ClearML Announces Extensive New Orchestration and Scheduling Capabilities for Optimizing Control of Enterprise AI & ML

Customers can now fully utilize GPUs for maximal usage with minimal costs – expediting time to market, time to revenue, and time to value

ClearML versus Other Solutions

SAN FRANCISCO, CALIFORNIA, UNITED STATES, October 16, 2023 /EINPresswire.com/ -- ClearML, the leading open source, end-to-end solution for unleashing AI in the enterprise, today announced it has released extensive new capabilities for managing and scheduling GPU compute resources, regardless of whether they are on-premise, in the cloud, or hybrid. Customers can now fully utilize GPUs for maximal usage with minimal costs, resulting in optimized access to their organization's AI compute – expediting time to market, time to revenue, and time to value.

ClearML's newest capabilities bridge the gap between machine learning teams and AI infrastructure by abstracting infrastructure complexities and simplifying access to AI compute. Customers can now take advantage of the industry's most advanced orchestration, scheduling, and compute available to manage and maximize GPU utilization and allocate resources effectively. These enhancements include:

- Optimal Utility: To enable customers to better visualize, manage, and control resource utilization more effectively, ClearML has rolled out a new Enterprise Cost Management Center that intuitively shows DevOps and ML Engineers everything that is happening in their GPU cluster and offers teams a better way to manage job scheduling and maximize fractional GPU allocation and usage as well as project quotas. For example, companies will be able to better understand their GPU usage on NVIDIA DGX™ or DGX™ Cloud machines to maximize utilization of resources and manage costs. In addition, customers can now manage the provisioning, splitting, scheduling, pooling, and usage of GPU resources.

- Extensive Flexibility: The company is set to release new web app capabilities, which will set up a secure Jupyter Lab or VS Code IDE directly in a secure ClearML UI browser window, making it much easier for data scientists to remotely work securely and seamlessly. ClearML already ensures that DevOps and Engineers do not need to be engaged for every user requiring a machine when using ClearML's permissions settings, credentials management, and automatic provisioning. The event history for every cluster is automatically logged for easy auditing and overall governance. Lastly, ClearML's policy management provides DevOps Engineers with easy tools for managing quotas and GPU over-subscription, in addition to job scheduling and prioritization.

"GPU management is critical for companies from a resource allocation and budgeting point of view. By pooling or splitting GPUs to run multiple models on a single GPU, companies can more efficiently serve end users and control costs," said Moses Guttmann, Co-founder and CEO of ClearML. "Our enhanced GPU-handling capabilities, bundled with a robust enterprise cost management center, provide customers with improved AI and ML orchestration, compute, and scheduling management as part of the ClearML MLOps platform – which, right out of the box, replaces the need for a stand-alone orchestration solution while simultaneously encouraging ML team members to self-serve. Engineers now have built-in capabilities to monitor, control, and provision compute resources more easily than ever."

Guttmann noted that ClearML provides seamless, end-to-end MLOps, with full orchestration, scheduling, and compute management integrated into the entire ML lifecycle. In this way, ClearML supports and serves entire AI, ML, and DSML teams, unlike other solutions that only serve DevOps with no secure built-in ML/DL tracking and monitoring capabilities for other team members. In addition, ClearML can be installed on top of Slurm, a widely used free and open-source HPC solution, as well as Kubernetes and bare metal, for managing scheduling and compute. ClearML is also completely hardware-agnostic and cloud-agnostic, maximizing the freedom of companies in choosing and using their vendors and optimizing costs by allowing hybrid on-prem / cloud combinations.

Massive Scalability Takes Center Stage

In addition to helping companies manage their on-premise compute usage, it should be noted that ClearML's Autoscaler capabilities already enable companies to manage cloud and hybrid setups more efficiently, using automatically provisioned cloud machines only when and as needed. For cost management, teams can set budgets for resource usage, with limits set by types, nodes, and idle timeout. Even provisioned GPUs can be automatically spun down if the machine has been idle for a predetermined amount of time, saving energy and costs. Extra budget-conscious teams also have additional options to ensure that cloud machines use spot instances or are not zone limited, with automatic re-spinning when spots are lost, seamlessly continuing running jobs without any external intervention.

"As GPUs become increasingly important for enterprise AI, it's critical to get resources into the hands of the professionals who need them, securely and efficiently," Gutmann noted. "When

coupled with the budgets and per project/group quota limits, ClearML's role-based access control with LDAP integration and SSO allows worry-free access to compute resources for everyone as a self-serve option, while controlling costs."

Next Steps
Get started with ClearML by using our free tier servers (https://app.clear.ml) or by hosting your own (https://github.com/allegroai/clearml-server). Read our documentation here: https://clear.ml/docs/latest/docs/. You'll find more in-depth tutorials about ClearML on our YouTube channel (https://www.youtube.com/@ClearML) and we also have a very active Slack channel (https://join.slack.com/t/clearml/shared_invite/zt-1kvcxu5hf-SRH_rmmHdLL7l2WadRJTQg) for anyone that needs help. If you need to scale your ML pipelines and data abstraction or need unmatched performance and control, please request a demo. To learn more about ClearML, please visit: https://clear.ml/.

About ClearML
ClearML is used by more than 1,300 enterprise customers to develop a highly repeatable process for their end-to-end AI model lifecycle, from product feature exploration to model deployment and monitoring in production. Use all of our modules for a complete ecosystem or plug in and play with the tools you have. ClearML is an NVIDIA DGX-ready Software Partner and is trusted by more than 150,000 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.

Noam Harel
ClearML
email us here
Visit us on social media:
Twitter
LinkedIn
YouTube
