WEKA Data Platform for AWS

Cut Your AWS Data Costs in Half
Accelerate Time to Results by 10X

Introduction

A new wave of workloads is moving to AWS for machine learning, high-performance computing, and various industry-specific applications. Enabled by innovation in high-density compute, fast networks, and GPU acceleration, builders are creating new ways to get their work done. Creative studios are building their entire content production pipeline on elastic cloud resources, but artists require a "local feel" when editing video files and rendering at 120 frames per second. Pharma researchers are using new tools like high-resolution cryogenic electron microscopy to analyze protein structures to discover new viral treatments. Still, they need the ability to rapidly analyze hundreds of thousands of individual files at petabyte scale. Generative AI developers training and tuning their large language models need high-speed access to the millions of small files that make up the massive data sets used in training.

A recent study from Enterprise Strategy Group revealed that "performance requirements could not be met" is now the #1 reason holding cloud migrations back (for the first time ahead of "implementing security measures"), with "too costly to migrate" tied for the #2 spot! Most technology leaders are realizing that legacy approaches to data are holding back some of their most strategic projects. In one recent global survey of AI adoption, leaders cited "data management" (32% of respondents) as the biggest inhibitor to AI success.

Reduce Costs and Accelerate AWS Workloads

The WEKA Data Platform delivers the performance and scalability needed to run performance-intensive applications in AWS affordably. WEKA provides a software solution purpose-built for high-performance and AI workloads running in AWS. Used by leaders across every industry vertical, the solution is helping organizations accelerate AI model training, tuning, and inference while reducing costs. WEKA software eliminates the need to over-provision resources in AWS to meet demanding performance requirements. It removes the need to manage multiple copies of data across a pipeline, simplifying and accelerating data pipelines while dramatically reducing costs. Customers who use WEKA for AI, HPC, and other performance-intensive workloads in the cloud reduce infrastructure costs by 50%, accelerate data-intensive workloads by 65%, and accelerate data pipelines by 90%, leading to increased revenue by as much as 20%.

High-Performance Data That Also Saves Money

In AWS, the WEKA Data Platform runs in the customer's tenant (usually deployed as a subnet within their VPC), utilizing a cluster of Amazon EC2 I3en instances with local NVMe storage to form a high-performance storage layer. The single WEKA namespace extends to an Amazon S3 bucket to add massive capacity at an affordable cost. The entire data set is available to the applications without the need to move or copy data. Applications residing either in traditional EC2 instances or containerized in Amazon Elastic Kubernetes Service (EKS) can access data in the WEKA namespace through multiple storage protocols – S3, NFS, SMB, POSIX, and CSI. WEKA stores hundreds of petabytes of data in S3 object storage while keeping the most recent version of the working set of data needed for data-hungry applications in NVMe storage. It leverages tiering technology to efficiently and optimally place data in NVMe storage to drive high performance when applications need it.

The Data Platform that Scales up…and Down

WEKA auto scaling provides customers with the ability to add scale performance of the WEKA system independently from capacity. WEKA software leverages EC2 Autoscaling to add new I3en instances to the cluster as needed to deliver near-linear increases in performance. Autoscaling also enables customers to scale the cluster back down (or completely away) when projects are not running. This ensures customers don't have to over-provision storage resources to meet a performance requirement, and they don't pay for resources they don't use. Scaling the single WEKA namespace capacity through Amazon S3 is simple and scales to hundreds of Petabytes or even Exabytes of capacity.

Fast, Easy Deployment Automation

WEKA deployments are automated using AWS CloudFormation or Hashicorp Terraform scripting. CloudFormation templates create a private subnet with the customer's VPC, set-up EC2 Autoscaling groups, register and start up a predefined set of EC2 instances in the cluster, create the requisite S3 bucket, and deploy WEKA software to the cluster.

Accelerate Hybrid Cloud Workflows

Customers use WEKA for a wide variety of hybrid cloud use cases, including cloud bursting, cloud archive and backup, disaster recovery, and migrations. The same WEKA Data Platform software that is available for AWS also runs in customer's data centers and in any cloud, including Microsoft Azure, Google Cloud, and Oracle Cloud. WEKA Snap to Object enables simplified data operations by eliminating the need to create, move, and maintain multiple copies of data across data silos located on-premises and in the cloud. Snap-to-Object creates a self-describing snapshot of the entire WEKA environment - including data and file system metadata and replicates that snapshot to another object store either in a customer's data center, in AWS or any cloud. The WEKA Zero Copy Architecture ensures customers need only maintain a single copy of data to deliver their performance requirements for every stage in their data pipeline. With WEKA, customers can maintain a single copy of data, maintain consistency across any location, and ensure high-performance, and low-latency access to data from anywhere in the world.

