White paper

virtana

Does AI Adoption Have You Concerned?

You Need an AI Data Fabric Copilot

virtana.com

Executive Summary

This paper discusses the significant impact and rapid adoption of Generative AI (GenAI) across various sectors since the release of tools like ChatGPT. This adoption spans many enterprise functions, including marketing, legal, and software development. Analysts estimate that 54% of companies had integrated generative AI into their business processes as of November 2023, which is unprecedented in terms of production IT technology adoption, given that ChatGPT was released only one year earlier.

As GenAI continues to take hold, using third-party Large Language Models (LLMs) and other AI services poses risks, such as unintended data leaks and costs. This exacerbates challenges already associated with managing critical data access for enterprises and introducing new compliance concerns associated with AI implementation. Datacenter management must be upgraded for AI as an important part of an enterprise's strategy to mitigate these risks and costs.

To address these complexities, visibility into and management of the Data Fabric of the AI Data Center (AIDC) will be critical. The AI Data Fabric represents the entire path from the application to the data and back; tuning it is key to achieving maximum performance of the AIDC. Enterprises need an AI Data Fabric Copilot!

Introduction

While early forms of Artificial Intelligence and Machine Learning (AI/ML) have been in use for a while in healthcare, equipment monitoring, logistics, marketing, and financial systems, only in the last year and a half has the integration of AI tools and techniques proliferated throughout all enterprise functions after the release of ChatGPT and other similar tools. This new Generative AI (Gen AI) has wide applicability and is being adopted everywhere in almost every field, demonstrating immediate value in process improvements and creative work and enhancing human-computer interaction and experience.

However, the adoption of GenAI has occurred through both intentional and accidental paths: Marketing is using ChatGPT to improve copy; legal is validating positions in Paxton; and software developers are generating code using tools like Starcoder and GitHub Copilot. Services like MS Office and Google Workspace have woven AI into every aspect of their products, and AI is now used virtually everywhere. Every application will have AI, and organizations will face severe competitive pressure if they do not implement a pragmatic strategy.

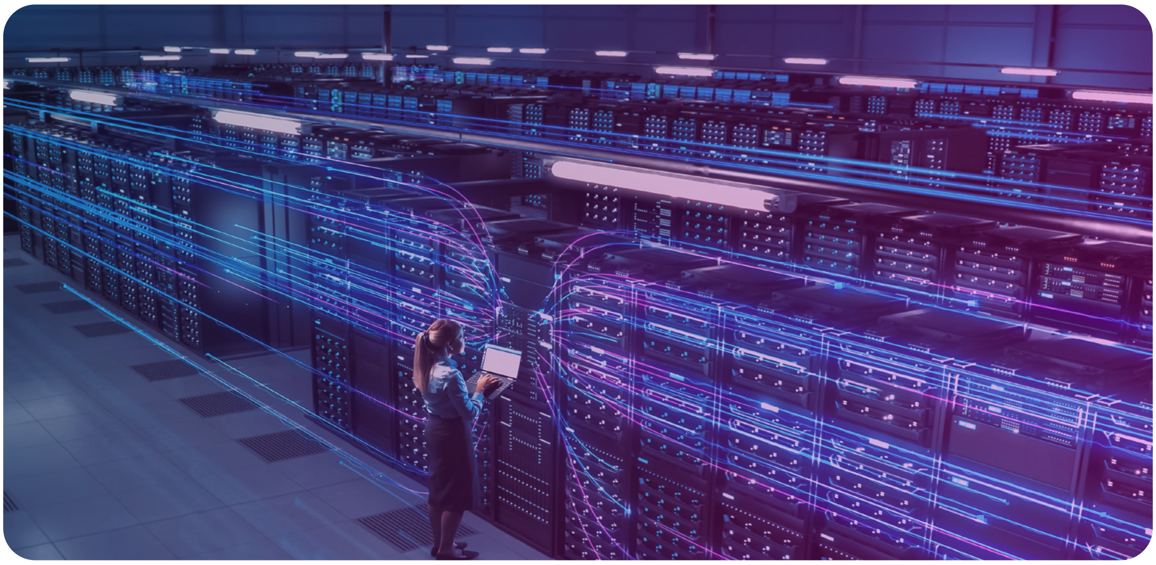
Evidence of this rapid adoption can be seen in the fact that fifty-four percent of companies had used generative AI in their business by November 2023, just one year since ChatGPT was released (PwC, 2023). Unfortunately, most IT organizations are not fully aware of all AI services and software in use and their environment, and few have the tools necessary to manage this new and important component of the IT estate, indicating that actual adoption may be significantly higher.

