Web Data Integration: A new source of competitive advantage

An Ovum white paper for Import.io

Publication Date: 29 January 2019

Author: Tony Baer

Summary

Catalyst

Web data provides key indicators into a company's competitive landscape. By showing the public face of how rivals position themselves and providing early indicators of changing attitudes, sentiments, and interests, web data complements traditional enterprise data in helping companies stay updated on their competitive challenges. The difficulty is that many organizations lack a full understanding of the value of web data, what to look for, and how to manage it so that it can live up to its promise of complementing enterprise data to provide fresh, timely views of a company's competitive landscape.

Ovum view

Web Data Integration applies the discipline and automation normally associated with conventional enterprise data management to web data. Web Data Integration makes web data a valuable resource for the organization seeking to understand their competitive position, or understand key challenges in their market, such as the performance of publicly traded companies or consumer perceptions and attitudes towards products. Web Data Integration ventures beyond traditional web scraping by emulating the workings of a modern browser, extracting data not otherwise accessible from the simple parsing of HTML documents. Automation and machine learning are essential to the success of Web Data Integration. Automation allows the workflows for extracting, preparing, and integrating data to be orchestrated, reused, and consistently monitored. Machine Learning (ML) allows non-technical users to train an extractor for a specific website without having to write any code. With Web Data Integration, web data now becomes a first-class citizen in the enterprise data environment that can be used to generate insights, complementing or augmenting the traditional data that sits in data warehouses or is utilized for big data analytics. Import.io provides a good example of a solution that seeks to transcend conventional web scraping with a SaaS-based Web Data Integration solution.

Key messages

- Conventional "web scraping" techniques that parse HTML documents are missing the big picture that web data can provide.
- Web data can yield valuable insights if treated with the same degree of discipline for extraction, transforming, and cleansing as enterprise data.
- Web Data Integration (WDI) is a new approach to acquiring and managing web data that focuses on data quality and control.

The value of web data

The new reality

Enterprise organizations are entering a whole new era of data-driven competition. Data need no longer be used just for record keeping, reporting or directional decision-making but can inform every part of the business, in real time. The products and services of tomorrow are going to be data-driven: their design and behavior determined by patterns in data. Companies need to look beyond internal

© 2018 Ovum. All rights reserved. Unauthorized reproduction prohibited.

data sources and commodity data providers when selecting the data sources that they will use to solve business problems. These traditional data sources are too limited in scope.

The opportunity

The web is the largest repository of content ever created and has incredible potential as a business data source. Some of the best data about a company's customers, competitors and suppliers exists out there, on the web.

The problem

The web was built for humans to read, it was not built for machines. Websites are not required to have machine-readable APIs that serve data in a standardized format. The traditional method for making a website machine-readable is known as "web scraping", it is as old as the web itself and has not evolved much over the years.

Web scraping projects are notoriously complicated, expensive and labor-intensive. They require organizations to employ engineers to write custom software for every type of web page that they want to target. In addition to engineers, organizations will also require subject matter experts to conduct quality assurance (QA) by comparing the spreadsheets of extracted data with the actual web page(s). Compounding the challenge, by the time that the subject matter expert finally gets a chance to review the data, the web page(s) in question may have changed, and screenshots of the page(s) may not be available. Consequently, conducting QA on the data is not as easy or cut-and-dry compared to the QA processes used for spotting bugs in software development projects.

In addition, web scraping projects are brittle: they are not resilient to change on the target website and they break easily; this is caused by the fact that extraction rules are hard-coded and are only ever informed by a sample of pages from the website. If the website changes, or the engineer did not sample a sufficient number of web pages when writing the extraction rules, the organization is left with incomplete, poor quality, unreliable and out-of-date data. These gaps or errors are often not caught until the data is analyzed by business users, and in many cases, the fixes may introduce new extraction errors, making the cure worse than the disease.