Maximum Flexibility, Minimum Hassle

The WEKA Data Platform can be deployed in multiple configurations to enable maximum business flexibility. WEKA software can be deployed on a set of dedicated EC2 instances within the customer's VPC, enabling maximum performance, while reducing costs. For applications like large language model training that rely on large GPU-accelerated clusters running on P4d instances, WEKA can run in converged mode on AWS for maximum resource efficiency. WEKA Converged Mode for AWS runs on the application instances themselves, relying on job scheduling services like Slurm to allocate and manage resource use between the application and the WEKA environment. Using WEKA Converged Mode for AWS, customers can eliminate the need for a dedicated storage system entirely, reducing costs even further and maximizing resource utilization of GPU-accelerated infrastructures.

The WEKA Data Platform for AWS Workloads

Build the Next Great AI Application

AI applications of every sort - generative AI, natural language processing, speech recognition, autonomous vehicles, and more - require massive data sets to train their models. However, the focus of most AI organizations has been on building large GPU fleets with high-speed 400GbE network interconnects with low latency. Legacy data infrastructures that proliferate most AI deployments today - either in the cloud or in data centers cause throughput bottlenecks and IO blender effects, slowing down the entire AI pipeline. It's become a well-known challenge that most GPU-accelerated deployments driving AI model training are starved for data, resulting in GPUs sitting idle up to 75% of the time, waiting on data operations. It's no wonder many AI organizations consider data infrastructure to be a tax.

Organizations like Midjourney, Stability.ai, WeRide, and the Center for AI Safety rely on WEKA to accelerate LMM model training and tuning because WEKA provides the fastest, most scalable data platform for AI and HPC applications in the cloud. These customers rely on WEKA software to deliver the performance developers expect and the scalability and simplicity the cloud promises - for all stages of model development in the cloud. Whether developing next-generation assistants, virtual reality applications, video games, or advertising campaigns, the WEKA Data Platform, saturates large GPU fleets used for model training and tuning, accelerates AI data pipelines, and dramatically reduces the cost of data infrastructure required to build great AI applications in the cloud.

Build a Studio in the Cloud

The cloud is transforming every aspect of the media & entertainment industry. Production companies are inventing entirely new ways of doing business, entirely in the cloud and relying on a globally distributed, highly collaborative pool of artists. As they move to the cloud, studios must continue adapting to higher frame rates, higher resolutions, and evolving techniques like tracking, rotoscoping, and keying. The storage demands of post-production from multi-camera capabilities, stereoscopic virtual reality content capture, higher dynamic range, and increased color depth are breaking traditional approaches resulting in poor artist experience, slow production timelines, and over-budget projects.

The WEKA Data Platform enables production studios and creative agencies to transform their businesses in the cloud completely, continue to build innovative and award-winning content, and facilitate artist collaboration on a truly global scale. Using WEKA software, production studios like Untold Studios, Preymaker, and Parliament VFX can shift to a full "studio in the cloud" business model. They can accelerate content creation pipelines from months to days and reduce infrastructure costs by as much as 65%. To facilitate artist productivity and collaboration, customers use WEKA to achieve a 10x performance improvement for creative applications like Foundry Nuke, Autodesk Flame or Maya, Adobe, Thinkbox Deadline, and Houdini.

Accelerate Drug Discovery

The speed and agility of the cloud, combined with new techniques such as cryogenic electron microscopy (cryo-EM) is transforming how modern pharmaceuticals get done - accelerating time to market, and reducing the infrastructure costs required to discover new treatments. However, the process of getting to the 3D output in a cryo-EM workflow may involve multiple iterations of processing in the computation pipeline to resolve the 3D image appropriately. In many cases, the entire workflow, including scanning a new sample will need to be repeated in order to achieve the appropriate results. This results in a massive amount of data that needs to be stored and retrieved at high speed to make the pipeline work efficiently. Today, many organizations that use CryoSPARC are taking advantage of the ability to rent GPUs and scale the compute functions in the cloud in an elastic manner for processing micrograph images.

WEKA enables researchers who rely on structural biology software tools Relion, CryoSPARC, Schrodinger, and Fluent Biosciences to reduce project times from weeks to days and accelerate time to new drug discoveries and scientific insights.

Accelerate Applications and Reduce Data Costs in AWS

In the cloud, at the edge, or in hybrid deployments, WEKA helps you optimize your AWS environment by accelerating performance-intensive applications, reducing costs, increasing resource utilization, and simplifying data operations. According to a recent study conducted by the Enterprise Strategy Group, WEKA customers:

- Accelerate performance-sensitive and data-intensive applications by up to 10x
- Complete data pipeline analysis for HPC and AI in the cloud by 90%
- Drive faster project velocity leading to a 1% to 20% surge in revenue
- Streamline data operations leading to a 64% reduction in staff hours dedicated to data pipeline management orchestration
- Reduce cloud costs by an average of 38% annually
- Reduce environmental impact with 260 tons of CO2 saved per PB