Web Data Integration initiative, it is important to consider the following questions:

Why is web data necessary to your organization?

The first step in any data analysis effort is to determine the questions that need to be answered. For Web Data Integration efforts, the same principle applies. The web provides an unprecedented wealth of information but finding the right solution for the job requires a clear understanding of the business need and exactly how web data will help.

What kind of web data do you need?

How much web data will you need? Where can you find it and how quickly do you need it? Is it a one-time collection or will there be ongoing web data needs? Note that these requirements are based on your technical and business needs and are essential in your evaluation of the solutions different vendors have to offer.

How can you access the data?

Most think of web data solely as the content that's visible when you navigate to a webpage. However, there may also be data that's hidden on the page or that requires logins or other interactions before being exposed.

How will you use the data?

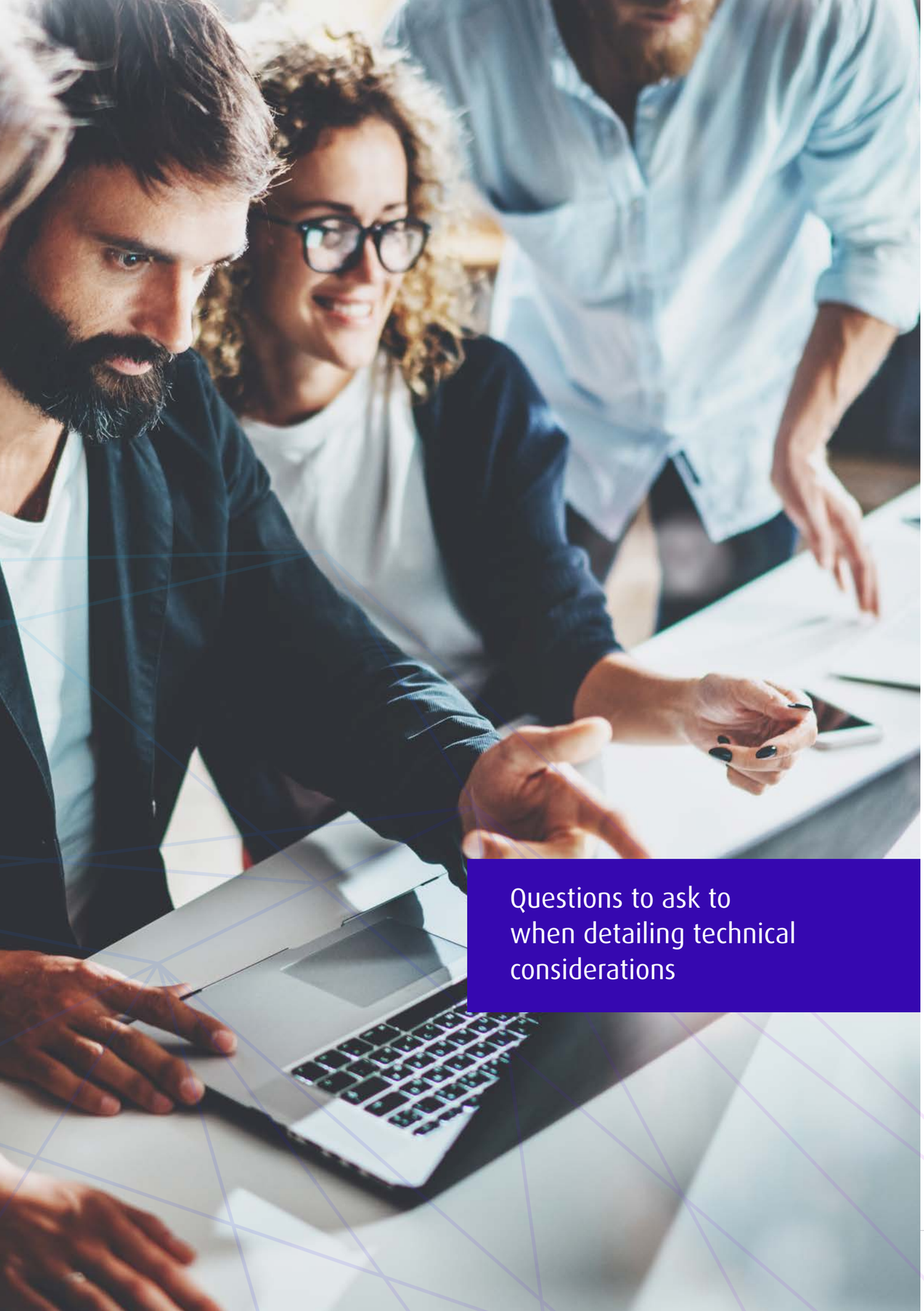
Will the web data be used to integrate directly into an application or business process, or will it be used to support an analytical investigation? If you are looking at web data to guide decision-making, it is important that you ensure that it is high quality.