Enterprises cannot build on incomplete and poor-quality datasets and so web data projects are either abandoned before they ever reach their full potential or become prohibitively expensive as the organization battles to fix the problems as they arise, diverting technical personnel away from other core activities.

The way it should be

Companies should be able to build products and services on web data with confidence and without worrying about data quality or reliability issues. Web data should be a first-class citizen in the enterprise data environment.

Table 1. Selection of real-world use cases for web data

Monitor competitor product pricing strategies over time	Monitor industry blogs, social media, news sites etc. in order to understand sentiment and mood towards particular products or companies	Build data sets for training machine learning models for new data driven product development using real world, user generated content	Monitor and enforce minimum advertised pricing polices across retailers
Track consumer reviews to understand product quality trends and perceptions	Understand share of market by finding and matching customers that are doing business with competitors, useful for marketplace providers for real estate, events, food delivery, etc.	Track the opening and closing of physical store locations as they are added and removed from online store finders	Monitor government websites for regulatory updates, legislative agendas and approval notices affecting certain products or industries
Identify sellers of counterfeit goods on secondary marketplaces and automatically issue de-listing requests to marketplace operators	Monitor product market trends for informing new product releases (what materials to use, what price, what product mix etc.)	Perform online due diligence on companies and people in the supply chain in order to monitor for sanctions risk	Determine or predict sales and inventory status at competitors to understand competitor product performance and to identify opportunities

Source: Import.io

Web data: the misconceptions

There is a misconception that web data is low value data. This view is likely caused by the limited effectiveness of traditional web scraping practices that rely on error- or gap-prone manual coding, rather than issues with the inherent value of the web data itself. Traditional web scraping approaches that simply parse HTML documents will fall short, especially with modern websites that use JavaScript, APIs and other non-HTML assets for the display of data. For instance, parsing the HTML of an online shopping page will not necessarily capture the additional information about a product that appears when the cursor hovers over an image or when product-size or -color variations are selected. The result is that organizations limiting themselves to traditional web scraping will gather limited datasets, and therefore, the analysis of the resulting web data will yield only cursory results that will not reflect the full value of the data that exists on the web.

Web data: the opportunity

There is more to web data than limiting collection to the content that is visible on individual, static web pages. There are valuable insights to be gained by probing beneath the surface. For instance, by actually interacting with a product page and adding the item to a shopping basket it is possible to reveal in-basket pricing, which can be different to list-pricing; by selecting delivery options it is possible to view delivery prices; by interacting with an online calendar and specifying a date it is possible to reveal availability information.

By applying robust data quality and validation practices, such as filtering and harmonization, web data can become as valuable as any other enterprise data.

Introducing Web Data Integration

What is Web Data Integration?

Web Data Integration (WDI) is a new approach to acquiring and managing web data that focuses on data quality and control. A WDI *project* treats the entire web data lifecycle as a single, integrated journey from the initial identification of data requirements and the actual process of web data extraction through the intermediary stages of data preparation, cleaning, analysis and visualization and then finally into data integration and consumption by downstream applications and business processes.

The purpose of adopting a WDI strategy is to allow the enterprise to use and build on web data with the same high levels of trust and confidence that are associated with internal enterprise datasets. WDI solves for the fragility and friction that accompanies traditional web scraping projects where there are no common tools, processes or standards joining: the business team that owns and defines the problem to be solved, the engineering team that writes the scripts to get data from the website and the data analysts that test and explore the dataset for patterns and insights.

Requirements for a WDI solution

To deliver on the promise of enterprise-quality web data, the WDI project and technology solution must meet several different requirements. These requirements can be organized into five groups according to the different stages of the web data lifecycle including identify, extract, prepare, integrate, and consume.

Identify

The project team must have a clear understanding of the business problem and exactly how web data will help. It is useful to know at the beginning of a project if the web data will be used to integrate directly into an application or business process or whether the web data will be used to drive an analytical investigation. The value of web data increases as it is gathered over time and it becomes possible to perform time-series and trend analyses on the data. For instance, time series analyses can be especially useful when analyzing the competitive impacts of significant events, such as a sports team championship playoff win; a surprise celebrity endorsement; or the launching of a limited-time sale or product promotion.