With this adoption, the key challenges are Costs, Efficiency, & Reliability, which include:

- Model training requires scale-out, high-performance clusters that are generally only available in the cloud. How much of this training work should be in-house, how much can or cannot be in the cloud, and why?
- Resource usage for Inference and model tuning is lower and more readily deployed on-premises.
- AI on the Edge will have inferencing and training for non-LLM models.
- How does one obtain a single integrated view of this hybrid environment and the data flow in a hybrid architecture of this nature to optimize cost, performance, and overall ROI?
- Given the hybrid nature described above, how can an enterprise appropriately manage data access, ensuring appropriate access controls are in place at every stage?
- How can critical data access be correlated with end-user activity when many data access transactions are driven by machine instances, like AI Chatbots?
- How does one balance the inventory of costly and power-hungry new chipsets from Nvidia, Intel, AMD, and others to ensure that device and technology selection is consistent with overall business objectives in terms of both increased productivity and overall cost management?
- How can an enterprise even ensure that they can appropriately manage the budget that is marked for AI infrastructure?
- What about balancing the various components in the AIDC to achieve maximum performance and ROI?

These concerns have resulted in enterprises moving to a hybrid AI strategy underpinned by an AI Data Fabric that:

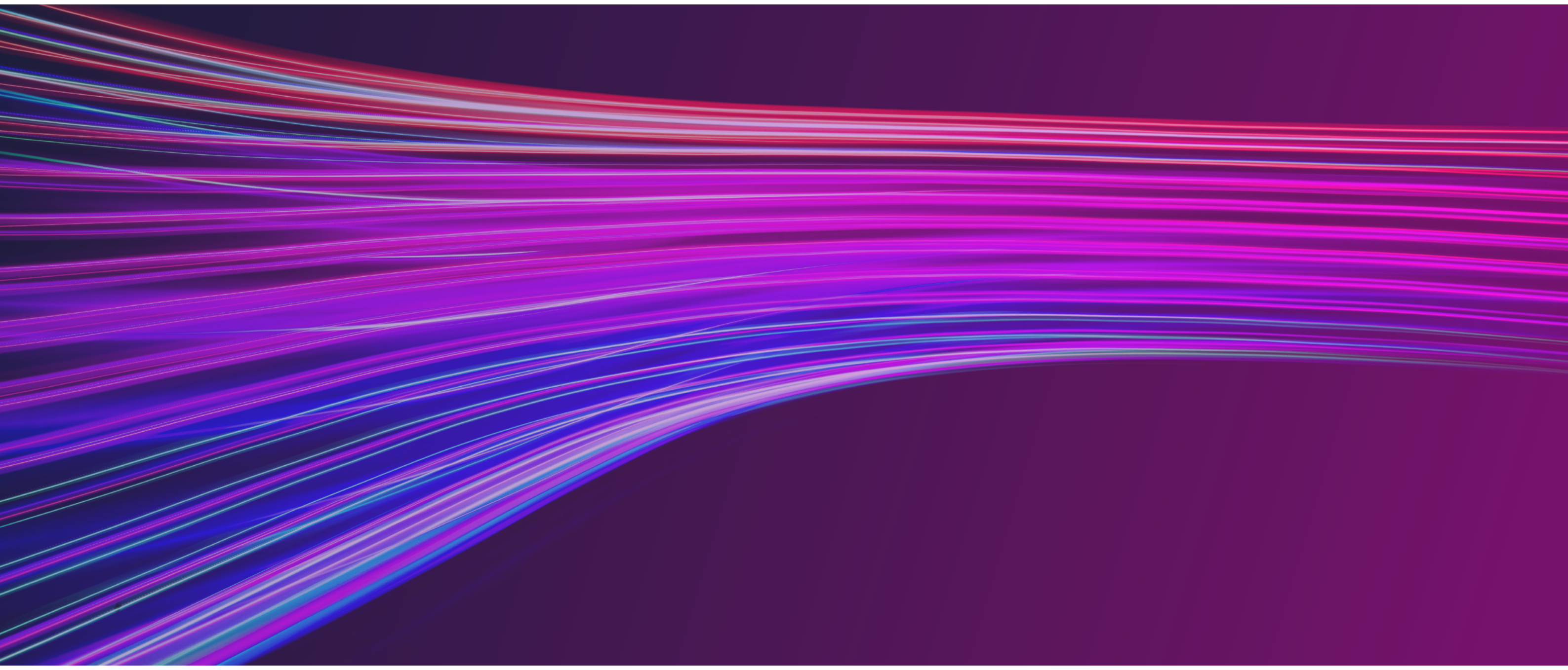
- Repatriates parts of the AI architecture to on-premises deployments to accommodate privacy, security, and cost concerns.
- Distributes the risk and costs over multiple locations, addressing localized behavior and geographic regional requirements like GDPR.
- Deploys multi-party combinations of Cloud AI services, AI data centers (AIDC), edge pods, and on-premises infrastructure.
- Implements an over-arching AI Data Fabric that enables this strategy via diverse protocols for interconnects – PCI/NVMe, Fiber Channel, EFA, InfiniBand, RoCEv2, iWarp, and CXL.



The Need for AI Data Fabric Monitoring

In the AI world, the 'AI Data Fabric' represents the entire path from the application to the data and back, which includes:

- The Application environment scheduled by Kubernetes & GPU add-ons like Slurm.
- The communication between the executing application code and the data.
- The storage entities like memory arrays, near memory, flash, HDD/SDD etc.
- The databases like Rag, RDBMS, NoSQL, etc. that manage & serve the data at rest.