Do you need a commercial or in-house solution?

Are you looking to move more responsibilities to non-developers? Are you hoping to get faster turnarounds on web data requests? Do you know what the infrastructure requirements are, and, if so, are those resources currently available? These questions will help you start planning for the people, processes, and technologies that you'll need.

As a result of answering the questions above, you'll be well positioned to determine whether you need a commercial Web Data Integration solution, and, if so, the specific requirements you should look to have it addressed.

import.io



Questions to ask to when detailing technical considerations

Developing technical requirements

Once you've identified the web data needed to answer your business questions, it's time to dive a bit into the details and more technical considerations.

How much self-service is your organization prepared to take on?

Do you have the staffing resources and expertise needed to build, maintain, and use commercial software? For example, will end-users be able to work with raw extracted data? Will the software offer the capabilities needed to enrich data and make it consumable by non-technical users? Do you value a solution that allows non-technical users, including business experts, to easily participate in tuning the system to extract the data required? Does the software allow you to create, automate, and schedule specific data transformation processes, so that they can automatically be applied to every extraction?

In addition, Web Data Integration software providers should give you the flexibility to choose between self-service and managed-service approaches. Make sure the software doesn't create gaps between the extraction of raw data and the delivery of the high-quality data your business needs.

Should we build it ourselves for complete control?

A big advantage to building your own extraction tools is the level of flexibility and customizability you have. You can define exactly what data you want to access and at what frequency. This allows you to tailor your tool to the exact scope of your initiative.

If you're trying to answer a relatively narrow question that requires a small dataset, or monitor websites on an ad-hoc basis, developing an in-house data extractor can be a simple, viable approach. Before embarking on a project to build your own web data extractor, here are some questions to consider:

- How long will it take to build these extractors and get the data you need?
- Does your team have experience in writing code for web data extraction?
- What will it cost to have the infrastructure—including servers, networking, and storage—needed to support your ongoing data extractions?
- What if you need more web data? Will the tool that you build, and its underlying infrastructure effectively scale?
- What happens when the websites you need to extract data from change? How quickly can your team rewrite the code?
- Are you OK with gaps in datasets when an extraction process fails?
- What is the opportunity cost of having your engineers spend time on extractor development rather than other efforts that are more core to your business?
- What happens if the engineers who build the extraction code leave your company?
- Who is going to be responsible for maintaining your in-house extraction tool? Do they have the skills and time they need to keep your web data extraction processes flowing seamlessly?

Do you need an on-premise or SaaS solution?

What types of license and subscription models work best for my business and budget? How are fees calculated, and do they continue to make sense if new use cases arise over time? How much hardware do you need to procure for on-premises solutions? How do you ensure scalability and reliability as your web data needs grow? What happens if your IP addresses get blocked by the websites you are extracting data from?

Do you need a visual interface or developer tool?

Think about your users. Your ideal vendor will be different depending on if your intended user is an analyst, developer, or data scientist. The tool must be designed to make the user more productive.

Within your organization, subject matter experts know what data will provide value and where the data is located on the web. If these staff members are the ideal target user, they should be able to use the chosen solution to quickly create web data extraction workflows and gain value—even if they don't know how to code or have other technical skills. This requires a solution with an intuitive, graphical user interface. On the other hand, if your target users are developers, make sure these team members have advanced developer-oriented features like: the ability to write custom web data extraction rules using XPath, the ability to write custom logic using JavaScript and the ability to use APIs for data integration.

The solution that you select should have capabilities for running and automating workflows across the entire Web Data Integration process, including data extraction, preparation, and integration.

What sort of data security do you need?

Consider your security requirements for data storage and processing. Should data be centrally stored in the cloud or on an individual users' hardware and circulated via email? How would you mask personally identifiable data that may be extracted from the web in order to be GDPR compliant?

What integration points need to be supported?

