RECURSIONPHARMA.COM

CONFLUENT

Recursion Accelerates Drug Discovery with New Data Pipelines Based on Confluent Cloud

RECURSION

Headquarters
Salt Lake City, UT

Industry
Pharmaceuticals

Challenge
Accelerate the drug discovery process by streamlining the analysis of biological image data

Solution
Use Confluent Cloud and Kafka to create a scalable, highly reliable data pipeline based on real-time event streaming

Results

- Drug discovery pipeline stages accelerated
- Flexible, highly available data pipeline established
- Stable operation in production since launch

Advances in machine learning, image processing and data science have opened unprecedented opportunities to accelerate the discovery of new drugs that hold the potential to dramatically improve and even save millions of lives. Recursion Pharmaceuticals has seized these opportunities by building a massively parallel system that combines experimental biology, artificial intelligence, automation and real-time event streaming. This system has been used to process over three petabytes of biological image data that Recursion has generated on its automated robotics platform. It applies machine learning to help the company understand mechanisms of action, spot possible toxicities and discover and develop drugs for treating inflammation, rare and infectious diseases, immune disorders and other conditions.

The core data pipelines at Recursion start with fluorescence microscopy images captured during experiments with cells and reagents. As Recursion scaled their high-throughput screening (HTS) lab, the volume of experimental data increased significantly and bottlenecks in the batch system became apparent. Processing the data from a single experiment – potentially more than 8 terabytes – did not begin until all images were available. This introduced delays and made it impossible for the laboratory to obtain real-time quality control metrics on the images. The company evaluated several open-source options, including Apache Storm and Apache Spark, but had concerns about operational complexity and migrating existing microservice logic.

Recursion decided to use event streaming with Confluent Cloud and Apache Kafka to minimize administrative overhead, enable faster iterations on experimental results and simplify migration by reusing existing microservices. "We decided to use Kafka Streams and Confluent Cloud to orchestrate all of our existing services. This was an attractive option because it was low overhead. Also, with both Spark and Storm we would have had to still introduce a queue system, but by adopting Confluent Cloud we got a queue with a great persistent log model and a stream processor." says Ben Mabey, VP of Engineering at Recursion. "The scale and robustness of the system we built with Confluent Cloud have played a key role in accelerating our success in our mission of discovering new treatments and has helped us bring new treatments to human clinical trials".

"The scale and robustness of the system we built with Confluent Cloud have played a key role in accelerating our success in our mission of discovering new treatments and has helped us bring new treatments to human clinical trials."

Ben Mabey, VP of Engineering

Business Results

Drug discovery pipeline stages accelerated. Recursion has already made significant strides in accelerating drug discovery, with more than 30 disease models in discovery, another nine in preclinical development, and two in clinical trials. With the old batch system used for processing experiments, extracting features would take one hour for small experiments. With Confluent Cloud and the new streaming approach, the company has built a platform that makes it possible to screen much larger experiments with thousands of compounds against hundreds of disease models in minutes, and less expensively than alternative discovery approaches.

Flexible, highly available data pipeline established. "Lack of visibility and robustness made our previous pipeline difficult and costly to operate," says Mabey. "By building on top of Confluent Cloud and Kafka Streams, we created a flexible, highly available and robust pipeline that provided a clear migration path leveraging our existing microservices."

Stable operation in production since launch. "The system we architected with Confluent Cloud has been in production for over a year and it has been stable since we introduced it," says Mabey. "Having a central orchestrator that is robust to node failures has stabilized our operations substantially. Throughout it all, we have never had to deal with any sort of deployment or operational headache. Early on we encountered a minor configuration issue and Confluent support alerted us to the issue and helped us resolve it."

Technical Solution

A core element of the data pipeline that Recursion built with Confluent Cloud is Dagger, an event-driven workflow and orchestration library for Kafka Streams that enables Recursion engineers to orchestrate services by defining workloads as high-level data structures. Dagger combines Kafka topics and schemas with external tasks for actions that are completed outside of the Streams applications.

The workflow that Recursion orchestrated using Confluent Cloud and Dagger starts with microscopy images being captured in the lab with image events being published to Kafka. These image events causes the images to be uploaded to the Google Cloud Platform and results in further events and tasks being created. Worker nodes on Google Kubernetes Engine are autoscoled based on the number of tasks that need to be processed. A worker is matched with the next task in the queue – for example, processing the next image, generating metrics and extracting features that then flow downstream to Recursion's convolutional networks and machine learning models. As images are processed with this event streaming architecture, quality control issues with the image are detected in real time and cumulative metrics are aggregated for reports.

For Recursion, the key objectives transitioning to an event-streaming platform for a mission critical business workflow included minimizing operational overhead while ensuring operational stability. Mabey explains that Confluent Cloud and Kafka have been instrumental in

meeting these objectives. "Kafka provides the backbone for how all of our systems communicate with each other reliably and gives our engineers the flexibility to think and design systems at a higher level," Mabey says. "And Confluent Cloud makes all of that so much easier because we don't have to manage Zookeeper or worry about operational aspects of our Kafka deployment. Confluent Cloud is crucial to the cadence and robust workflow we have established."

> *"The scale and robustness of the system we built with Confluent Cloud have played a key role in accelerating our success in our mission of discovering new treatments and has helped us bring new treatments to human clinical trials."*
>
> *Ben Mabey, VP of Engineering*

Learn More About Recursion Pharmaceuticals

www.recursionpharma.com