Companies are increasingly using web scraping to support a wide range of initiatives including sentiment analysis, price monitoring, market research, and financial assessments. Scrapinghub has been meeting this demand for more than a decade with an array of tools, services, and open source technologies. Each day thousands of companies and more than a million developers rely on Scrapinghub tools and services to extract the data they need from the web. And each month those tools and services are used to crawl more than eight billion pages.

To strengthen its position as a market leader, Scrapinghub recently launched a new product, AutoExtract, that provides customers with AI-enabled, automated web data extraction at scale. Scrapinghub built AutoExtract on Confluent Cloud running on Google Cloud Platform (GCP), with an Apache Kafka®-based, event-streaming backbone for its service architecture. These technologies were chosen to shorten time to market, and to ensure reliability and scalability.

"With AutoExtract we’re using AI and machine learning to take web scraping to the next level," says Ian Duffy, DevOps Engineer at Scrapinghub. “A key advantage of Confluent Cloud in delivering AutoExtract is time to market. We didn’t have to set up a Kafka cluster ourselves or wait for our infrastructure team to do it for us. With Confluent Cloud we quickly had a state-of-the-art Kafka cluster up and running perfectly. We don’t have any worries about under-replicated partitions or ongoing maintenance. And if we run into any issues, we have experts at Confluent to help us look into them and resolve them. That puts us in a great position as a team and as a company.”

As a fully remote company with its employees working from locations all over the world, Scrapinghub depends on the ability of its teams to operate autonomously, rather than getting bogged down waiting for groups in other time zones. Confluent Cloud enabled the AutoExtract team to deploy and manage their own
“A key advantage of Confluent Cloud in delivering AutoExtract is time to market. We didn’t have to set up a Kafka cluster ourselves or wait for our infrastructure team to do it for us. With Confluent Cloud we quickly had a state-of-the-art Kafka cluster up and running perfectly smoothly. And if we run into any issues, we have experts at Confluent to help us look into them and resolve them. That puts us in a great position as a team and as a company.”

Ian Duffy, DevOps Engineer at Scrapinghub

Kafka cluster dedicated solely to their product offering, which accelerated and streamlined the rollout. “A big part of Scrapinghub’s growth strategy is giving teams autonomy and letting them do what they need to do,” says Duffy. “Confluent Cloud was a great fit for this strategy because it let us set up the AutoExtract infrastructure so that we had complete control over it, with clear separation from other services and no need to wait for others to provision resources for us.”

**BUSINESS RESULTS**

**Deployment time halved.** “When we started our move to the cloud we had the whole application running on Confluent Cloud and Google Compute Engine within a week,” says Duffy. “With Confluent Cloud we had it nailed in that week; without it we would have needed at least twice as long and it would have been a rush job.”

**Initial setup completed in minutes.** “Setting up Confluent Cloud was simple. We just went to the website, signed up, provided payment, and clicked ‘Create Cluster.’ Fifteen minutes later, if even that, we had credentials and we were ready to go,” says Duffy. “And, we saw the same interface as with our initial on-prem Kafka deployment, so almost nothing had to change but a few configuration parameters.”

**100% uptime post-launch.** “We’ve had no outages whatsoever since our launch,” Duffy says. “In fact, moving to Confluent Cloud helped us identify incorrect configurations in our existing Kafka deployment, so we could make the necessary changes to increase reliability and availability there as well.”

**Latencies minimized with no cloud vendor lock-in.** “A major benefit of using Confluent Cloud with GCP is that we have low latencies because our Kafka cluster is in the same availability zone as the rest of our infrastructure,” says Duffy. “Plus, Confluent Cloud offers that across the three major cloud providers, so if we make a change in the future, we’ll still have those low latencies with no problems.”

**TECHNICAL SOLUTION**

For development and testing, the AutoExtract team initially used a Kafka cluster deployed on servers run by Scrapinghub’s data center operations provider. Kafka provides the communications infrastructure for the many services that make up the full, compute-intensive AutoExtract data pipeline. As the
beta program wrapped up, the team began to look for a platform that would enable AutoExtract to scale horizontally to handle high traffic loads.

With a relatively small team, Duffy and his colleagues wanted to minimize ongoing maintenance responsibilities and had already begun taking advantage of managed services, such as Google Kubernetes Engine and Google Cloud SQL to do so. "Confluent Cloud offered cloud vendor independence, a more flexible pricing model, and support from experts in the Kafka space. It looked like the perfect solution, and so far it has been," says Duffy.

The Kafka producers in AutoExtract are web crawlers, and the consumers include AI and machine learning models that extract webpage content into a structured form, which is delivered back to the customer via FTP, AWS S3, or Google Cloud Storage, among other channels. Still in its early days, the service is handling around 12 requests per second. "As consumer demand grows, we expect to be handling 100 requests per second or more," says Duffy. "That's why we wanted Confluent Cloud – because it enables us to horizontally scale all of our components, without having to worry about Kafka maintenance. All we have to worry about is the software we develop ourselves, and we know how that runs very well."

Scrapinghub is currently working on plans to use Kafka Connect to stream logging data into BigQuery to enable real-time analyses and graphs of usage metrics.

“That’s why we wanted Confluent Cloud – because it enables us to horizontally scale all of our components, without having to worry about Kafka maintenance. All we have to worry about is the software we develop ourselves, and we know how that runs very well.”

Ian Duffy, DevOps Engineer at Scrapinghub

Learn More About Scrapinghub
scrapinghub.com

Learn More About Confluent
confluent.io