Enterprise Strategy Group™
by TechTarget

WHITE PAPER

Building an Efficient, Future-proof AI Infrastructure Platform with Pure Storage and NVIDIA

By Mike Leone, Senior Analyst
Enterprise Strategy Group

February 2023

Contents

Introduction

Organizations—particularly those with data scientists and others tasked with turning mountains of data into insight and action—are looking for new solutions to help drive improved business outcomes. Research from TechTarget's Enterprise Strategy Group points out two important issues: Most organizations still take weeks to gain insight from their data, and then take additional weeks to act on those insights.[1]

To shorten both time to insight and time to action, organizations are turning to artificial intelligence (AI) and analytics. While AI is hardly a new, unproven technology for most organizations, benefitting from production-scale AI workloads can be a formidable task. After all, standing up an AI pilot project in the public cloud may let organizations sample AI's benefits quickly and inexpensively, but those solutions often don't scale economically. That is particularly true when it comes to AI use cases that are both compute- and storage-intensive, such as in healthcare and life sciences, manufacturing, financial services, and any other industry marked by massive data sets and the need for extremely low latency. These two critical requirements can in some cases rule out public cloud solutions. Fast-growing organizations and those generating and processing more data encounter data models that are more complex and less forgiving to things like performance bottlenecks and latency. This is particularly true in analytics workloads, which benefit substantially from the predictable performance at scale delivered by AI tools.

To offset performance, latency, and cost concerns, organizations need to look for a modern, future-proof AI infrastructure built upon cutting-edge, scale-out architecture specifically designed for demanding AI and analytics workloads. At the same time, they need an infrastructure platform that can start small, if necessary, and scale quickly as workload requirements change with demand. This on-premises solution must deliver the benefits of a public cloud—agility, scalability, security—along with the upsides of on-premises infrastructure in terms of performance, resilience, availability, and cost predictability.

Understanding and Overcoming Infrastructure Challenges in Production-class AI

The growing availability of powerful, intuitive AI tools has prompted many organizations to start and learn from AI pilot projects. Those learnings have given organizations the confidence necessary to make AI a core element of their IT initiatives and to use AI as a linchpin for developing new products and services based on remarkable new insights.

However, moving from "science project" to production-stage AI can be challenging, especially with high-value AI workloads. It's not just a matter of lifting and shifting AI workloads from small, sandbox-type deployments to the organization's mainstream data center infrastructure. Too often, organizations decide to deploy mission-critical analytics workloads but run the risk of adding yet another infrastructure silo because legacy systems simply can't provide the sustained performance for big data and other analytics use cases.

There are several reasons why new thinking—and new infrastructure—makes sense when AI workloads need to be scaled to broader deployment. For instance, legacy/traditional infrastructure usually lacks the compute and storage "oomph" to support those workloads and trying to load AI on top of existing infrastructure can interfere with day-to-day operations. With that in mind, existing storage infrastructure often isn't sufficient to handle AI workloads in the same way it handles legacy data center-based workloads. Issues such as model development and training, deployment and operations, and feedback and evolution require more than just additional or faster storage. They

[1] Source: Enterprise Strategy Group Research Report, *Cloud Analytics Trends*, March 2022.

require storage optimized for AI workloads to avoid issues like configuration rigidity, I/O inflexibility, and data isolation.

One of the common challenges associated with enterprise infrastructure is the need to constantly scale or configure hardware because it doesn't meet the needs of the new AI/software frameworks. Instead of standing up new infrastructure every time a new AI workload or tool is deployed, organizations need infrastructure that consolidates all AI data pipelines, regardless of the stage of the process or whether multiple workloads need to be supported.

Although deploying AI proofs of concept in public clouds often makes sense, organizations may run into important challenges when trying to scale those production-class AI workloads. For instance, trying to predict the compute and storage resources needed for AI workloads can dramatically inflate cloud costs, making budgeting tricky and elusive. In some cases, it's leading to the repatriation of workloads. In fact, Enterprise Strategy Group research indicates that 57% of organizations have repatriated workloads from the public cloud back to on-premises environments.[2] And drivers related to AI infrastructure for organizations considering these workload migrations include processing bottlenecks, storage latency, and the need to tightly integrate an AI infrastructure stack on their own. These issues are leading businesses to seek out comprehensive analyses of total cost of ownership of an appropriately configured on-premises solution compared to an equivalent cloud-based deployment. Other considerations in evaluating on-prem versus cloud deployment include security, data mobility, access, data governance, data ownership, and data sovereignty.

If data must migrate from an on-premises data center or a co-location facility to the cloud, latency can be killer, especially for real-time use cases, and it is a prime example of the impact of data gravity. Data gravity is where a large data set attracts applications, compute power, services, and other data. This proves to be a big challenge for cloud environments, making purpose-built AI-oriented infrastructure essential and placing a premium on the evaluation and selection of the right storage platform—one that can be tightly integrated with an underlying compute infrastructure that can yield deterministic performance in on-premises or colocation facilities.

