April 2014, HAPPIEST MINDS TECHNOLOGIES

Website Scraping

Author
Ritu Banerjee

Copyright Information

Contents

Abstract

The term “Application Programming Interface (API)” has been around the market for a long time but of late like “Big Data”, “Cloud Computing” the term “API” has become a buzzword which is on the mouth of countless technology industry pundits.[1] Thus, the idea of how some software components should interact with each other is not new. Application Programming Interface (API) has gained popularity in the market recently. It has become a buzzword amongst the technology industry pundits[1] just like “Big Data” and “Cloud Computing”.

Back in ’90s websites had just one distribution channel - HTML, but the evolution of API, websites can now reach thousands of users through an API channel available anywhere. Moreover back in ’90s there were just simple websites, now we have the web, mobile and device applications running through the internet. The power of information and data that can be used is much powerful than it was earlier, "Not having an API today is like not having a website in the 90s" says Martin Tantow co-founder of 3Scale.[2]

The Trend:[3]

- 1995: At the beginning there were just websites
- 2005: First websites with API as add-ons; they were an additional access/interface to their existing data: e.g Yelp
- 2008: The API has overtaken website traffic, the API is more important than the website: e.g Twitter
- 2009: The API is the product and websites/webapps have become web services: e.g Twilio, Simplegeo
- Future: Web services will become an open platform, applications will turn into platforms, everythingwill be programmable and expandable

The internet evolution has turned Web into the largest public data source in the world. It is designed to make it easy for people to find information. But people use information differently than computers do. Many businesses today are focused on leveraging their own data. Organizations are increasingly looking at extracting relevant information from the world wide websites and generate business value. The kind of information organizations want to harness include sales leads, competitive intelligence, market intelligence, news, creative content, company and sector performance data, enhanced ecommerce operations, gather contact information for use in marketing and promotional campaigns, Organizations need solutions that support better and faster decision-process. But often while finalizing the business strategy, businesses realize that they have websites but not APIs. This gap makes it difficult for companies to extract the data easily for creating new channels. In such scenario with the help of extraction solution it is possible to extract data from the existing website with the consent of the website owner and expose the content as structured APIs. But, getting to that data, and transforming it into something relevant and usable is not as easy as it

[1] http://software.intel.com/sites/billboard/article/api-transformation
[2] http://siliconangle.com/blog/2010/12/02/the-api-market-is-taking-a-big-shape/
[3] http://siliconangle.com/blog/2010/12/02/the-api-market-is-taking-a-big-shape/

may sound. Information must be searched, located, filtered, and extracted and then provided as structured APIs.

With the growth of Web Scraping API (WSAPI) solutions, not only developers and startups but also large companies with international activities recognize that there is more to Web Scraping API solutions than just marketing hype. WSAPI solution allows an organization to extend their existing web based system as well designed structured set of services for creating diverse channels.

However, a complete extraction of data from website through the medium of an API is still a vision. Many organizations in order to locate, capture, and store high volumes of information they need from targeted websites, still uses Traditional web data extraction methods:

- Manual "Cut and Paste" – time consuming and prone to human error
- Development of "Homegrown" Tools - require skill and investment of resources to maintain

Questions also arise about security and quality of service or, whether the offered services can effectively meet the company demands of maintaining and supporting the business processes. Professional providers of Website Scraping API solutions for enterprise customers must address these challenges to provide a transparent and cost-effective solution.

Introduction

The internet is a universally accessible resource for millions of people. The problem is that the overwhelmingly rapid growth of the internet has challenged the capabilities of traditional search engines and technologies. This has increased difficulties for a business in maintaining timely intelligence for strategic planning purposes. Many a times while making such strategic planning a business tends to realize that while developing their website they did not give importance to building an API. As they cared more about maintaining their public-facing website and focused less and did not concentrate on their structured data feeds, it has led to shortcomings when the opportunity was presented to extract data for creating new channels for business growth. Web data extraction is the process of transforming the useful content on websites into valuable business assets. There are several web extracting software that has emerged in the market which helps to address this problem. The software aids in extracting structured content from a web page and exposes the required services as APIs and makes it useable for further processing.[4]

For example: A Tours and Travel company envisions extending their service to their customers to book train tickets with railways agency who does not have APIs but has a website. In such a situation a simple software meant for extracting data from the web pages (web scraper) will be useful to associate the Tours and Travel Company with the railway agency. The software would extract the required data, turn them into structured APIs and expose.

