REPORT REPRINT

Import.io bolsters AI-driven capabilities for web data integration

FEBRUARY 21 2019

By Paige Bartley

Web scraping has traditionally been done with scripting, and data must still be integrated and cleansed before it is useful in analysis. Import.io offers a SaaS platform that performs the full lifecycle of web data collection: from extraction to analysis.

Web Data Integration

A new approach to acquiring and managing Web Data

Web Data Integration Journey

In today's data driven economy, organizations can no longer rely solely on internal data. They must also leverage external data sets to drive decisions or be at risk of being disrupted. In fact, a recent NVP survey showed an astonishing 79.4% of executives feared disruption from data-driven competitors.

The world's biggest repository of external or alternative data, the internet, is a source of tremendous potential. But the data on the web is neither structured nor organized to be consumed programmatically, rather – it is formatted to be read by humans. To drive insight and innovation this data needs to be accessible, machine readable and consumable at scale.

Organizations currently trying to programmatically access and integrate web data are doing so by using legacy web scraping tools. Unfortunately, web scraping tools are incomplete and insufficient to deliver on the promise of web data. They can only access a fraction of the data on the internet, they provide little in the way of data quality, and they must be integrated with other tools to deliver real value. This leaves organizations either missing the opportunity to leverage web data or with incomplete data access, poor data quality, unreliable and out of date data, high costs and uncertain business risks.

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

- **IDENTIFY** the URL where your data is located. Simply point and click to show us what data you need. Alternatively, our machine learning based auto suggestion feature makes "one-click to data" a reality.
- **EXTRACT** displayed or hidden content from anywhere on the web. Behind a login, across multiple pages or require interactions, Import.io can extract exactly what you need, when you need it.
- **PREPARE** extracted data by exploring, assessing and refining the data quickly. Cleanse, normalize and enrich the extracted data using 100+ spreadsheet like functions and formulas.
- **INTEGRATE** prepared data with a library of APIs to support seamless integration with internal business systems and workflows or deliver it to any data repository to develop robust data sets for advanced analytics capabilities.
- **CONSUME** prepared data with graphs and charts to find answers and glean insights. Analyze data with change, comparison, and custom reports.

Import.io is the leading Web Data Integration provider, delivering the world's data directly to enterprises, fueling business insight and competitive advantage. The Import.io Web Data Integration platform provides you with everything you need to identify, extract, prepare, integrate and consume high-quality, comprehensive web data into analytics platforms and business applications. The company delivers data to more than 850 customers worldwide from millions of web sources.

Summary

The concept of web scraping is not new, but the process of deriving insight from data collected from the internet has traditionally been a multistep process requiring multiple tools and techniques. Data must be extracted, typically using scripting, then it must be prepped and integrated and put into a format that can then be fed into a preferred analysis tool. Scraping data is simply the first step in a longer workflow, and if that workflow takes too long to complete, the data being extracted and analyzed may no longer be relevant to the intended use case. Import.io aims to offer a SaaS offering that addresses the entire web data collection workflow: data extraction, integration and cleansing, and analysis. The company wants to demonstrate value with a cohesive platform that can help organizations operationalize the process of web data integration and systematic data assurance, rather than having it be a tedious, ad hoc process.

451 TAKE

As data, and the ability to derive insight from it, becomes the competitive differentiator for business, organizations are looking to supplement the data they already manage and analyze internally with external and public sources, such as those found on the web. When it comes to deriving insight, being able to consistently integrate and cleanse this data, rather than just collect it, isn't just half the battle: it is the battle. Many free tools and frameworks exist for the various steps of the web-scraping-to-analysis workflow, but Import.io is betting on its full-lifecycle approach to deliver value in terms of time savings, consistency and the ability to scale far beyond what any DIY approach could offer. Still, convincing enterprises can be a challenge; the mentality that free is always a better value pervades. However, competition in this space suggests that a real business pain point is being addressed and that there is a real value proposition.

Context

Import.io was founded in 2012 in London with its original offering: a simple web data extractor. Since then, the company has moved its headquarters to Los Gatos, California, and broadened its focus and platform to support the full web-data-integration lifecycle, from data collection to analysis. With the premise that the internet is the enterprise's largest potential source of information, the company is aiming to provide a single environment where formerly ad hoc collection, integration, cleansing and analysis workflows spanned multiple tools and products.

Prior to its recent acquisition of Connotate (see Strategy section below), Import.io had roughly 800 paying customers, 60% of which are in the US, 30% in Europe, and 10% in the rest of the world. Because web data use cases are so diverse, Import.io customers, unsurprisingly, span industry verticals and range from small startups to large enterprises. The common denominator of clients isn't firm size or specific industry but, rather, the appetite for external data sources. The company currently has 50 full-time employees plus 20 contract workers, with additional offices in Boulder, Colorado and London.

Import.io closed a $15.5m series B round of funding in December 2018, which was led by Talis Capital in the UK. Its funding, to date, totals roughly $38m. Other investors include Delin Capital, ipgroup, OpenOcean, Oxford Capital, Wellington Partners and AME Cloud Ventures.

Products

The namesake and flagship offering of the company is the Import.io SaaS platform, which has a simple enough premise: to provide a single environment where the entire web data integration workflow can be conducted: from initial data collection to analysis and insight. Most available tools, especially free tools and frameworks, focus only on the initial – scraping – step of the process, rather than downstream data integration, cleansing and analysis steps. They are labor-intensive and disjointed. The data-extraction process itself poses perhaps the least technical difficulty; as is the case with many data-driven enterprise initiatives, it's the integration of data into business applications and preparation of data that take up inordinate amounts of time and skill. The platform seeks to unify and operationalize what was, formerly, a chain of ad hoc processes and tools within most organizations. The Import.io platform can be leveraged as a standard SaaS log-in environment, or its capabilities can be embedded in other applications via APIs.

