THE MOZENDA CLOUD

COREY YOUNG

Copyright © Mozenda, Inc.

mozenda

THE MOZENDA CLOUD

COREY YOUNG

Today's web is becoming increasingly complicated as websites evolve to offer users a more interactive and fluid experience. This poses difficult challenges to traditional data extraction systems and scripts that cannot scale to meet business' needs, and which are limited in their ability to capture data from modern websites. At Mozenda, we solved many of these challenges by building innovative software running in a secure, scalable cloud system that eliminates the traditional hassles associated with web scraping.

Table of Contents

Copyright © Mozenda, Inc.

Benefits of Cloud

The Mozenda Cloud provides many of the typical advantages of cloud computing. We take care of the details so you don’t have to; so you can get back to solving your business challenges.

With a cloud solution, initial up-front capital expenditures are reduced or eliminated because there is no need to purchase and maintain hardware, no software to install and upgrade, and no data backups to perform. Security of your data, around-the-clock monitoring of the system, and problem resolution are taken care of for you.

Cloud solutions are flexible, allowing you to scale your costs with growth rather than buy everything up front before the scope of your project is fully understood. This allows you to increase your usage as your needs increase, or scale back as needed with little upfront commitment.

Cloud solutions offer the luxury of working from anywhere with increased collaboration because the system can be accessed at any time, from any location.

Responsible Web Scraping

Traditional methods of scraping used "headless" browsers and scripts that pummeled websites to get data as quickly as possible. This would cause significant spikes in traffic to the target website and increase the likelihood that the website would detect and block scraping. In an effort to thwart scraping, and spoof the scripts, some companies would change their websites to return inaccurate and unofficial data thereby damaging the integrity of the data.

Additionally, while larger websites with hundreds, or even thousands, of servers can handle an enormous amount of web traffic, many websites with only a handful of servers cannot. Scripts that bombard these smaller websites with impunity cause performance problems at best, or bring them to their knees at worst. This behavior is irresponsible, in our opinion.

At Mozenda, we scrape responsibly. This means we don’t overload websites that can’t handle the load and we traverse web pages the same way a user would, with a browser. At Mozenda, we study traffic patterns and rankings of almost every website our customers scrape. By understanding the traffic patterns and rankings of a website, we can better adjust the frequency with which we perform requests to a website. This means we are able to responsibly perform a larger number of concurrent requests to larger websites while at the same time performing far fewer concurrent requests against a small website. We can also adjust this as-needed, automatically, so our customers don’t have to.

When many users want to capture data from the same website, our proprietary request handling system intelligently brokers all requests to the website so that each Agent performs “their fair share” of requests while Mozenda maintains reasonable request levels based on the total traffic to the website.

Browser-Based Web Scraping

The most important way Mozenda plays nice with a website is to simply act like a human visitor to the website. Our system mimics human behavior as accurately as possible by rendering each web page in a browser and performing actions on the page like a human would. For example, rather than simply performing a request to the website to get a page that the user would get to by clicking a link with their mouse, our system actually clicks the link with a mouse. Not only that, our system even simulates mouse movements and keystrokes.

Scraping in this way has several significant benefits. First of all, it reduces the complexity of the scraping engine because, by performing the click like a user, we allow the web page to perform its activities, such as showing or hiding content, or executing code without interruption or special processing. In other words, the website behaves exactly as it was intended to behave.

Second, by scraping in this manner, we reduce the likelihood of the website deliberately providing inaccurate data or changing navigation paths because it has detected an automated script. No website wants to make it difficult for its users to get what they need and the same is true when you scrape responsibly, in the way it was intended and at a rate that is reasonable.

A third benefit to this method of scraping is that we can stop and restart scrapes without complications. Many websites maintain user state as the user navigates the website. Mozenda monitors and records the precise path an agent takes so we can follow the same path when the agent is restarted. This helps tremendously with error handling. When an agent runs into an unexpected error condition, we load the agent to the precise page that caused the error, providing the user with a visual error resolution system. Then, after modifying the agent to fix the error, the agent can be restarted where it left off.

Remaining Anonymous

Even though we scrape responsibly, some websites are particularly sensitive to any scraping. Our system knows which websites these are and automatically sends requests to these websites through anonymous proxy systems to reduce the likelihood of the scrape being blocked.

For customers who consider anonymity critical to their business and who view anonymity as an important competitive advantage, Mozenda has you covered. Our sophisticated proxy solution consists of large pools of IP addresses on disparate geographically diverse networks constantly available and frequently rotated to ensure freshness.

