SOLUTION BRIEF

Drive AI Computing with Run:ai and WEKA

Summary

- Get the most of your data and GPU resources
- Support hybrid cloud deployments
- Drive faster time to production and even more immediate outcomes

Use Cases

- Machine Learning & Deep Learning
- Computer Vision
- Genomics
- High Tech
- Manufacturing

Simplify Orchestration and Acceleration for AI Workloads with Run:ai AI Computing Platform and WEKA Data Platform

Artificial Intelligence (AI) is creating new business opportunities for companies in every industry. It is no surprise that Deloitte AI Institute's 2022 State of AI in the Enterprise found that 76% of their survey respondents are planning to increase their expected AI investment in the next fiscal year.

But operationalizing machine learning (ML) and deep learning (DL) requires the ability to process massive amounts of data from different sources in the shortest possible time. As a result, a new wave of accelerated compute (e.g., GPU) devices on the market and rapidly adopted by enterprises. Furthermore, the 10-100x performance over traditional CPUs provided by these GPUs coupled with a massive increase in Deep Neural Network (DNN) models. As a result, it resulted in a Cambrian explosion in AI.

As your organization expands the scope and scale of its AI efforts, success hinges on the ability to deploy complete AI stacks easily—both on-premises and in the cloud—while ensuring that you deliver:

- Optimized application and data performance and management
- The ability to move data easily between locations and Cloud providers
- Optimized and tested with AI and deep learning applications, network architectures, and commercially available GPUs.

But the reality of today's AI can be a long way from this ideal.
When companies scale their AI operations, they often struggle with the following challenges:

- Inefficient resource usage. Data scientists and engineers are struggling to effectively use and get access to GPU resources and wasting time waiting for mundane tasks to finish. Additionally, they are unable to move data between locations efficiently.

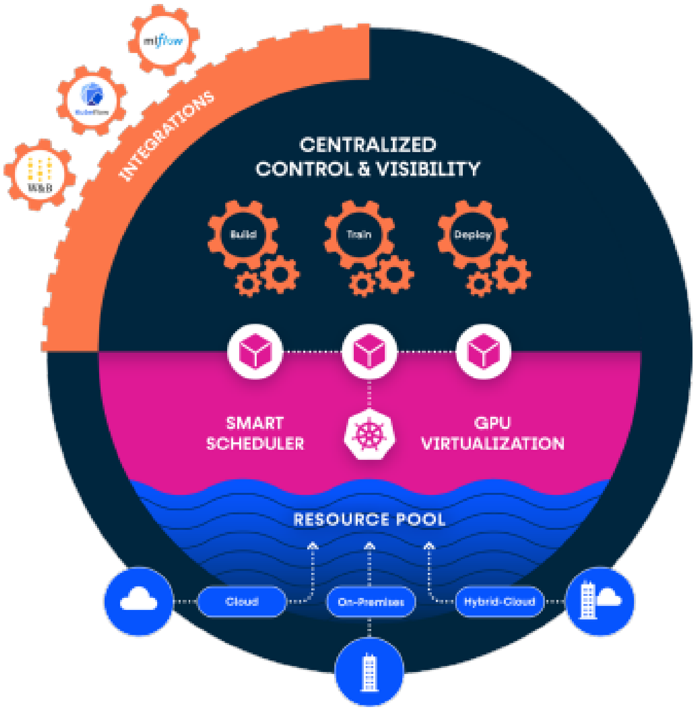
- Diverse tools. Complexity in dealing with the significant differences in the tools and operational processes across different Cloud providers for containerized AI workloads. These tools make it difficult for teams to take advantage of the resources in each location.
- Complex and slow deployments. Deploying a new AI stack to support experiments or production is complex, time-consuming, and error-prone.
- Data-starved GPUs can have insatiable appetite for data. But, unfortunately, data bottlenecks mean that GPUs often sit idle because data sets aren't ready or data platforms can't keep delivering data fast enough. These delays increase costs and delay results.