The Import.io platform is modular, offering functionality for five progressive steps as they correspond to the web-scraping-to-insight workflow: identify, extract, prepare, integrate and consume. Not all customers will utilize every module of the Import.io platform. While 'identify' and 'extract' are core web extraction capabilities that practically all customers leverage, some organizations have other preferred tools in place to 'prepare' or 'consume' data. Examples would be self-service data prep environments for blending and cleansing, or self-service analytics tools for visualization and insight. Generally, smaller customers use the platform's capabilities end to end, while larger enterprise customers are more likely to have existing, complementary software deployments that can help handle the integration, prep and analysis aspects of the web data integration workflow.

Still, the maximal value proposition of the Import.io platform is achieved when it is used cohesively, to the full extent of its capabilities. If web data extraction, integration and analysis can occur in a single SaaS environment, the business has more control and clarity over the process, leading to insight that is more accountable. The Import.io platform aims to unify formerly disjointed tools and processes, and from a data governance perspective, this consolidated control streamlines auditing and oversight, which is increasingly necessary for data privacy and data protection requirements such as the EU's General Data Protection Regulation and copycat legislation in regions around the globe.

Strategy

The company's initial product in 2012 focused on web data extraction, but this soon proved (at least in isolation) to be insufficient to compete with the proliferating market of free tools and frameworks available for that specific functionality. The opportunity in the market was to create a single platform for a series of traditionally brittle, hand-coded processes that were difficult to scale and adapt, which resulted in scraped data that had to be tediously integrated, prepped and cleansed before being fed into an analytics platform or tool. Import.io positioned its platform approach as 'web data integration,' emphasizing functionality for the traditionally time-consuming steps that need to be taken once data is scraped. The addition of embedded analytics capabilities into the platform created an 'end to end' environment that could handle the entire workflow from data collection to delivery of insight.

Earlier this month, the company announced the acquisition of Connotate, a web-data-extraction company focused on enterprise clients and AI-driven extraction methods. The rationale for the acquisition was threefold: in acquiring Connotate, Import.io is onboarding more than 70 enterprise customers, retaining nearly the entire roster of Connotate talent, and expanding its intellectual property portfolio with eight patents. Strategically, the Connotate acquisition is especially significant given Connotate's focus on AI and machine-learning-driven extraction techniques, which will expand Import.io's capabilities to automate scraping and extraction of data from increasingly complex website structures. Embedded AI and machine-learning-driven functionality is becoming a key differentiator between data management software offerings, so the acquisition comes at a pivotal time. It should help Import.io fortify native extraction capabilities, which originally used more of a traditional, heuristic model.

Competition

While Import.io does much more than web scraping or extraction of data, the product is inextricable from this capability. In the eyes of the enterprise, that puts it in competition with the multitude of freely available ('freemium') web-scraping tools, packages and frameworks. Python and R are common languages for scripting, Scrapy and GoogleScraper provide frameworks, the Beautiful Soup Python library converts data into parsible HTML, and products such as Scraper API can help circumnavigate CAPTCHAS and blocked IPs. Other scraping tools include Portia, Kantu, ParseHub and Octoparse. Today's websites, with complex code and structure, often require a daisy chain of individual tools to emulate and render interactive content into machine-readable format. These tools, however, are focused primarily on scraping and scraping alone: not the entire workflow leading to data integration and final analysis.

For integration of data, Import.io often competes with incumbent Excel spreadsheets, or (to some extent) providers such as Informatica and Talend. Data preparation and blending is a market of its own, and many organizations have either stand-alone or embedded self-service data prep functionality: Alteryx, Trifacta, Paxata, Datawatch, ClearStory Data and Unifi Software are all examples of this category. For visualization and analysis, common BI and self-service analytics tools such as Tableau, Qlik, Microsoft Power BI and Oracle Analytics Cloud all overlap with Import.io's analytics capabilities.

If final data output is the main consideration, Import.io competes with data-as-a-service providers such as PromptCloud and Scrapinghub. These players orchestrate and manage the web-scraping workflow, delivering the final data product to the customer in the desired format. Because they are services rather than SaaS platforms, the enterprise cedes control of the actual management of the data collection, data integration and data formatting process. Additionally, these services simply deliver the data in the agreed-upon format; there are no native analytics or visualization capabilities, and the organization must feed the data into existing BI or analytics tools. However, it should be noted that Import.io also offers a fully managed as-a-service option.

SWOT Analysis

STRENGTHS

A single SaaS offering for the entire workflow from web data extraction to analysis allows organizations to unite and scale what were formerly, in most cases, ad hoc processes. A single platform approach allows for faster time to insight than individual tools and manual methodology.

WEAKNESSES

The flipside of Import.io's approach is that its platform's functionality potentially overlaps with numerous other enterprise software investments. The company will have to clearly demonstrate ROI for prospects that already have preferred analysis or prep tools.

OPPORTUNITIES

Risk-reduction use cases for extracting web data, such as vendor risk assessment, are growing. If Import.io can demonstrate value for these uses while continuing to strengthen its product controls for privacy and data protection, it should be well positioned to benefit from these trends.

THREATS

Web scraping and data privacy are active topics of litigation worldwide. As the legal landscape develops, there is always the possibility that Import.io's business model could be threatened by case law decisions that place restrictions on the automated collection and analysis of web data.