

Technical Research Study

PROWESS

# Navigating Data Locality, Pipeline Performance, and Operational Complexity in Enterprise AI Infrastructures

An independent architectural evaluation of Dell Technologies, NetApp, Everpure, and VAST Data across distributed enterprise environments.

# Executive Summary

As enterprise AI initiatives move from pilot phases to production scale, infrastructure strategy faces a critical turning point. Organizations are no longer evaluated solely on peak theoretical throughput or isolated lab benchmarks. Long-term success now depends on architectural resilience, data locality, and sustainable operations.

In this report, Prowess Consulting presents an independent evaluation of solutions from Dell Technologies, NetApp, Everpure, and VAST Data strictly against real-world shipping documentation, enterprise data gravity realities, and cost implications. This comparison examines how these solutions perform in distributed enterprise data environments, including their strategic architectural impact, data-locality handling, pipeline acceleration, synchronization overhead, and multi-vendor integration.

A central finding of our evaluation is that architectural claims often diverge significantly from actual capabilities available in shipped products. Many platforms rely on temporary data staging, closed namespaces, or background replication loops that create hidden operational friction. As AI pipelines scale, architectures must minimize data movement and eliminate unnecessary synchronization dependencies to prevent administrative friction, storage bloat, and costly pipeline delays.

**TOP FINDING**

Operational risk, not feature lists, drives the true cost of AI data platforms. Open in-place access, which provides the ability to process data where it resides without copying or moving it, reduces complexity, while proprietary wrappers, synchronization, and closed namespaces create friction at scale.

This report provides an objective audit of how each platform's design choices impact multi-year operations, helping IT leaders separate marketing roadmaps from proven, production-ready execution across five architectural pillars.

# Defining the Modern AI Data Platform

For this study, we define a modern AI data platform as an integrated infrastructure and software ecosystem that provides unified data access, cross-estate query federation, pipeline acceleration, and open toolchain compatibility without forcing data migration or requiring proprietary query engines. Establishing this precise definition is essential because vendor marketing has blurred traditional boundaries, requiring clear architectural benchmarks to separate genuine multi-cloud capabilities from proprietary closed-stack designs.

Evaluating major industry alternatives reveals divergent design philosophies across the market. NetApp's evolving ecosystem strategy, comprising its recent data processing acquisition and platform announcements, and VAST Data's prescriptive global namespace (incorporating VAST® AI OS and VAST DataSpace) address distributed data through centralized control points. These architectures rely on tightly coupled software layers or distinct staging boundaries. Similarly, Everpure pairs an intelligent discovery layer with all-flash data planes to manage performance. While these approaches offer specific administrative benefits, they can also introduce architectural friction, reliance on future product releases that may not yet be available, or forced data movement that conflicts with distributed data gravity.

To audit how platforms perform against these real-world requirements, this study evaluates the architectures of four leading vendors: NetApp, VAST Data, Everpure, and Dell.

# Differentiation of Platform Architectures from Vendor Messaging

Real architectural differences still exist between platforms, but promotional narratives often obscure how systems actually function in production. Enterprise architects must look past vendor messaging to evaluate shipping documentation, separating verified capabilities from roadmap promises.

Across the industry, vendor messaging has increasingly adopted terminology that makes it hard to judge platforms based on promotional descriptions alone. For example:

- NetApp promotes an architecture designed to "make your distributed enterprise data agent-ready without moving it, copying it, or rebuilding your storage architecture," supported by the launch of the NetApp® AI Data Engine (AIDE).[1]
- VAST Data relies on an "open ecosystem that gives organizations the flexibility to choose the technologies that best meet their performance, operational, and business requirements."[2]
- Everpure describes its offering as an enterprise data cloud operating on a unified data plane and an intelligent control plane.[3]
- The Dell AI Data Platform provides an open, modular foundation designed to turn distributed, siloed data into AI-ready assets.[4]

Because this terminology makes architectural evaluation challenging, we conducted this technical research study to cut through the marketing language and provide enterprise IT leaders with an objective, documentation-backed analysis.

