KESQUE whatify

Running Hundreds of ML Jobs in Parallel With Ease: How Whatify Simplified Its ML Pipeline Using Kesque

The Challenge

Whatify (formerly Firefly.ai) provides powerful business predictions and recommendations using a next-gen AI engine. They deal with many long-running ML jobs, some taking several hours to complete. They wanted a work queue to manage these jobs. After surveying available solutions, they found that many of them made it difficult to send jobs to a changing number of workers or didn’t even support multiple workers without some complex combination of services. Expensive ML-capable worker machines could sit idle when there was a job in the queue because the messaging system needed to be reconfigured before it could send to an additional worker.

In addition, like most other companies, Whatify has other messaging needs. They have event-driven applications using asynchronous messaging between their microservices. They wanted a single message solution that could handle their long-running ML jobs as well as processing events with low latency.

Whatify didn’t want to get locked into a cloud-vendor specific solution. And being a customer focused company, they didn’t want to be bogged down managing open source software, preferring to focus their time on providing the best AI-powered predictions and recommendations for their customers.

Highlights

- Whatify generates predictions and recommended actions using next-gen AI technology. They needed a work queue for their long running ML jobs (up to several hours each)
- They wanted to be able to seamlessly scale in and out the number of workers processing their ML jobs
- Using a Kesque Dedicated Plan, they were able to quickly build a work queue for their long-running ML jobs using a Pulsar shared subscription that will scale up to hundreds of workers
- Whatify also uses their Kesque plan to process events with low latency, avoiding cloud provider lock-in since Kesque is powered by open-source Apache Pulsar

> **“As a big fan of using managed services, using Kesque's cloud messaging service was the obvious choice. It allowed us to focus on our core product and reduced the overhead of maintaining the infrastructure ourselves. The Kesque team is very knowledgeable and responsive, and we've found Kesque's web application very useful as well.”**
>
> Gilad Ivry - Chief Architect @ Whatify

The Solution

Apache Pulsar provides a powerful work queue with its shared subscriptions. Using a shared subscription, new workers can be dynamically added to the work queue and jobs will automatically get distributed to the new workers. And if a worker goes away, the shared subscription sends the job to the next available worker. Pulsar supports multiple subscription types, not just shared. Using different subscription types and combinations of subscriptions, it is easy to support other patterns, such as pub-sub.

The power and flexibility of Apache Pulsar subscriptions were exactly what Whatify was looking for, but operating a powerful, but complex distributed messaging system like Apache Pulsar was going to slow them down. Then they discovered Kesque. With the Kesque Dedicated plan, they unlocked the power of Pulsar's shared subscription, supporting up to 500 workers to their ML work queue. And with Kesque's fully integrated dashboard, it only took a few minutes to get started. Since Kesque supports all Pulsar subscription types, they were also able to support their event-driven applications easily.

In addition, Whatify uses the Prometheus metrics add-on which allows them to gather detailed metrics about their Kesque service, including how many jobs are waiting in the work queue. Feeding that data into an auto-scaling group in their cloud provider, they are able to dynamically adjust the number of ML worker nodes based on how many jobs are waiting to be processed. If the backlog of jobs is low, they can scale down workers and save money, and if the backlog grows they can scale up more workers to make sure they can meet their service-level agreements.

Kesque was the perfect fit for Whatify: easily handling the ML jobs with support for auto scaling, meeting other messaging requirements is a single solution, avoiding cloud provider lock-in, and freeing up their team to do what they do best--providing next-gen AI powered predictions and recommendations.

Contact

sales@kesque.com

For free account or schedule demo
please visit: kesque.com

Ottawa, Canada

Follow us

KESQUE

Formerly known as Kafkaesque