[4] http://www.mozenda.com/web-scraping-software
http://www.connotate.com/uploads/Final-10-Reasons-White-pape08-07-5.pdf

Current Challenges

3.1 Traditional Extraction Method

3.1.1 Manual extraction process

The manual web data extraction process has two major problems. Firstly, it can't measure costs efficiently and can escalate it very quickly. The data collection costs increase as more data is collected from each website. In order to conduct a manual extraction, businesses need to hire large number of staffs, this increases the cost of labor significantly. Secondly, each manual extraction is known to be error prone. Further, if any business process is very complex then cleaning up the data can get expensive and time consuming. The below figure explains the errors and data cleanup processes problems with manual method.

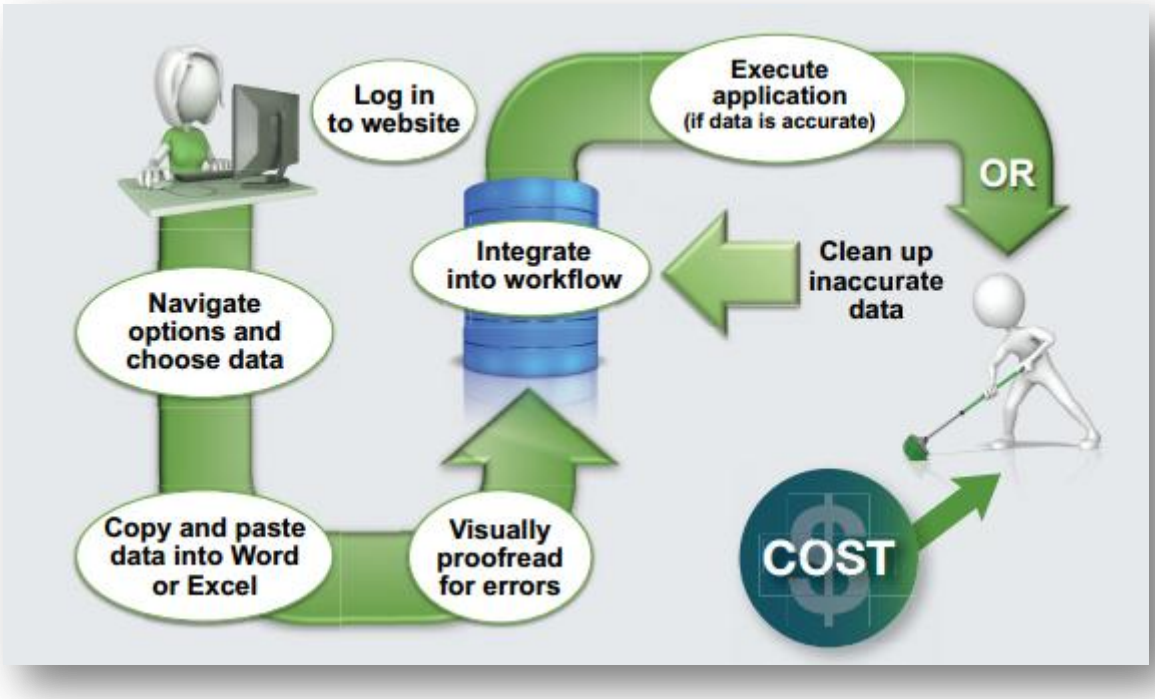

Figure 1: Manual Process introduces human error, requiring costly data cleanup to ensure quality

[Source: From Top-Line Growth to Bottom-Line Profits: 10 Reasons to Use Automated Web Data Monitoring and Extraction by Ryan Mulholland, President, Connotate]

3.1.2 "Homegrown" Tools

In the past, many companies have chosen to design and build "Homegrown Tools" to automate the manual task of extracting data. But, for companies who have not, developing such systems from scratch can be highly time-consuming, complex and costly. Even for those who have developed custom solution, ongoing maintenance and updates can be expensive. The process requires considerable amount of resources to work on building the tool, and doing so either puts extra burden on programmers to stay on track without impacting other business critical tasks or lock up expertise in the tool maintenance rather than work on more value adding projects.