At the outset of the project, the project team should ask themselves, what trends they expect to see in the web data as it is collected over time. The Identify activities of a WDI project are generally led by the business owner who will have subject matter expertise regarding the target websites, the desired data and how the data will solve the business problem. If working with web data is new to an organization, it is possible for an outside consultant to lead an internal team through an Identify process.

Extract

Modern websites today do more than simply render HTML documents; today, websites are effectively complete software applications. Consequently, parsing just the HTML content of a webpage is akin to viewing only the tip of an iceberg. Web data extraction must be capable of fully *emulating* the operations of a modern web browser, including: rendering CSS (Cascading Style Sheets); processing

© 2018 Ovum. All rights reserved. Unauthorized reproduction prohibited.

JavaScript; interpreting network traffic (including API calls made from a web page); securely authenticating the user session; storing cookies; and maintaining session integrity.

Ideally when presented with a web page, the WDI solution will be able to automatically generate an extractor that targets the likely data of interest on the page. As noted below, this is a golden opportunity for employing machine learning that provides a more adaptable approach compared to traditional methods of hard coding static rules. Ideally, a hybrid approach that combines automatic extraction with supervised point-and-click training and supplemented by the ability to write extraction rules, is the most practical approach. Import.io's hybrid approach is described later in this document.

The Extract training process involves building a workflow that allows automated extraction agents to navigate through a website, extracting data along the way. Workflows are required for scenarios involving pagination, infinite scroll, list-detail patterns, form-filling, click interaction, authentication, and so on. As part of the training process, the user should have the chance to adjust or calibrate the automated workflow through manual point-and-click training on the web page. Once extractors are set up and gathering data they should be continually monitored to ensure that they are still running and still extracting the data that is expected.

The automated approach to extracting web data has many advantages, such as:

- Allowing non-technical users, including business subject matter experts, to easily participate in training the system to extract the required data from a website
- Enabling the system to automatically adjust (or self-heal) the extraction routine as the target website changes; and
- Making the re-training of the extractor a rapid process.

Prepare

The data preparation step, which has long been a staple of cleansing or harmonizing enterprise data, is a critical part of the Web Data Integration process as well. The Prepare process involves data wrangling, performing tasks such as splitting or combining columns, de-duplicating rows, interpolating gaps in sparse datasets, harmonizing data formats, schemas or structures, and generally cleaning the data. With the data harmonized and cleansed, and rules set up for this process to be repeated on new data coming from source, the stage is set to create calculated fields, run extraction on a regular schedule, and set up data pipelines that allow pre-built transforms to be uniformly applied on every extraction. As part of the Prepare stage business analysts should be able to perform exploratory analysis and data visualization, allowing them to see summary statistics and understand data distributions so that they can determine the data wrangling operations that need to be performed.

An important part of the process is establishing a quality assurance (QA) loop; once data extraction is set up and running, QA should be an ongoing activity that involves manual sampling of extracted data: having people visually check the data against screenshots of the web page taken at extraction time. Making QA an integral part of the Web Data Integration workflow is critical to validating that extracted and cleansed web data paints a reliable picture of what exists on the web.

