MERIT

BLOGS

Web Scraping Done Right: Best Practices to ensure Ethical Data Collection & Web Scraping

Key Highlights

- The pandemic has accelerated digitisation, which has also resulted in a data surge online
- Using this data for analytics and decision making is a no-brainer

- While data mining and data collection are central to any analytics effort, the key is to make sure quality data is being collected
- One approach to increase the quality of data being used is web scraping since we mostly scrape the data we want from proven, well-known and

This website uses cookies to improve your experience.

Accept Reject Cookie Policy

The Covid-19 pandemic has shaken the world out of its complacency and catalysed change in many different areas. We now know that disasters can come in any form and can last for a really long time. Urgh!

Wave after wave of the pandemic has had a deep impact on the economy at large. Thanks to globalised supply chains, a lockdown in one part of the world can impact the operations of another company in some other geography. The daily routines of people and organisations have changed, for good or for worse. Of course, companies, governments, and entrepreneurs have responded with newer offerings. Some new trends that have emerged as a result are:

- Remote working even in traditional industries such as manufacturing
- Accelerated digital transformation of processes
- A boom in e-commerce and at-home service economy
- The resilience of supply chains
- Behavioral change in how people spend their time

A massive surge in online data generation

One McKinsey report titled 'The COVID-19 recovery will be digital' highlighted why most enterprises were digitising at least some parts of their business – primarily to protect employees and serve customers in new ways. This was further propelled by a skill shortage and the need for AI-based automation to maintain and improve business operations.

As a result, in just eight weeks of the pandemic's start, digital adoption leaped forward by about five years. For every business, be it manufacturers, banks, retail stores, healthcare service providers, and even schools – digital delivery of services became crucial.

Even though lockdowns have gradually lifted, customers have become

This website uses cookies to improve your experience.

Accept Reject Cookie Policy

data generated was 1.7 megabytes per second per person in 2020. Overall, internet users were generating about 2.5 quintillion bytes of data each day.

Get a Free Consultation

Leveraging the Data Surge

The good news, of course, is that now businesses have more data for running analytics and drawing insights for making informed decisions.

The data analytics market is expected to grow at a CAGR of 25.7% from USD 15.11 billion in 2021 to USD 74.99 billion in 2028.

At the same time, the estimated cost of poor data quality is expected to go up to $3.1 trillion yearly in the US alone. Needless to say, poor data quality can massively impact the quality of insights generated from data.

To improve the quality of data, merely aggregating it from different sources is not enough. It needs to be:

- Relevant
- Reliable
- Complete
- Timely
- Accurate

But given the volume of data and the heavy resource crunch, ensuring data quality is proving to be a challenge. The answer lies in automating data harnessing using what is called Web Scraping.

Web Scraping 101

This website uses cookies to improve your experience.

Accept Reject Cookie Policy

extraction, web harvesting, or screen scraping, can be used to look for and collect a specific type of data based on the specific need of an enterprise.

According to Techopedia, it is a **form of data mining** that is fast becoming a popular tool for collecting aggregated data such as weather reports, market pricing, auction details amongst others. The data thus collected is exported to MS-Excel, a database, or an API.

How Web Scrapers Work

On submitting the URLs from which the data is to be collected, the web scraper will load the entire HTML code. The entire website, with CSS and JavaScript elements, may be accessed if an advanced scarper tool is used.

Users can specify the data they need or let the scraper extract all data on the page before running the project. This data is then output in CSV format and in the case of advanced scrapers, other formats such as JSON can also be used to feed to an API.

Ethics of Web Scraping

Last but not least, there is one thing that MUST be followed. All your data scraping efforts must be ethical.

Here are few approaches to ensure the Web Scraping process is completely transparent and ethical:

- Use a Public API when available and avoid scraping all together if the data you're looking for is available through the API
- Pass your data through a user agent string to identify who you are
- Scrape data at a reasonablerate and throttle/control the number of requests per second. The website owner must not think it is a DDoS

This website uses cookies to improve your experience.

Accept Reject Cookie Policy

- Don't scrape private data – Look at the site's robots.txt and analytics needs to avoid scraping data from sensitive areas.
- Ideally, you must provide a user agent string, that gives the data owner a way to contact you if necessary
- Develop a formal Data Collection Policy

Developing a formal Data Collection Policy

It's important to develop a formal Data Collection Policy to guide developers and technology teams. This is crucial to ensure all developers abide by best practice.

Policy implementation should include regular audits on robots and their underlying code followed by updated briefings to the relevant team members. This practice is key to ensuring that ethical collection is kept centralised and consistent.

Merit Data & Technology: A Trusted Web Scraping & Data Mining Partner, with a deeply ethical approach

Though automation makes web scraping sound easy, it is not as straightforward. Some of the challenges include:

- The different formats and designs used by different websites requiring web scrapers with varying functionality and features
- The possibility of websites protecting data with captchas and other methods
- Ensuring ethical data collection by ensuring that the scraper selects only publicly available data as it is illegal to extract data not available in the public domain

This website uses cookies to improve your experience.

Accept Reject Cookie Policy

- **The Right Infrastructure:** Web scraping needs the right tools and skills to help you meet your business outcomes. While small projects may be manageable, for large data sets, customized scripts and software will be required to collect the right kind of data. An experienced team like the one from Merit will begin the process by understanding your needs, the purposes for which you need the data, and then create customised tools to deliver the right data in the format you need.

- **Scalability of Data Collection:**As your needs grow, you will need to scale up your data collection process as well. By outsourcing it to a reliable partner like Merit, you can keep your costs low but increase the value as well as scale up or down based on your business needs.

- **Data Quality Validation:** As mentioned earlier, web scraping does not overcome the quality issues of data. An experienced team of data scientists can ensure and validate data quality before it is used for analytics and decision-making.

- **Greater Focus on Core Functions:**While you leave data collection, data validation, and data processing to the experts, you can continue to focus on the core areas of your business and improve your team's productivity and efficiency using the data and analytics we provide.

Merit Data & Technology has been delivering data solutions to clients for over 15 years across a range of industries, from maritime to construction, fashion and E-commerce. The company has developed a number of automated data collection solutions, **in addition to machine learning tools** that help our clients transform raw data into usable and valuable intelligence.

Our Managing Director and CEO, Con Conlon is speaking at **OxyCon**, a two-day conference on the Future of Web Scraping. To register for this session where Con will be speaking with Alan O'Neil of The Data Works, book your place here.

Get a Free Consultation

This website uses cookies to improve your experience.

Accept Reject Cookie Policy

Related Case Studies

01 /

Sales & Marketing Data Analysis and Build for Increased Market...

A leading provider of insights, business intelligence, and worldwide B2B events organiser wanted to understand their...

Read More >

This website uses cookies to improve your experience.

Accept Reject Cookie Policy

Merit Data and Technology Limited is a direct subsidiary of Merit Group Plc.

Copyright © 2022 Merit Data and Technology Ltd, all rights reserved.

This website uses cookies to improve your experience.

Accept Reject Cookie Policy