How Atomwise Accelerated and Innovated Drug Discovery and Time to Market with Weka and AWS

aws partner network

Executive Summary

Atomwise is a pharmaceutical research company that uses artificial intelligence (AI) for broad life sciences use cases such as oncology, infectious diseases, neurology, biotechnology, cardiology, as well as many rare diseases.
Like many companies in the healthcare and life sciences sector, Atomwise ingests petabytes or more of unstructured data regularly. To develop new drugs, the company develops 3D, structure-based models of molecules for deep learning applications and builds training models for millions of structures. Getting this unstructured data to the AI training model was a bottleneck in the workflow and the infrastructure was limited by the I/O. This resulted in longer AI model training times and longer compute run time.

Amazon Web Services (AWS) and Weka helped Atomwise lower costs and enable faster training times by providing the flexibility of spot instances and leveraging Amazon Simple Storage Service (Amazon S3) for snapshots. Atomwise achieved timely access to the data sets supporting multiple concurrent projects, enabling the company to accurately train the models with the right data and iterate faster.

Eliminating Constraints in High Performance Workloads

Atomwise has discovered a number of small molecule hits for many undruggable targets. Such discoveries begin by combining structure-based drug design with deep learning. Atomwise then develops AI models to help create better medicines faster—but increasingly, data and bursty workloads require massive compute resources to process the volume and variety of data needed to develop the AI models.

Atomwise applies convolutional neural networks for molecular recognition—similar to neural networks for image recognition—which looks at atoms and groups them based on function. This is a global model that can be applied to any target. Using this model creates many data points to explain how well the compound binds to given proteins. It can typically result in 4,000 protein targets, three million compounds, and 15 million overall experimental measurements. The training models rely on millions of structures for Atomwise to process. Before 2017, Atomwise's process was sufficient because data volume was low, and only two developers were on staff.
By mid-2019, data volumes reached 30 million files and required 12-plus developers. Atomwise engaged Weka to help lower the barrier to training and testing on custom data sets, fulfilling Atomwise's objective to offer a platform that enables faster research and discovery.

Atomwise

ABOUT ATOMWISE

Atomwise uses deep learning for structure-based drug discovery and has developed a pipeline of small-molecule drug candidates advancing into preclinical studies.

AtomNet® technology has been used to unlock undruggable targets, and Atomwise is tackling over 600 unique disease targets across 775 collaborations spanning more than 250 partners around the world.

Weka Helps Atomwise Eliminate Performance Constraints

With such high-volume data processing and storage needs, Atomwise needed a solution that would be scalable and sustainable. “We wanted to train a model on the 30 million files we had, but the models are fairly large, with 30-50 epochs, a timeline of up to four days, and a lot of random-access-file lookups. GPUs are quite fast and hungry for data—you want to feed them as much data as you can. Our bottleneck came in the I/O to the file system: if you have a faster file system and faster I/O, you get faster training times. That’s the problem we needed to solve,” said Jon Sorenson, PhD, VP of Technology Development, Atomwise.

Our bottleneck came in the I/O to the file system: if you have a faster file system and faster I/O, you get faster training times. That’s the problem we needed to solve.

—Jon Sorenson, PhD, VP of Technology Development, Atomwise

To solve Atomwise’s challenge of still-sluggish data on existing systems and needing to get data into AI models easier, Weka implemented its clusters to support Atomwise’s Amazon Elastic Kubernetes Service (Amazon EKS) architecture. The Weka clusters are made up of eight i3en.xlarge Amazon Elastic Compute Cloud (Amazon EC2) instances. The company added the flexibility to spin up zero to 10,000 compute instances per day if needed.

Weka helped Atomwise solve the challenge of integrating the two systems together: Atomwise containers (which run on Docker) and the Weka system, which accessed the Atomwise data to feed it into the AI models.

Weka's ability to leverage Amazon EC2 NVMe instances as well as the Amazon S3 buckets as part of the same global namespace for Atomwise brought in the economics and scalability needed to the pipelines. Weka organized the Amazon EKS service for scheduling on-demand jobs, Kubernetes orchestration, as well as storage persistence.

“Atomwise’s Atomnet platform is a transformative solution, leveraging AI, GPUs, AWS, and Weka Data Platform for drug discovery. Weka is honored to have Atomwise as a customer as well as a technology partner, where AI is used for precision medicine, research, and drug discovery for life-threatening diseases like cancer and Sars-COV-2019,” said Shailesh Manjrekar, Head of AI and Strategic Alliances, Weka.

Atomwise’s Atomnet platform is a transformative solution, leveraging AI, GPUs, AWS, and Weka Data Platform for drug discovery.