With such diverse infrastructure, enterprises need a way to understand and manage the AI Data Fabric, whether on-premises, at the edge, or in the cloud that provides:

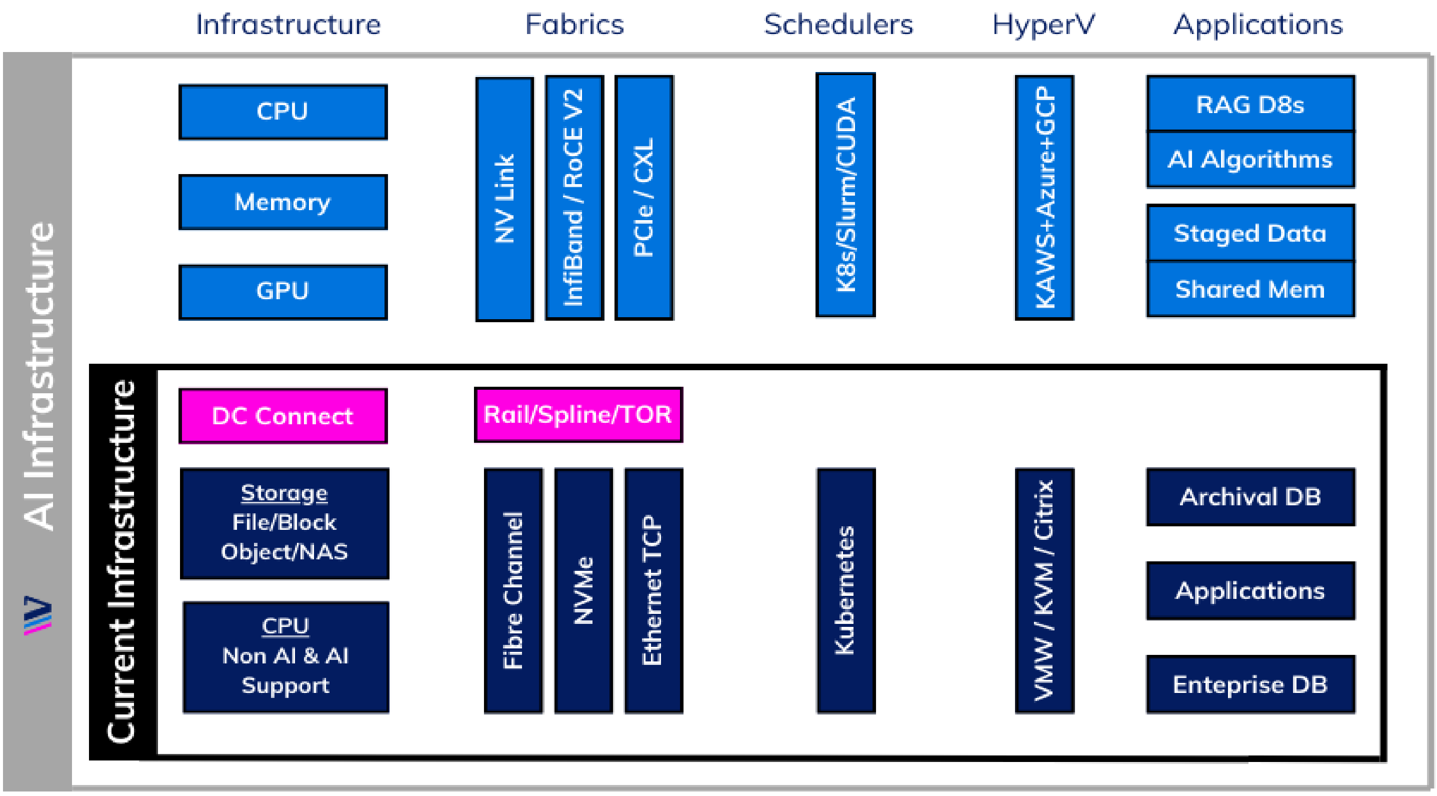
- Visibility into how the AI Data Fabric is performing currently and trending over time.
- End-to-end telemetry and correlation across all AI Data Fabric components (PCI, Ethernet, Fiber Channel, and InfiniBand).
- Proactive problem detection and resolution recommendations.
- TCO analysis and capacity planning.

However, enterprises do not currently have suitable IT controls to provide this level of AI Data Fabric observability. They are left with real concerns about managing costs, optimizing performance, and understanding the accuracy of their AI tools and how to troubleshoot issues, all while managing the risks these tools present to data privacy and regulatory compliance.

Conclusion

AI tools and services promise vast improvements in employee productivity, systems automation, data analysis, content creation, and code development. However, these gains are not achievable without incurring significant operational risks if AI is deployed without appropriate observability and IT controls. Enterprises need a way to understand and visualize their AI Data Fabrics to measure and fine-tune performance, describe and adjust fabric topology, detect and remediate failures, track and manage AI costs, and ensure data compliance. While individual AI software and hardware vendors provide discrete tools to operate their individual products, there is no integrated solution for monitoring and managing the AI data fabric. Enterprises are deploying and operating in the dark.

They need a Data Fabric Copilot!



Current infrastructure environments are typically made up of the components in the black box. AI infrastructure environments include all those and the more specialized components for advanced AI workloads.