1 in 3 organizations believe their IT team has a problematic shortage of existing skills in AI.

Lastly, for those organizations requiring on-premises AI solutions, taking a do-it-yourself (DIY) or build-your-own AI solution approach is problematic and tricky to pull off. This is especially true in ensuring the right mix of compute and storage resources, while keeping in mind the lack of sufficient AI expertise. Also, a DIY approach is not as simple and efficient to scale with evolving future requirements as it seems at first glance. Recent Enterprise Strategy Group research highlights that 1 in 3 organizations believe their IT team has a problematic shortage of existing skills in AI.[3] The fact that the AI skills gap is real and significant means that organizations need solutions that come pre-designed with the right mix of hardware and software to deliver the right solution and desired outcomes out of the box.

Requirements for a Production-ready AI Solution

Designing, building, and deploying an enterprise-like solution for AI production workloads is no trivial task. In essence, organizations want the best of both worlds. On one hand, organizations want and need a "cloud-like" solution that can be quickly deployed, offers extensive scalability, and is easy to manage. On the other hand, however, production-class AI workloads must run on purpose-built infrastructure that offers predictable, deterministic performance with high availability and rock-solid security. General-purpose compute and storage on-

[2] Source: Enterprise Strategy Group Survey Results, *2021 Data Infrastructure Trends*, September 2021.

[3] Source: Enterprise Strategy Group Research Report, *2023 Technology Spending Intentions Survey*, November 2022.

premises infrastructure no longer offers the performance and scalability needed for AI workloads, but a cloud deployment comes with substantial price-performance tradeoffs that often are difficult to predict.

In these instances, organizations should consider an AI-optimized on-premises solution that is built upon right-sized compute and high-capacity, high-IOPS, low-latency storage. Because AI workloads tend to expand dramatically—and often without immediate visibility—it is vitally important that this AI infrastructure delivers fast, easy, and nearly infinite scalability. It also must be deployed quickly and offer enterprise-class functionality, including multi-layered security, a scale-out storage architecture, built-in replication and snapshots, hybrid cloud support, and easy data management. Additionally, the solution should provide integrated hardware acceleration for compute-intensive AI workloads, such as analytics, simulations, imaging, and machine learning. Also, today's solutions should align with environmental, social, and governance (ESG) issues since AI-driven use cases power through massive amounts of data, which requires the underlying storage infrastructure to be efficient in reducing the overall carbon footprint and supporting long-term, upgradeable lifecycles to meet ESG guidelines.

Organizations also should look for a technology partner that can deliver end-to-end support that reduces the need to find and hire large numbers of infrastructure specialists. The solution also should have agility engineered from the start to ensure long-term relevance without massive new CapEx investments. Working with strategic and knowledgeable partners will enable organizations to gain the necessary guidance to plan and implement an AI-optimized infrastructure so organizations can start with relatively small production-class workloads and grow quickly, easily, and cost efficiently, as conditions dictate.

AIRI//S:
The Collaborative AI Solution from Pure Storage and NVIDIA

Any organization looking to make AI-driven workloads an integral and strategic part of its applications portfolio must answer an essential question: How should compute and storage infrastructure be modernized and right-sized to be considered truly AI-ready, now and in the future? As noted above, there are several essential requirements and capabilities for IT infrastructure to be able to handle production-class AI workloads and deliver the economic and operational value that AI promises to achieve. From a big-picture perspective, those requirements also must:

- Simplify the move from pilot to production-class with an integrated, validated design that uses best-of-breed components and engineering philosophy for optimized storage and compute.
- Start small, if appropriate, and scale non-disruptively over time.
- Achieve cloud-like agility, but in an on-premises solution that offers the best of both environments.
- Offer performance-optimized storage and accelerated computing systems, both individually and, especially, in concert.
- Identify and work with the right technology partners that offer state-of-the-art technology components optimized for demanding AI workloads to provide a solution that works out of the box.

Pure Storage and NVIDIA have developed an integrated AI-ready infrastructure: AIRI//S.[4] This solution combines NVIDIA's DGX™ A100 or NVIDIA DGX™ H100* system, NVIDIA end-to-end networking, and Pure Storage's FlashBlade//S storage platform[5] and is designed to simplify the task of scaling pilot or sandbox AI projects into production status quickly and without glitches. Since AIRI//S leverages NVIDIA DGX, it includes the best-of-breed AI tools that power every DGX system, including NVIDIA AI Enterprise and NVIDIA Base Command. NVIDIA AI Enterprise includes cloud-native developer software such as data science libraries, optimized frameworks, and pre-trained models. NVIDIA Base Command is the operating system of the AI data center, providing a single pane of

[4] Source: Pure Storage, Introducing AIRI//S™: Modern AI Infrastructure.