—Shailesh Manjrekar, Head of AI and Strategic Alliances, Weka

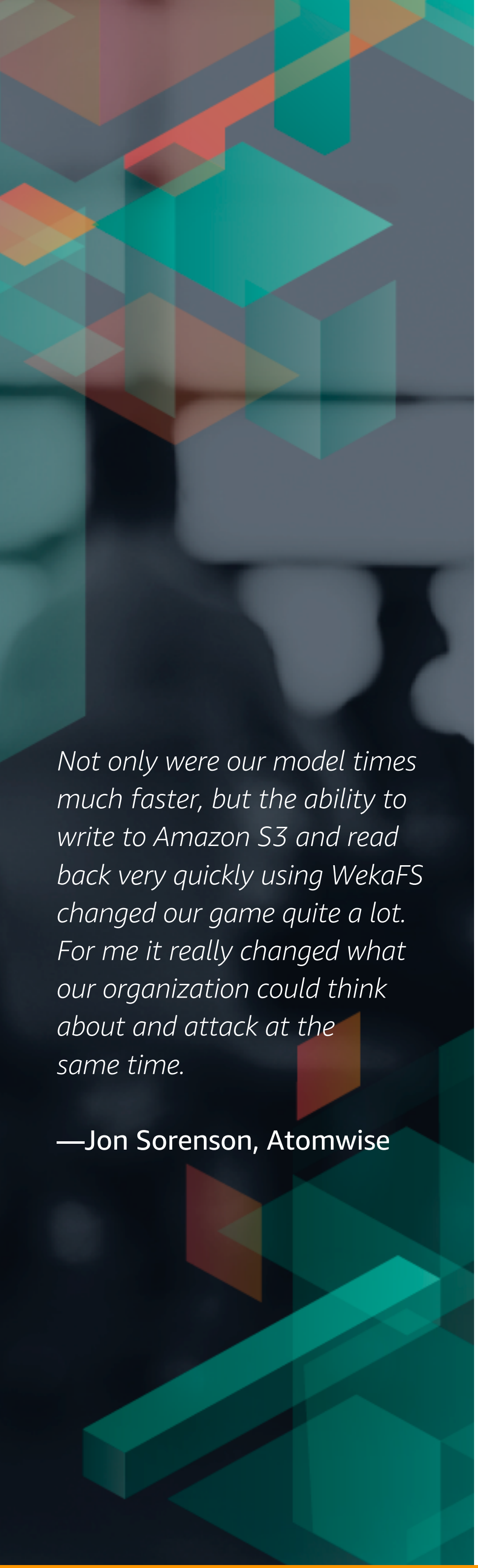
Fast, Scalable I/O and the Drugs of Tomorrow for Atomwise

Atomwise’s GPU-based data pipelines had unique challenges with an immense need for on-demand elastic computing. For its storage I/O, the company must import data, clean the data, generate descriptors, and package the data for use before the neural networks can be trained. Each of these stages have varied storage requirements that typically would result in storage silos. Weka's ability to handle mixed workloads allowed them to uniquely meet Atomwise's varied storage I/O requirements, resulting in faster experimentation and insights.

After a quick set up to connect a server and sync the existing Amazon S3 data, there was successful I/O performance for shared file access. Creating and copying 1GB files saw a 40x speed improvement, from 332 to 8.2 milliseconds (ms). Small file access saw the most improvement. Atomwise accesses thousands of files daily, so cumulatively these speeds saved a lot of time: leveraging the in-place Amazon Elastic File System (Amazon EFS) structure, Weka sped up access to small files by 3x, or 4.8 ms. The impact to its customers was that the model training times dropped by up to 2x.

Weka's performance capability allowed Atomwise to actually save money by moving to a higher-performing—and more expensive per instance—GPU model on AWS. The greater efficiency and performance utilization from the GPUs allow Atomwise to complete models much faster.

“Not only were our model times much faster, but the ability to write to Amazon S3 and read back very quickly using WekaFS changed our game quite a lot. We could now consider experiments that earlier—because of all these headaches—might take us three months to figure out how to run. Now we can do this exact same experiment in less than a week. For me it really changed what our organization could think about and attack at the same time,” said Sorenson.



> *Not only were our model times much faster, but the ability to write to Amazon S3 and read back very quickly using WekaFS changed our game quite a lot. For me it really changed what our organization could think about and attack at the same time.*
>
> —Jon Sorenson, Atomwise

ABOUT WEKA

Weka's data platform was built to address the storage challenges posed by modern applications that leverage cloud and GPU compute. Built on WekaFS, the Weka Data platform offers the simplicity of NAS, the performance of SAN or DAS, and the scale of object storage. No more compromises between simplicity, speed, or scale. With its unique software-defined architecture, customers can run on-premises, natively on the cloud, orchestrate their data effortlessly between the two. The ability to mix and match Flash and Disk offers customers the best economics. For more information, go to weka.io

WEKA