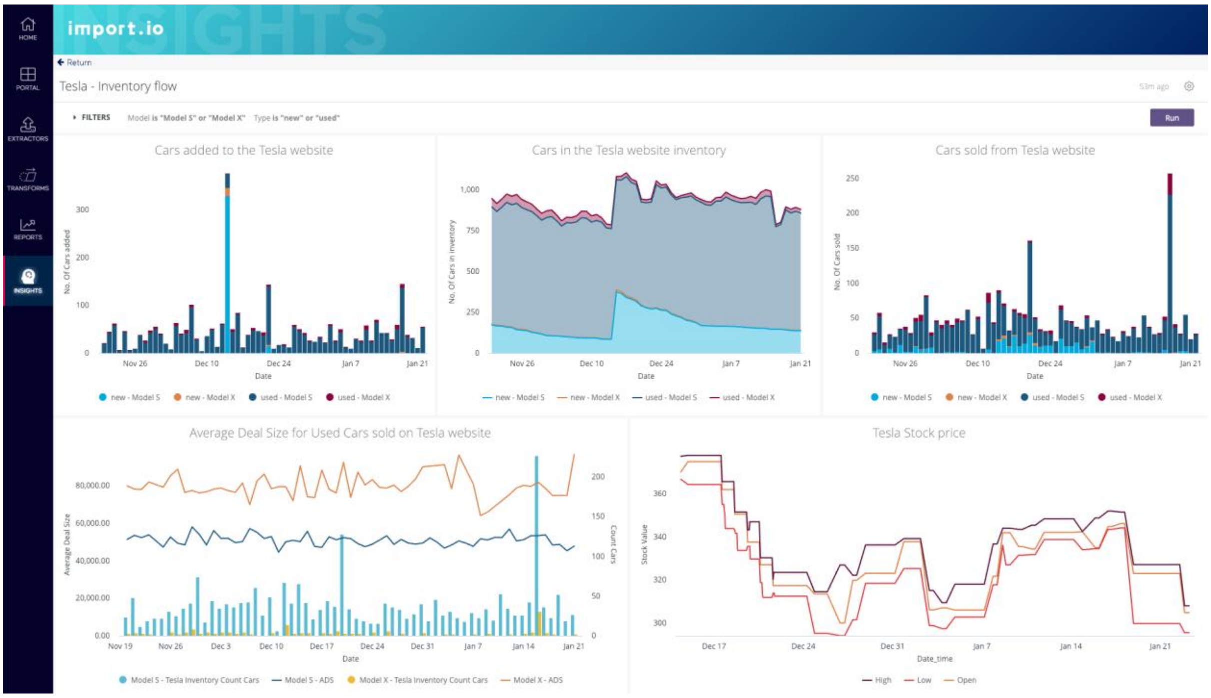
Integrate

APIs are a key component for making web data accessible as they enable external tools or dashboards to be used to control the entire WDI process. Additionally, APIs make web data consumable; they allow web data to be queried using SQL and/or integrated into third-party business intelligence solutions or enterprise applications. For example, with an API, requests for web data from a particular page or pages can be made directly from an application in real-time and the resulting data can be re-integrated for display to end-users.

Consume

Once web data is extracted, prepared, and integrated, it can be consumed in much the same manner as business intelligence data. Trends or KPIs can be visualized graphically and displayed in dashboards or embedded into enterprise applications, changes in the underlying data can be used to feed rules-based notifications, and the datasets themselves can be used to develop new data-driven products by serving as training data sets for ML models. Figure 1 shows an example of a web data dashboard, used by a hedge fund that is monitoring and analyzing the vehicles available for sale on the Tesla website.

Figure 1. Web data dashboard example



Tesla - Inventory flow

FILTERS Model is "Model S" or "Model X" Type is "new" or "used"

Source: Import.io

Import.io's approach to Web Data Integration

Automating the workflow of Web Data Integration

Import.io offers a SaaS-based (cloud) solution that helps enterprises harness web data by treating it with the same robust extract, preparation/transformation, and integration practices that would be expected for conventional BI analytic data or big data. It adapts or extends these practices to the world of web data. Import.io's web data extraction capability is designed for the modern web, which often requires website interactions during the extraction process. For instance, Import.io fully emulates the function of a modern web browser, going beyond parsing static HTML content by rendering web components such as CSS and processing JavaScript. Reusable workflows can be designed to emulate web interactions, such as using available APIs calls to external web apps or data sources or authenticating user sessions in order to perform interactions such as checking adaptive drop-down menus or filing out web forms and storing cookies.