[5] Source: Pure Storage, FlashBlade//S: Storage Built for AI.

glass to manage AI development workflow, job orchestration, scheduling, cluster management, acceleration libraries, and more.

AIRI//S is a disaggregated, modular, scale-out storage[6] solution for organizations trying to cope with large, fast-growing data sets typically associated with AI workloads. It is designed for quick installation and implementation, speeding time to value and simplifying user onboarding.

Among some of the key attributes of the solution are:

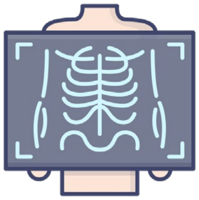
- Full-stack, pre-integration for AI workloads to maximize AI training performance.
- Sustained performance at scale to help organizations keep pace with the expanding scale and scope of AI workloads.
- Future-ready design to deliver long-term economic value without costly mid-cycle upgrades of CapEx equipment.
- Sustainability features that promote data center efficiency, such as lower power consumption in both the storage system provided by Pure Storage and the NVIDIA DGX A100 or the new NVIDIA DGX H100 system.
- The NVIDIA DGX H100 system features eight NVIDIA GPUs and two Intel® Xeon® scalable processors and delivers substantially more performance than its predecessor.

Empowering AI Transformation in Key Industries

Because AI workloads are increasingly used as mission-critical solutions, it is important to understand how Pure Storage and NVIDIA have jointly delivered real-world AI solutions based on AIRI across different industries. For instance:

Automotive: A Europe-based engineering lab of an autonomous driving solutions company uses sophisticated AI to create precise mapping and simulation for navigation. AI and machine learning drive the accuracy of the company's products and services. They need powerful, modernized, and AI-optimized infrastructure to feed the analytics engine and handle extremely large data sets. To deliver the performance, scalability, availability, and security for their AI workloads, the company selected AIRI jointly developed by Pure Storage and NVIDIA and powered by NVIDIA DGX systems. The AI-ready infrastructure allowed the company to provide its customers with ultra-accurate navigation and mapping solutions without performance bottlenecks that could have impacted the reliability and user experience of their products.

Healthcare: For a leading global supplier of MedTech solutions, improving patients' health outcomes is heavily dependent upon sophisticated imaging solutions. Imaging solutions are well known to generate very large quantities of unstructured data, putting particular emphasis on the need for cutting-edge storage and performance-centric accelerated computing systems. The supplier worked with Pure Storage and NVIDIA to deploy AIRI to activate real-time analytics and AI/machine learning. The AI-optimized infrastructure was used to help speed the development of an AI production system for intelligent models for software development. They combined the Pure Storage and NVIDIA solution built with NVIDIA DGX and used Red Hat OpenShift to quickly deploy container-based application development rather than rely on traditional runbooks, thus reducing human error and speeding application development.



[6] Source: Pure Storage, *Eliminate Traditional Scale-out Storage Limitations*.

Public Sector: An Asia-based higher-education institution focusing on science and technology research needed to upgrade and update its AI computing facility to help students do their research projects more efficiently. To make that goal into a reality, the institution used AIRI AI-ready infrastructure with Pure Storage and powered by NVIDIA DGX systems to accelerate projects, simplify management, and support a wider range of projects with scale-out storage architecture. Additionally, the Pure Storage and NVIDIA solution helped the institution's IT team monitor storage usage and demand spikes anytime and anywhere, thus improving efficiency of overall infrastructure management.

Pure Storage and NVIDIA have worked together for years to jointly develop and deploy AI-optimized infrastructure across a wide range of industries and AI use cases. Their ability to collaborate on everything from system design and ESG functionality to global service and support is an important asset for organizations looking to deploy AI for a range of applications and use cases.

Conclusion

Moving from AI pilots and proofs of concept to production systems is an important and potentially complex transition, heavily dependent upon choosing the right infrastructure platform to base the solution on. While using the cloud to stand up early-stage AI trials often makes sense, it usually doesn't stand up to the performance and cost-efficiency demands for data-intensive production AI workloads.

Instead, organizations should strongly consider the benefits of building AI solutions with on-premises infrastructure—as long as that infrastructure is purpose-built for AI workloads with high performance, lightning-fast storage, low latency, and easy scalability down the road.

Pure Storage and NVIDIA have teamed up to engineer, integrate, and deliver a purpose-built AI infrastructure platform designed to stand up to the rigors of production-class AI. The AIRI//S infrastructure solution avoids the many headaches of legacy infrastructure or DIY solutions with pre-integrated, validated, easy-to-deploy, and easy-to-scale solutions that are future-proof for continuous innovation.

About Enterprise Strategy Group
Enterprise Strategy Group is an integrated technology analysis, research, and strategy firm that provides market intelligence, actionable insight, and go-to-market content services to the global IT community. © TechTarget 2023.

✉ contact@esg-global.com