The MLOps Solution

WEKA has joined forces with Run:ai to help you address these challenges, increasing the success of your machine learning and deep learning efforts. Research teams can gain on-demand access to resources for their entire AI workflow, from building the model to training to inference.

Run:ai Advantages

Run:ai helps organizations simplify and deliver faster on their AI journey from beginning to end. An AI Computing Platform powered by cloud-native operating system can support running your AI initiatives on-premises, on edge, or in the cloud. The Run:ai Atlas platform gathers the compute and GPU resources in a centralized resource pool and then uses a Kubernetes-based Smart Workload Scheduler to ensure dynamic allocation of resources. Deep integration with GPUs allows the effective sharing of these resources across the entire AI workflow. AI practitioners can easily consume resources in a self-service model using the built-in engines to build, train and deploy or by using 3rd party integrations, such as MLflow, Kubeflow, Weights & Biases, and many more.



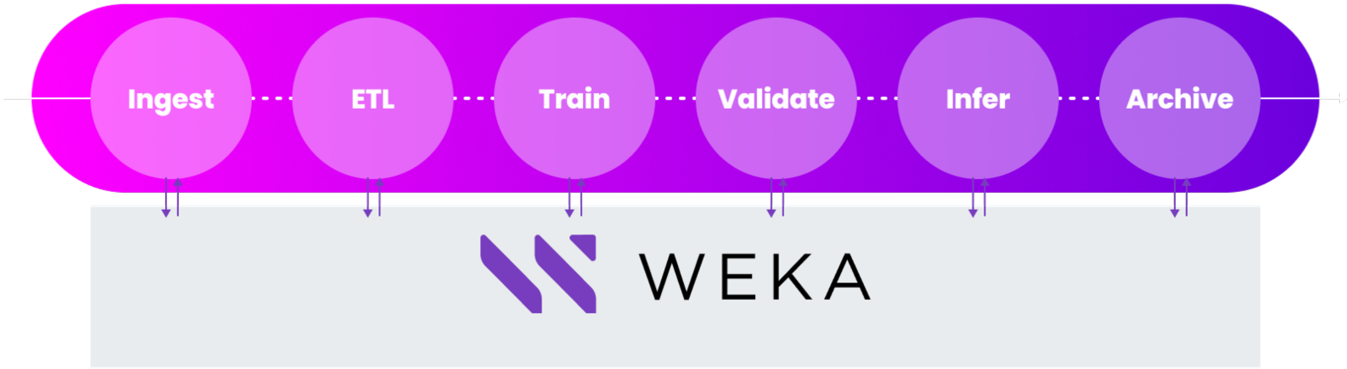
Benefits of Run:ai include:

- Centralize Control and Visibility – Run:ai Atlas offers dashboards and analytics, giving IT insight across all resources and workloads. Align resource allocation to business goals by setting policies and priorities across departments, projects, or users.
- Optimize GPU Utilization and ROI – Automated resource management and efficient sharing of GPU resources enable organizations to achieve higher utilization and increase value per GPU.
- Build on a Truly Open and Extendable Platform – Use the built-in workflows optimized for the complete AI development lifecycle or extend the platform by seamlessly integrating with any AI and ML applications, including MLOps tools.
- Accelerate Hybrid Cloud – Run:ai has the unique capability to deliver centralized control and visibility across resources on-premises or in the cloud, enabling organizations to make hybrid cloud AI infrastructure a reality.

WEKA Advantages

WEKA has built a software-defined Data Platform that leverages cutting-edge cloud, compute, storage, and fast networking technologies to unleash the value of your data.

WEKA Data Platform delivers fast access to data when needed across the AI workflow. Our patented architecture scales to exabyte scale and eliminates the need for data copies, reducing operational complexity, enhancing pipeline efficiency, and increasing GPU utilization.



With WEKA, a single data platform supports all popular data access methods, including the POSIX file system, NFS, SMB, S3, and GPU Direct Storage (for direct data movement between GPUs and storage). In addition, because the platform is software-defined, it enables both on-prem and cloud deployment.