3.2 Security

Companies who wish to use scraping solution services for their existing web based system have concerns to do so. They have high requirements, regarding the process that is followed to secure the company data. The services must be reliable in order to avoid putting the business processes at risk.

3.3 Quality

While scraping data from website as per the organizations requirement the quality of the scraped data also needs to be taken into consideration. If extraction process is not conducted accurately lot of information may get lost and the quality would suffer noticeably.

Definitions

4.1 Website Scraping

Web Scraping (web harvesting or web data extraction) is a computer software technique to extract information from websites. Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser, such as Internet Explorer or Mozilla Firefox.

Web Scraping is closely related to web indexing, that indexes information on the web using a bot web crawler and is a universal technique adopted by most search engines. In contrast, Web Scraping focuses more on the transformation of unstructured data on the web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web Scraping is also related to web automation, which simulates human browsing using computer software. Uses of Web Scraping include online price comparison, contact scraping, weather data monitoring, website change detection, research, web mashup and web data integration.[5]

4.2 Website Scraping API (WSAPI)

A Web Scraping service is involved in large-scale Web Scraping on the basis of custom requests. It allows access to scraped data to its clients using an API and provides companies with fresh structured data which is integrated into their systems. [6]

4.3 Our Definition

Web Scraping is a technique to extract structured data from websites. WSAPI is the platform that enables an organization to extend their existing web based system, as well designed set of services for creating new channels, developer integration or partner integration. It helps to offer clean and structured data from existing websites, so that the data can be effortlessly consumed by disparate systems. The data that is being exposed through these APIs can be monitored, transformed and controlled easily. The inherent design helps developers to incorporate website changes without affecting the extraction logic by moving them to configurations.

[5] http://en.wikipedia.org/wiki/Web_scraping
[6] http://promptcloud.com/web-scraping-api-software-google.php

Why Scrape the Web?

There are many specific reasons why businesses may want to scrape their website; one of the vital reason being the unavailability of APIs. Some of the other major reasons which may lead a company into scraping their website are:

5.1 Expand Market Share

Due to the lack of availability of APIs the possibility of collaborating with business partners is limited. By exposing the data available in their website as APIs enterprises can open up new channels, possibilities to expand the market share and increase sales.

5.2 Enter New Markets with Early go-to Market Strategy

API being the long time strategy, Web Scraping solution can potentially enable organizations to build an early go-to market strategy.

5.3 Access to Renewed and Structured Data

Scraping the website of the organization through a Web Scraping solution gives an organization the chance to access renewed, structured and up to date data through the scraped APIs.

Benefits of Scraping Solution

In order to remain competitive, businesses must be able to act quickly and assuredly in the markets. Web Scraping plays a big role in the development of various business organizations that use the services. The benefits of these services are:

6.1 Low Cost

Web Scraping service saves hundreds of thousands of man-hours and money as the use of scraping service completely avoids manual work.

6.2 Less Time

Scraping solution not only helps to lower the cost, it also reduces the time involved in data extraction task. This tool ensures and gathers fast results required by people.

6.3 Accurate Results

Web Scraping solutions help to get the most accurate and fast results that cannot be collected by human beings. It generates correct product pricing data, sales leads, duplication of online database, captures real estate data, financial data, job postings, auction information and many more.

6.4 Time to Market Advantage

Fast and accurate results help businesses to save time, money and labor and get an obvious time-to-market advantage over the competitors.

6.5 High Quality

A Web Scraping solution provides access to clean, structured and high quality data through scraping APIs so that the fresh data can be integrated into the systems.

Some Business Use Cases

- **Gathering data from multiple sources for**
  - Market analysis and Lead generation
  - Research
  - Data Integration

- **Helps monitoring of**
  - Competitor's inventory information
  - Stock prices
  - Order status from ecommerce portals
  - Opportunities

- **Automation of repetitive tasks**
  - Procuring inventory
  - Getting product reviews on message boards and forums

Happiest Minds Website Scraping API (WSAPI) Solution

8.1 Happiest Minds Website Scraping Solution

Happiest Minds' has developed a Website Scraping API (WSAPI) solution that enables to create alternative API channels for the existing websites of an organization. Happiest Minds created a robust API infrastructure for managing and monitoring the API channel. The solution is developed to help an organization in:

- Packaging their existing web solution and to support their business strategy
- Building successful mobile app strategy
- API enable their solutions for their customers or partners. API enabled solutions for their customers or partners.