API integrations, JSON file formats, CSV files, Amazon S3 uploads, or direct database loads? After the web data is identified, extracted, and prepared, how will it then be used and what format does it need to be in? Will it be integrated directly into an application or business process? Or will you need to feed a data lake or data warehouse for analytical consumption? Do you need a production-ready API to integrate data into your internal and external services?

Before making any investment in Web Data Integration, make sure to have a comprehensive technical understanding of the way you will want the data to be structured, modeled, and integrated into your IT infrastructure.

What about the commercial considerations?

How much are you willing to spend to find the right solution? Look for vendors whose price includes not only the software, but the levels of service you may require. Assess the amount of internal resources you have and their capabilities. Most vendor pricing will primarily depend on the number of websites, volume of data needed, and frequency of collection.

What level of data quality do you need?

Trustworthiness is a critical factor in determining whether using data for business decisions will be helpful or harmful. What steps do you need to take to ensure a high level of data quality? Do software providers offer service level guarantees on the accuracy of data? What's needed to clean, standardize, and optimize it? Are QA processes in place to manually sample extracted data and compare it against screenshots to ensure reliability? Another point to consider is the timeliness and completeness of the data. If it takes too long to conduct quality assurance on the web data, the data might grow outdated before it is ready to use.

The problem with raw extracted data from poor quality web data management software is that it can be inconsistent, incomplete, and in some cases irrelevant. The top complaint amongst business analysts, data scientists, and other users is a lack of high-quality data from in-house or existing web data solutions. Cleaning and normalizing data from different websites isn't easy. But when it's done well, the benefits and competitive advantages delivered can be huge. In fact, those results are a big reason to make the case for Web Data Integration.

Look for software providers that offer service level guarantees that specify data quality measurements.

Creating a requirements document for vendor evaluations

Now that both business and technical aspects have been covered, you should have a good understanding of what you need to get started. The next step is to consider the requirements for the various tools, technologies, and techniques that are available to get the data you need.

Narrow your list of candidates by considering the following questions:

Does the vendor support both self-service and managed-service approaches?

For legal reasons, will the vendor provide indemnification to customers, protecting them from any legal risk associated with getting data from the web?

Is the solution SaaS based or on premises?

Can the vendors' solutions access hidden web data that's not visible to the eye or web data that's generated dynamically?

How does the vendor test and verify the web data that's extracted? How do they ensure data quality?

What type of support does the vendor provide? Will they be available to help if things go wrong?

How scalable is the software? Can it keep up with your Web Data Integration needs if your volume of data grows 10-fold or 100-fold?

Licensing is another major consideration in any software decision. Ask what your options are and determine which licensing model is best for your particular use case. A self-service model may be ideal for some needs, while managed service approaches could prove more effective for others. Before you commit to any Web Data Integration solution, it's also vital to consider how your needs might change in the future. Can the vendor's software scale and support greater complexity?

Conducting a proof of concept

While a standard product demo can provide a taste of a solution's capabilities, it may not be enough to help you decide if the software is right for you. Look for vendors that will tailor a demo to your specific use case. Give your vendor a sample of websites from which you need to extract data. Do everything you can to ensure the demo is focused on what you need to see, rather than what the vendor wants to show you. While planning the demo, these are the three things that should be defined:

Effort and duration

Give the vendor a deadline to meet. Typically, it should not take more than two weeks to deliver a proof of concept.

Project scope

Restrict the scope of artifacts requested, focusing on the subset of websites and data types that you need to address a specific problem.

Success criteria

Make sure that the success criteria include a means to evaluate whether the data provided meets quality requirements.

Meanwhile, sign up for a trial and have the people who'll actually be using the software do a full evaluation of the product. Make sure that the software helps the user be more productive and covers all aspects of the Web Data Integration process—including extraction, preparation, integration, visualization and analysis. The product must be designed for self-service—meaning that users can execute the entire process on their own. If the analyst needs to rely on other staff to extract, integrate, and prepare the data, the process becomes fragmented and this may introduce errors and delays. Your team should also test the software to see how it meets their real-world requirements and performance needs. Both the demo and the evaluation are important in helping you assess an offering's technical capabilities, as well as the vendor's commitment to quality service and response times.

Pro tips

Take time and try to actually use the software. Ideally, an evaluation will approximate your real-life use cases as closely as possible.

Agree on pricing and commercial terms before embarking on a proof of concept.

