HealthManagement
Promoting Management and Leadership

New Survey Unveils The State of AI Infrastructure at Scale, Exposing GPU Utilization Challenges

Research Report Highlights Key Business AI Insights, Benchmarks, and Priorities

Are you navigating the AI infrastructure market, and need clarity on how the automation and orchestration of AI/ML workflows at scale can elevate your operational efficiency, enhance productivity, and reduce costs? ClearML's latest survey conducted with the AI Infrastructure Alliance and FuriosaAI to understand AI and technology executives' biggest pain points in moving AI/ML to production.

The new report, called "The State of AI Infrastructure at Scale 2024: Unveiling Future Landscapes, Key Insights, and Business Benchmarks" dives into respondents' current scheduling, compute, and AI/ML needs for training and deploying models as well as their AI framework plans for 2024-2025. In this report, we show that while most organizations are planning to expand their AI infrastructure, they can't afford to move too fast in deploying Generative AI at scale at the cost of not prioritizing the right business use cases.

We also explore the myriad challenges organizations face in their current AI workloads and how their ambitious plans for the future signal a need for highly performant, cost-effective ways, such as GPU partitioning, to optimize GPU utilization as well as the need to harness seamless, end-to-end open source AI/ML platforms to drive effective, self-serve compute orchestration and scheduling with maximum utilization.

After reading this report, you'll understand:

- How executives are planning to expand their AI infrastructure
- The critical insights, benchmarks, and key challenges they face
- How they rank priorities when evaluating AI infrastructure solutions against their business use cases
- Why optimizing GPU utilization and GPU fractioning are major concerns and why mitigating compute challenges will be critical to success
- The importance of open source AI Infrastructure
- Why so many respondents (a staggering 74%) are dissatisfied with their current job scheduling and orchestration tools
- Why mitigating compute challenges will be critical to success and the bottom line

This extensive global survey includes responses from AI/ML and technology leaders at 1,000 companies of various sizes across North America, Europe, and Asia Pacific. Dive into the future of the AI Infrastructure landscape by downloading your copy here.

Source & Image Credit: ClearML

Published on : Tue, 30 Apr 2024