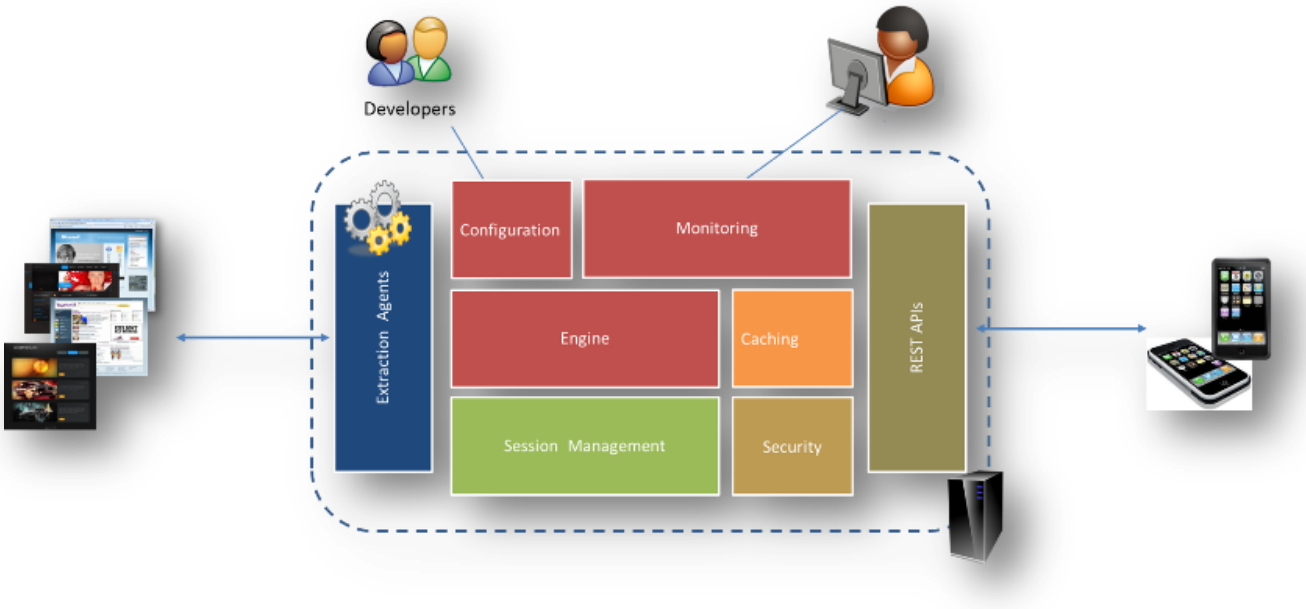

Figure 2: Website API (WSAPI) Solution Overview

Unique proposition of Happiest Minds' WSAPI solution

- Define data sources from various websites without using any programming language
- Designer tool for developers to easily set- up and configure web sources
- IDE Support for easy API creation and test
- Easy deployment option
- Allows cloning an existing data source and modifying it to define new data

8.2 Key Features of the Solution

Happiest Minds, with a strong understanding of the market and the need of its customers, has developed the solution that blends very few distinctive feature sets. Some of the important features of the solution are:

- DSL for easy development of routes
- Intelligent Agent and Test Driven Development
- Security: Custom Token, Basic & OAuth
- Data Caching Strategy
- Identifies changes(website) and notifies stakeholders
- Scheduled Data Extraction
- Service Monitoring

8.3 Key Benefits of the Solution

Happiest Minds has a history of developing solutions keeping "Customer Happiness" foremost. It has been our endeavor to not just develop and deliver a solution to customer but to develop a solution that would benefit the customer in larger scale. Some of the benefits that customers can reap from Happiest Minds WSAPI solution are:

- Enable creation of a virtual API channel
- Reduce development time
- Easy and quick setup through IDE plugins
- Scheduled extraction
- Extensive monitoring
- Cloud enabled
- Increases efficiency and productivity
- Ease of maintenance
- Faster go-to-market