Check to see if the vendor offers additional, compelling capabilities, such as reporting, charting and dashboards.

Implementing the solution and using the data

Having documented your requirements, goals, and milestones, it's time to implement the solution and start turning the web into a business-critical data source. Your business needs can change so it is important to make sure that you have a documented change management process in place, this allows you to keep your project flexible and responsive and helps to ensure that things will get corrected if any problems arise. Start small and scale up gradually.

Once web data is extracted, prepared, and integrated, it can be consumed in much the same manner as business intelligence data.

Trends or key performance indicators (KPIs) can be visualized graphically, displayed in dashboards, embedded into enterprise applications, fed into rules-based notifications, or used as datasets for training machine learning models.

Measuring business value

Every project needs to be measured and for this we've come full circle to where we began. We started this process by identifying the business objectives and value that can be realized through the utilization of web data. Use the anticipated value as a way to measure your return on investment (ROI). How will web data make an impact on your business, clearly identify metrics that you expect to be positively impacted via your use of web data. Look for activities that can be measured. For example, online retailers may be able to realize measurable improvements in a number of areas through their use of web data. They may see improvements in the number of customers acquired, average purchase amount, overall profitability, repeat purchase rates, customer churn rates, and customer lifetime value. Manufacturers may increase revenue and market share by eliminating price erosion by monitoring distribution channels for MAP compliance. Fund Managers can improve investment returns by leveraging alternative data sources to fuel investment models.

On the other hand, broader metrics like improved customer satisfaction or loyalty, while obviously good for the business, may be harder to quantify and not be as well-suited to an ROI calculation.

Conclusion

Executives in companies from a broad range of industries are quickly realizing the value that can be found in datasets that reside outside of their organizations' walls. As a result, many are turning to the web as a key source of intelligence.

In today's data-driven business climate, executives should be able to make informed business decisions based on web data—without having to worry about data quality or reliability issues. Web Data Integration solutions can yield quality data, while eliminating much of the complexity, effort, and cost associated with custom built solutions. Too often, these projects lacked common tools and processes, which left a disconnect between the business users that had a problem to solve, the engineers who wrote the scripts, and the analysts looking for patterns and insights. Web Data Integration removes the fragility and friction that accompanies traditional web data extraction projects and allows the enterprise to use web data with the same level of trust that are associated with internal enterprise datasets.

High-quality Web Data Integration solutions enable the speedy and repeatable automation of web data capture and aggregation. Now more than ever, these capabilities are essential for teams looking to employ web data at scale in order to support critical business functions.

Making the right choice for your business takes research and preparation, but those efforts can pay off in multiple ways. When you find the vendor and software that's best for your needs, you maximize your ability to reap both short- and long-term benefits. By following the advice in this buyers' guide, we hope you can find the solution that's ideal for your company's unique requirements.

Appendix: popular use cases for web data integration software

At the enterprise level, Web Data Integration is being employed in a wide variety of ways—and the opportunities are virtually limitless. In this section, we describe some of the most common applications of Web Data Integration.

Monitor products and prices

Gather comprehensive product and pricing data to monitor competitors online. Continuous price monitoring can also provide valuable insights into competitors' product portfolios and strategies.

Track customer sentiment

Tap into customer demand by gauging customer sentiment towards products. Leverage social media to predict how the activity around a specific product can affect its performance in the market.

Gain market intelligence

Automate the harvesting of market data to stay up to date with emerging trends and make better decisions for capitalizing on new market opportunities.

Expand addressable markets

Keep up to date on current property inventory levels. Track pricing, details, and availability for properties in specific areas of interest across all channels.

Aggregate market data

Market information is freely available on the web but is found across hundreds of websites. Receive a continuous stream of corporate operational data and eliminate the need to manually comb through multiple websites and online databases.

Guide product strategy

Identify problem areas and optimize product mix by understanding pricing and sentiment from customers and fans. Uncover new white space and rapidly enter new markets by identifying opportunities for products at certain price points.

Conduct economic and investment research

Reliable Web Data Integration that produces data fit for normalization and analysis can power investment decisions on a near real-time basis.

Perform risk management

Beyond reputation monitoring, many risk management measures can be enhanced with Web Data Integration. Web data can be used to complete background checks on employees and suppliers, and to manage counterparty risks. These solutions can also be used to improve your understanding of existing customers.


The buyer's guide to web data integration
import.io
Learn more at import.io