Import.io provides a visual environment for automating the workflow of extracting and transforming web data. After specifying the target website url, Import.io's web data extraction module provides a visual environment for designing automated workflows for harvesting data, going beyond HTML parsing of static content to automate end user interactions yielding data that would otherwise not be immediately visible. These workflows automate the crawling of web pages; the orchestration of interactions that are conducted for extracting data; and the scheduling of extractions to make Web Data Integration an ongoing process. Once the web data is extracted, Import.io provides full data preparation capabilities that are used for harmonizing and cleansing the web data and offers a library of over 100 Excel-like functions enabling the end user to build custom formulas that create new calculated fields in the dataset.

For consuming the results, Import.io provides several options. It has its own visualization and dashboarding module to help business analysts gain the competitive insights that they need, and it also provides APIs that offer full access to everything that can be done on their platform, allowing web data to be integrated and consumed using external applications, such as customer experience management software, segmentation models, or other analytics applications and tools.

The role of machine learning

The benefit of automating web data extraction workflows is that they can be trained by non-technical subject matter experts rather than software engineers. Machine Learning (ML) makes this a more adaptable process that is less likely to break when the underlying websites change.

Here's how it works. The user can point Import.io to a URL and the product then highlights the data of interest, and the user either accepts or rejects the result, alternatively the user points-and-clicks at the desired content on the web page and the product automatically generates extraction logic. In most cases, the user will start with the fully automated approach and then, using point-and-click training, will fill in any gaps. Either way, this hybrid approach can far more readily adapt to changes in web pages compared to the manual approach of coding of static extraction logic that will break when the underlying website changes.

Takeaways

While some enterprises already use web data to gain insights about their rivals and understand the changes occurring in their competitive landscapes, in most cases, web data is underutilized because of the perceived low value and low quality of the data. Typically, that is due to the fact that most organizations limit their efforts to conventional "web scraping" projects that simply parse HTML documents and fail to treat the integration of web data as a single holistic journey. Such traditional approaches can often miss valuable information that is available on a website and lack support for using processes and methods that allow multiple stakeholders or teams from business and IT to effectively collaborate.

When treated as a single, holistic workflow (from web data extraction to insight) with the same level of data validation discipline that is normally accorded to conventional BI data or big data, web data can yield valuable insights. For instance, web data can be used for continuously staying a step ahead of

© 2018 Ovum. All rights reserved. Unauthorized reproduction prohibited.

the competition by monitoring pricing from rival retailers or manufacturers. It can also be used for judging the financial health of companies through indicators such as sentiment expressed in industry blogs, social media, or news aggregator sites. It can also be used by insurers to gauge risk through tracing product reviews to gain insights into product quality or perceptions. The data gained from the web complements conventional enterprise analytic data or big data by adding evidence or providing context. For enterprises willing to go beyond conventional "web scraping," Web Data Integration can provide a competitive edge by yielding hidden insights about the market.

Appendix

Author

Tony Baer, Principal Analyst, Information Management

tony.baer@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

The contents of this product are protected by international copyright laws, database rights and other intellectual property rights. The owner of these rights is Informa Telecoms and Media Limited, our affiliates or other third party licensors. All product and company names and logos contained within or appearing on this product are the trademarks, service marks or trading names of their respective owners, including Informa Telecoms and Media Limited. This product may not be copied, reproduced, distributed or transmitted in any form or by any means without the prior permission of Informa Telecoms and Media Limited.

Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Informa Telecoms and Media Limited nor any person engaged or employed by Informa Telecoms and Media Limited accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard – readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Informa Telecoms and Media Limited.

Ovum
TMT intelligence | informa

CONTACT US

ovum.informa.com

askananalyst@ovum.com

INTERNATIONAL OFFICES

Beijing
Boston
Chicago
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
Paris
San Francisco
Sao Paulo
Shanghai
Singapore
Sydney
Tokyo