8.4 Target Customer Segments

- SME customers
    - Small ecommerce sites
    - Travel & Tourism portals
- Organizations with budget & IT resource constraints

Conclusion

Extracting data through scraping technology is a new evolving activity in the technology harvesting arena. Though many companies are still using manual process of extracting data but Web Scraping solutions will transform the traditional method of extracting data. With fast growth in this space the day is not very far when it will become a trend and majority of the organizations will realize the importance of scraping technology and how it significantly helps in staying ahead of the competition. With many players coming up in the market, Web Scraping solutions would sooner or later manage to completely eradicate the traditional method of scraping data. With Website Scraping API (WSAPI) solution from Happiest Minds, enterprise customers receive a solution that takes into account seamless extraction of data and provide desired results in less time.

References

1) **Advanced Web Data Extraction and Data Mining-** http://ficstar.com/wpcontent/uploads/resources/Ficstar-white-paper-062013.pdf
2) **The API market is taking a big shape –** http://siliconangle.com/blog/2010/12/02/the-api-market-is-taking-a-big-shape/
3) **IDC Vendor Profile- Ficstar: Simplifying Web Data Extraction –** http://ficstar.com/wp-content/uploads/resources/IDC_Vendor_Profile.pdf
4) **Five Questions to Ask When Evaluating Web Data Extraction Options by Vincent Sgro Founder and CTO, Connotate, Inc.-** http://www.connotate.com/uploads/Five-Questions-to-Ask.pdf
5) **From Top-Line Growth to Bottom-Line Profits: 10 Reasons to Use Automated Web Data Monitoring and Extraction -** http://www.connotate.com/uploads/Final-10-Reasons-White-pape08-07-5.pdf
6) **The API Transformation by John Tyrrell, Intel –** http://software.intel.com/sites/billboard/article/api-transformation
7) **Web Scraping Evolved: APIs for Turning Webpage Content into Valuable Data -** http://blog.programmableweb.com/2012/09/13/web-scraping-evolved-apis-for-turning-webpage-content-into-valuable-data/
8) http://en.wikipedia.org/wiki/Web_scraping
9) http://www.mozenda.com/web-scraping-software
10) http://www.iwebscraping.com/Benefits_%20webscreenscrapingcontentdataextraction.php
11) http://www.websitescraper.com/web-scraping-benefits.php

List of Figures

Acronyms

1) API - Application Programming Interface
2) HTML - Hyper Text Markup Language
3) WSAPI - Website Scraping API
4) HTTP - Hypertext Transfer Protocol
5) IDE - Integrated Development Environment

About Happiest Minds

Happiest Minds is a next generation ISO27001 Certified IT services company founded by Ashok Soota in 2011. The company is funded by Intel Capital, Canaan Partners and Ashok Soota.

Happiest Minds collaborates with clients to help them differentiate and win with a unique blend of solutions and services based on the core technology pillars of cloud computing, social computing, mobility and analytics. We offer customized, integrated services in the area of IT Services, Software Product Engineering, Infrastructure Management, Security, Independent Testing, and Technical Advisory Consulting

We combine an unparalleled experience, comprehensive capabilities in the following industries: Retail, Media, CPG, Manufacturing, Banking and Financial services, Travel and Hospitality and Hi-Tech with pragmatic, forward-thinking advisory capabilities for the world's top businesses, governments and organizations. Happiest Minds is privately held with headquarter in Bangalore, India and offices in the USA, UK, Singapore, Canada and Australia.

For further information, please contact contactus@happiestminds.com

Corporate Office
Happiest Minds Technologies Pvt. Ltd.
Block II, Velankani Tech Park
43 Electronics City
Hosur Road, Bangalore 560100, INDIA
Phone: +91 80 332 03333
Fax: +91 80 332 03000

United States
116 Village Boulevard, Suite 200
Princeton, New Jersey, 08540
Phone: +1 609 951 2296
2018 156th Avenue NE #224
Bellevue, WA 98007

United Kingdom
200 Brook Drive, Green Park, Reading
Berkshire, RG2 6UB
Phone: +44 11892 56072
Fax: + 44 11892 56073