ClearML Announces New Orchestration Capabilities to Expand Control Over AI Infrastructure Management and Compute Cost

New functionality delivers better oversight, management, and control over AI costs while maximizing usage of compute resources and visibility of model serving

SAN FRANCISCO, CA, US, March 18, 2024 /EINPresswire.com/ -- ClearML, the leading open-source, end-to-end foundational platform for unleashing AI in the enterprise, today announced two significant enhancements that give enterprises greater control over their AI/ML development lifecycle.

Because Generative AI introduces new risks over data privacy and ownership (plus substantial investment), ClearML is committed to providing businesses with enhanced control over their entire end-to-end AI/ML development lifecycle.

Today, the company is announcing:

1) A Resource Allocation & Policy Management Center, providing advanced user management for superior quota/over-quota management, priority, and granular control of compute resources allocation policies.

2) A Model Monitoring Dashboard designed for viewing all live model endpoints and monitoring their data outflows and compute usage.

These innovations are important; in a recent ClearML survey of global AI infrastructure leaders, optimizing GPU utilization and reducing the cost of inference were top of mind, with 93% of survey respondents noting that their AI/ML team productivity would substantially increase if real-time compute resources could be self-served easily by anyone who needed them.

ClearML's latest enhancements further enable an AI team's ability to maximize its compute infrastructure by enabling full flexibility and visibility into how compute resources are organized, allocated, and accessed. ClearML enables DevOps to deliver a broader level of self-serve access for data science, machine learning, and Generative AI stakeholders, accelerating the development-to-production lifecycle and delivering a faster time to value. The simplicity of ClearML's set-it-and-forget-it approach also lowers overhead for AI team admins, allowing them to focus on what matters most.

Now, instead of multiple stakeholders competing for the same resources and underutilizing them, organizations can optimize their compute, build dynamic hybrid clusters, reduce manual work for DevOps, and improve productivity. This results in faster shipping of models and higher ROI on AI infrastructure investments.

"It can be difficult to monitor and enforce stakeholders' compute usage and correlated budget," said Moses Guttmann, Co-founder and CEO of ClearML. "We are thrilled to offer these two new enhancements to help customers better utilize and reduce the cost of their compute and understand how their models are using compute in order to optimize their investments."

How the Resource Allocation & Policy Management Center Works
ClearML's focus on controlling users' resource allocation enables enterprises to maximize their compute utilization. Certified to run on NVIDIA AI Enterprise software, the ClearML platform allows companies to maximize their AI investment by boosting NVIDIA GPU servers (like NVIDIA DGX and any other GPU nodes) system utilization by up to 32% through quota and pre-emption capabilities. Using them in tandem with fractional GPU capabilities enabled by NVIDIA Multi-Instance GPU, more end users have access to GPU clusters, and those optimized workloads can use GPU compute more efficiently. Hybrid cloud configurations are also supported, increasing on-premises utilization while maintaining maximum flexibility and spilling workloads into the cloud to ensure high availability at the lowest cost to compute. ClearML is an NVIDIA AI Enterprise partner and NVIDIA DGX-Ready Software partner.

The new Resource Allocation & Policy Management Center delivers a frictionless experience and makes it easy for DevOps and AI leadership to dynamically control access to compute resources for user groups based on rules and logic. Users can be prioritized for (or limited to) certain resource pools, giving AI teams the ability to precisely manage each cluster, whether on-prem, in the cloud, or both. Horizontal scaling rules to support additional compute, such as cloud spillover or autoscaling, can also be applied via the new ClearML capabilities.

With ClearML, any Kubernetes cluster gains scheduling priorities, quota/over-quota, and

automation capabilities without changing the cluster. Other users can provision ClearML with all of the advanced capabilities on their Slurm, bare-metal, or virtual machines (VMs), giving AI teams the ultimate flexibility to define their resource pools and seamlessly manage their AI infrastructure usage according to their needs and specific setup.

About the Model Monitoring Dashboard

Gaining full visibility and control over AI/ML model inference is critical to an organization's data security and its compute infrastructure management. ClearML's Model Monitoring Dashboard shows live endpoints for all models – whether in training or production – providing AI teams with a complete overview and visibility of deployed models. Customers can now deploy internally developed models, as well as open-source LLMs or proprietary LLMs (such as NVIDIA Foundation Models), with a single line while creating maximum visibility of model usage statistics, performance, latency, and volume of requests. This makes it simple to understand which models are either underperforming or consuming too much compute, providing actionable insights into further optimization.

The demand for accelerated compute and inference capabilities at scale will only continue to increase for companies adopting Generative AI. According to ClearML's recent survey, AI leaders have already started considering their options. When asked how their company was planning to address a potential GPU scarcity in 2024, approximately 52% of respondents reported actively looking for cost-effective alternatives to GPUs for inference in 2024. Now, with ClearML's Model Monitoring Dashboard, companies can better manage inference demands to reduce costs, monitor and troubleshoot live models, and benefit from the efficiency of a single pane of glass.

Next Steps

Get started with ClearML by using our free tier servers (https://app.clear.ml) or by hosting your own (https://clear.ml/docs/latest/docs/deploying_clearml/clearml_server). If you need to scale your ML pipelines and data abstraction or need unmatched performance and control, please request a demo (https://clear.ml/demo).

About ClearML

As the leading open source, end-to-end solution for unleashing AI in the enterprise, ClearML is used by more than 1,600 enterprise customers to develop a highly repeatable process for their end-to-end AI model lifecycle. ClearML is an NVIDIA AI Enterprise certified partner and NVIDIA DGX-Ready Software Partner (https://www.nvidia.com/en-us/data-center/dgx-ready-software/) and is trusted by more than 250,000 forward-thinking Data Scientists, Data Engineers, ML Engineers, DevOps, Product Managers, and business unit decision makers at leading Fortune 500 companies, enterprises, academia, and innovative start-ups worldwide. To learn more, visit the company's website at https://clear.ml.

Adam Brett
Crenshaw Communications
Clearml@crenshawcomm.com

Visit us on social media:
Twitter
LinkedIn