Geographic Location Simulation

Mozenda’s proxy system can simulate the geographical location of a website user when scraping a website. This is important for some websites that only produce correct content when they detect the user browsing their website from a specific geographical location. For example, web retailers often provide different pricing based on the geographical location of the customer. Other websites will change their currency or even price based on location. Additionally, some social media websites and blogs regulate content they display based on location.

Mozenda can simulate location for any country in the world. Even though all Mozenda scripts run in the Mozenda Cloud, users have the ability to route their traffic to appear as if it is coming from inside their own company, if needed.

Safeguarding Customer Data

Because data is at the core of our business, keeping our customer's data safe and secure is of the utmost importance to us. In fact, nothing is more important to us than the safety and security of our customer's data. We rely on the expertise of C7, a world class data center located in Utah, to help us in this effort. We co-locate the majority of our servers at C7.

C7 has strict security policies and technology in place to ensure our data and infrastructure are both safe and secure, including SSAE 16, PCI, and HIPAA compliance audits, and 2 factor, 5 layer authentication, which requires a hand scan and four digit pin to gain physical access to our servers. Additionally, C7 employs 24 x 7 on-site security, video cameras throughout their facilities, background checks, access logs, and so forth. Security and transparency are at the core of their business and their infrastructure shows their commitment to these core beliefs.

How We Restrict Access

For our part, we restrict direct physical and electronic access to our servers to four specific Mozenda Operations personnel tasked with overseeing all data center operations. Electronic access to servers is allowed only through a VPN connection using IPSec, an industry standard for ensuring private, secure communications over the Internet.

Sometimes Mozenda employees who are not part of the Operations team need to access customer accounts to provide support and perform periodic maintenance on behalf of the customer. In these cases, employees use the same SaaS web interface used by customers and are limited to access the accounts during specific business hours and from specific locations. Each access of and modification to a customer's account is logged by our system, including the name of the employee and the time of the access, as part of our effort to maintain accountability.

PayPal and PCI Compliance

We integrate with PayPal for payment processing and undergo a periodic audit for PCI compliance. As such, we do not store any credit card information on our systems and do not have access to this information once a payment has been posted through our system.

Your Own Database

We give each Mozenda customer account a physically separate database. This compartmentalizes customer data so it is separate and secure from the data of other customers. This also simplifies how our applications access the data by virtually eliminating risks associated with cross-mingling of data. Rest assured, your data is truly yours.

Redundancy and High Availability

C7 utilizes a N+1 redundant power delivery infrastructure, battery and generator backup, and redundant power legs to each server cabinet. C7 data centers are highly connected with multiple Tier 1 and Tier 2 on-net providers.

We use HP ProLiant rack mount and blade servers almost exclusively because of their excellent support for high availability and their accurate alerting capabilities. We virtualize almost all of our data center processes using Microsoft Hyper-V running on Windows Server 2012. This gives us the flexibility we need to appropriately manage resources, balance load, and increase capacity without the need to frequently procure additional hardware.

All of our storage is redundant using RAID levels 1+0 and 6. We perform nightly differential backups and weekly full backups of important customer data to multiple locations, including off-site (out-of-state) using a secure connection.

Monitoring

To be sure our data center hums along in tip-top shape, we utilize a variety of monitoring tools, such as FrameFlow, Librato, and Windows performance counters, along with proprietary monitoring applications. We receive real-time alerts from these monitoring systems when specific thresholds are exceeded and take immediate action to resolve the issues before they turn into problems.

System Updates

We periodically maintain our systems between 11 PM and 3 AM, Eastern time, usually on a weekend. We do this to install security patches and update our hardware and software. While this is infrequent, it can mean that some services are not available during this time. We designed our system so that this should have minimal impact on most customers. Agents will pause while the system maintenance takes place and will resume once maintenance is complete.

In most cases, we will notify customers two weeks in advance, but sometimes security patches or other upgrades may be urgent and we may not always be able to give advance notice.

When we schedule maintenance, we will send an email to customers notifying them of the maintenance, including when it will take place, how long it will take, and what actions will be taken. Additionally, we will post this information to Status.Mozenda.com and provide periodic updates regarding progress during the maintenance window.

Conclusion

In this article, we discussed some features of the Mozenda Cloud that solve important historically difficult challenges to scraping websites. If you would like to know more, please give us a call at 801-995-4550 and we would be happy to answer your questions.