WEKA

SOLUTION BRIEF

# Level Up Your Generative AI Model Development

### Challenges

- Storing the large amounts of data needed to train generative AI models in the cloud can be expensive and time-consuming.
- Managing storage often requires expertise in various technologies, such as file systems, databases, and networking.
- As data sets scale, managing storage becomes more challenging to ensure performance and reliability are maintained

### Solution

- The WEKA® Data Platform takes the pain of managing the data sets in the cloud, helps you control your cloud costs, and helps you take full advantage of the power of your cloud.

### Benefits

- Avoid complex investments in infrastructure and get your development teams productive in the shortest amount of time.
- Deliver unbeatable file storage performance for your most demanding model development and training supporting high I/O, low latency, small files, and mixed workloads with zero tuning.

Are you looking to take your generative AI models to the next level? The key to unlocking their full potential lies in making efficient use of high-quality, diverse training data. The large amount of data to train models effectively, as well as acquiring, storing, and processing this data can be a challenge. The cost of storage and the time spent managing it are significant factors as the size of the data set grows, and the complexity of your models increases. As a result, the need to be storage experts and its cost can be a significant drag on startups focused on developing innovative solutions, enabling personalized experiences, improving automation and efficiency, and enhancing entertainment and gaming.

## The Cloud Data Challenges

However, there are several strategies that you can use to manage the data storage challenges for generative AI. For example, startups can use cloud solutions that offer flexible pricing models based on usage, or they can use compression techniques to reduce the size of the data set. But even in the cloud managing storage can be a complex and challenging task, requiring expertise in various technologies and disciplines. Key challenges are:

- **Complexity:** Storage management can be complex, especially when dealing with large and distributed data sets. Managing storage often requires expertise in various technologies, such as file systems, databases, and networking.
- **Cost:** Training a deep learning model with millions of parameters on a dataset of several terabytes could require hundreds or even thousands of GPU hours, which can translate to significant computational costs. The storage requirements could be even higher if you are dealing with multimedia data such as images or videos.
- **Pipleine efficiency:** data piplelines are often bottlenecked by storage systems that can't keep expensive, data hungry GPUs fully utilized. In this increasingly competitive landscape these innefieicencie can have a significant impact on time to market.
- **Scalability:** As data sets grow larger, managing storage becomes more challenging. Scaling storage requires careful planning and design to ensure that the system can accommodate increasing amounts of data while maintaining performance and reliability.
- **Performance:** Only a few storage offerings in the cloud offer the combination of extremely high performance and massive scalability your AI models need and oftentimes starting with the native files offerings of a cloud provider only lead to hitting a performance wall as your models scale.

For generative AI startups, managing cloud storage can be a significant hurdle to overcome. Storing the large amounts of data needed to train generative AI models in the cloud can be expensive and time-consuming. As a fast-moving startup, you need an "easy" button for cloud storage that simplifies storage management and streamlines workflows. With an easy-to-use cloud data platform, you can eliminate the complexity of storage management to save time, reduce costs, and improve productivity. This allows you to focus on developing innovative generative AI models instead of worrying about storage management.

## WEKA for Generative AI

The WEKA® Data Platform provides the fastest, most scalable file system for generative AI, delivering the performance developers expect and the scalability and simplicity the cloud promises - for all stages of model development on any cloud. WEKA lets you avoid complex investments in infrastructure and gets your teams productive in the shortest amount of time. Whether you're developing next-generation assistants, virtual reality applications, video games, or advertising campaigns, the WEKA Data Platform will simplify and accelerate your cloud data infrastructure so you can focus on creating models for more realistic and engaging content.

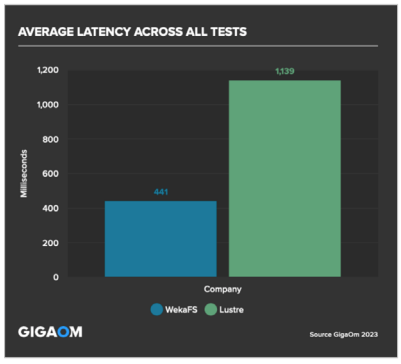
**Make data storage a development enabler, not a tax:** WEKA takes the pain of managing the data sets you need in the cloud, making it easy to have an entirely usable copy of any quantity of data in the cloud, so you can start training on that data without building out new infrastructure. And because a single WEKA deployment can support every step in the AI pipeline, developers spend less time waiting on storage and more time focused on model development.

**Optimize cloud storage costs:** WEKA helps you control your cloud cost by letting you use elastic compute resources to build or refine models in the cloud as needed. And then release—and stop paying for—the compute and storage resources when the development is complete.

**Maximize the use of your GPUs:** WEKA helps you take full advantage of the power of your cloud compute accelerators by keeping them fully utilized. It unlocks the full power of your GPUs to accelerate every step of your training lifecycle for the maximum acceleration of model development. WEKA eliminates data stalls and feeds GPUs with enough data to make yuor pipelines more effective and speeding up time to insight.

**Simplify your data footprint:** WEKA further helps you control your costs by optimizing resources and eliminates the need for multiple copies of your data across high performance and capacity tiers and seamlessly enables long-term retention either in the cloud or on-premises.

**Deliver the performance your developers and researchers need:** WEKA can accelerate and simplify the use of training data in a variety of formats, including images, video, text, and audio to deliver unbeatable file storage performance at any scale for your most demanding model development and training supporting high I/O, low latency, small files, and mixed workloads with zero tuning.




**GIGAOM** benchmarked the usability, effort, and performance of the WEKA Data Platform against Amazon FSx for Lustre on AWS. In this hands-on benchmark, they found that WEKA provided comparable or superior usability and outperformed FSx for Lustre at similar capacities by up to 300% or more. On some tests, WekaFS IO latency was less than 30% that of FSx for Lustre. Their usability tests also found WEKA to be a mature and easily found that WEKA outperformed FSx by 3x or more at similar capacities.

### For More Information or to Arrange a Free Trial

Visit us at https://www.weka.io/get-started or email us at info@weka.io.





WKA352-01 04/2023