HYPERION RESEARCH

EMAIL:
Info@hyperionres.com

Hyperion Research Special Study on Inference Practices

A Hyperion Research Special Analytical Service

Hyperion Research continues its in-depth, user focused research into AI practices, integration efforts, and support infrastructure with a special study specifically focused on inferencing needs, goals, and challenges. This upcoming survey-based study seeks to illuminate cross-industry use cases, purchasing patterns, and HW/SW expectations to support advanced computing end-user inferencing needs.

Information contained in this study includes the popularity of providers, cloud usage rates, hardware preferences, and future-looking expectation for inferencing technology. This survey is one in an ongoing series of studies taking a deep look at advanced computing AI users and their behaviors.

Key Questions

This study is designed to gain a further understanding of the current mindset of those who now have, or plan to have, integrated advanced AI into their HPC or advanced computing environment. Providers' status, budgetary considerations, integration drivers, and sentiment, both current and future-looking, will be key features in this effort. Below are selected key questions that represent the data that will be collected. These questions are only a fraction of those that will be addressed:

- What are the most important goals envisioned by integrating HPC-based AI inference into your organization's existing portfolio of HPC-based workloads?
- How does your organization support existing or planned HPC-based AI inferencing compute needs?
- Which server suppliers does your organization currently use or plan to use to meet its HPC-based on-premises AI inferencing needs?
- What specific processor/accelerator types does your organization use or plan to use to meet its HPC-based on-premises AI inferencing needs?
- With what type of system-level devices do you use or plan to use to support your HPC-based op-premises AI inferencing needs?
- If your organization is already leveraging AI technology within HPC-based processes, what percentage of your overall advanced computing annual budget is dedicated to procuring and maintaining AI inferencing capabilities?
- What is your organization's total annual budget (in US dollars) to support all of your HPC-based computational requirements? Please consider both on-premises and cloud-based resources.