Additional benefits include:

- Move and back up data easily. Advanced data management capabilities simplify backup data to the cloud or move data quickly and efficiently between clouds to meet operational needs.
- Tier automatically. WEKA can automatically tier cold data to low-cost object storage (on-prem or cloud) for better economics. All data remains in the namespace, and metadata stays on the flash tier for fast access.
- Ensure security. The WEKA data platform was architected to ensure the security of your data with advanced authentication, in-flight and at-rest encryption, and flexible key management.

Read more in the WEKA Data Platform architecture white paper.

What Does the Run:ai and WEKA Solution Means for Your Organization

With Run:ai and WEKA joint solutions, your teams can deploy new AI stacks whenever they need them—across private and public cloud clouds for short-term use or long-term needs—in less time and with less hassle.

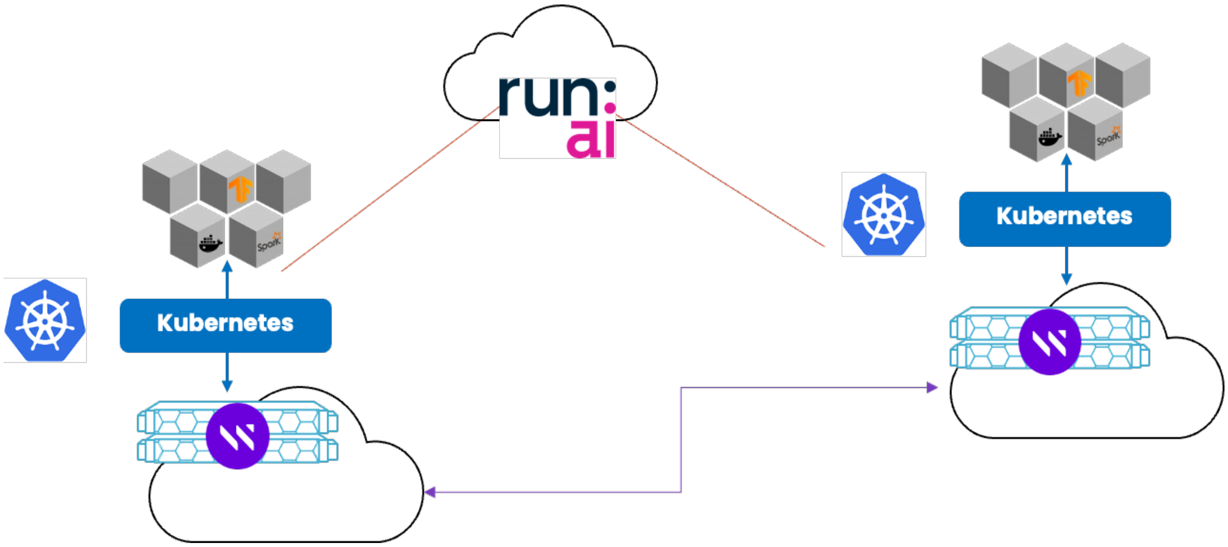


FIG. 1 Run:ai and WEKA allow organizations to create a hybrid cloud AI Compute Platform

Regardless of location, each new solution operates identically, with the same control plane, tools, and capabilities. As a result, your team can immediately use available GPU resolutions in the new stack without learning about any new tools, limitations, or other idiosyncrasies. No matter the underlying infrastructure, you will be confident that the solution will deliver the best possible performance from all the allocated resources. In addition, this combined stack will ensure the maximum return on your investment for cloud infrastructure and additional on-prem hardware.

This results in benefits such as:

1. faster training and validation task execution by as much as 50% through faster data access and improved utilization of GPUs.
2. As much as 5x quicker time to deployment across hybrid clouds.
3. Return on investment measured in months versus alternate approaches to MLOps and storage.

Get Started Today

If you are an organization investing in artificial intelligence, machine learning, and deep learning initiatives, Run:ai and WEKA can help simplify your journey with excellent ROI, better model quality, and an excellent time to production. We are excited to offer a complimentary technical consultation looking at your current and future needs including a proposal for proof of content. To learn more about this offer and more information, email us at mlops-solutions@weka.io or contact your authorized Run:ai and WEKA representatives.