To evaluate real-world system behavior against an established IT landscape, organizations must determine how each platform interacts with their existing estates. Each vendor approaches integration through a significantly different mechanism:

- NetApp's developing AIDE strategy is an ambitious move from storage into AI data activation. The company is currently focused on heterogeneous, in-place discovery without moving or copying files, limiting itself to data discovery and dataset curation.
- VAST SyncEngine solves the last mile data problem by cataloging, migrating, and preparing enterprise data, turning scattered assets into AI-ready pipelines at scale.[5]
- Everpure layers an intelligent discovery framework across third-party clouds and storage while keeping an all-flash data plane underneath.
- The Dell AI Data Platform pairs Dell PowerScale™, Dell ObjectScale™, and the Lightning File System with a federated control plane that queries file systems, object stores, warehouses, and clouds in place, leaving data where it lives instead of requiring a central copy.

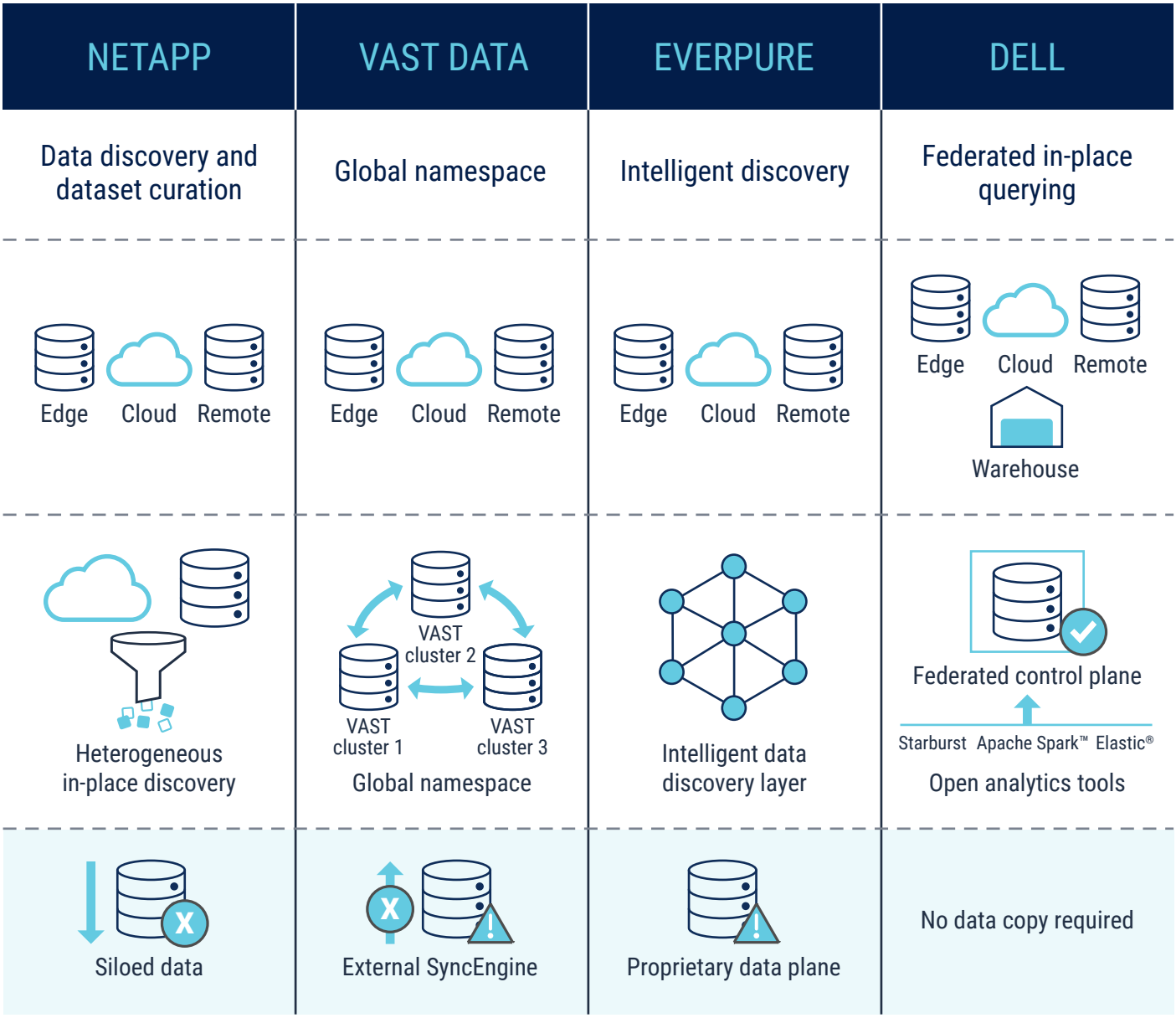



Figure 1 | Architectural approaches to distributed data estates

To successfully navigate these platform differences against an established IT landscape, organizations evaluating AI data platforms must assess essential operational criteria through five critical pillars:

1. Data locality constraints in distributed AI estates
2. Data movement and synchronization risks
3. Resolving the GPU bottleneck through in-place pipeline acceleration
4. Unified governance, security, and open interoperability in the AI era
5. Evaluating operational overhead and total cost of ownership (TCO) in enterprise AI

Assessing platform behavior through these pillars can help identify impacts on day-two operations and architectural costs.

# Pillar 1: Data Locality Constraints in Distributed AI Estates

Data locality remains a foundational constraint when building scalable enterprise AI pipelines. Research indicates that 93 percent of incremental enterprise data resides outside centralized public cloud environments,[6] making this an unavoidable reality for infrastructure strategy.

As data estates span multi-site data centers, edge locations, and diverse repositories, bringing compute to the data or querying the data where it lives determines whether a platform scales efficiently or chokes on network latency and redundant copies. Different platform architectures approach this challenge through contrasting structural designs:

- NetApp's strategy has evolved from traditional centralized network-attached storage (NAS) management toward a hybrid discovery model. With the AIDE metadata frameworks, NetApp addresses data locality by scanning and indexing distributed files and objects in place across hybrid multi-cloud environments. This enables analytics and AI tools to inspect datasets without an immediate, full-scale data migration. However, NetApp's framework remains limited further down the AI data pipeline, as active staging and tiering workflows continue to pull blocks into local storage pools rather than supporting true, uncompromised in-place processing for advanced GPU workloads.
- VAST DataSpace extends a global namespace across distributed VAST clusters using its VAST Global Access architecture. Remote data remains resident on the source cluster and is made accessible through satellite clusters that retrieve and locally cache data on demand. While VAST can also use SyncEngine and replication services to ingest, catalog, and synchronize data from external sources, DataSpace and Global Access primarily rely on distributed metadata management, leases, prefetching, and localized caching rather than wholesale background synchronization of datasets.
- Everpure approaches data locality by pairing an intelligent discovery and metadata management layer with an all-flash local data plane. Instead of federating queries directly across disparate external repositories in real time, Everpure's architecture abstracts storage locations through its unified control plane while caching or landing active datasets onto its high-performance flash layer for processing.
- The Dell AI Data Platform addresses distributed data gravity by pairing PowerScale, ObjectScale, and the Lightning File System with a federated control plane. Rather than forcing data into a central repository or relying on continuous background sync loops, it utilizes specialized analytics, processing, and search engines to query files, objects, warehouses, and cloud repositories directly where they live, including data residing on non-Dell storage infrastructure.

Evaluating platforms against this first pillar demonstrates that true data locality requires more than metadata visibility; it requires query execution engines that can operate across hybrid estates without triggering unnecessary data movement or synchronization friction.

# Pillar 2: Data Movement and Synchronization Risks

Reducing data copies within an enterprise architecture is fundamentally distinct from eliminating them. Rather than assuming platforms operate entirely without data movement, enterprise architects must evaluate how platforms handle data over its entire lifecycle, recognizing that continuous ingestion and movement remain a reality across almost any infrastructure.

NetApp's strategy reflects ongoing evolution as it navigates the transition from its initial AI data engine framework toward broader integration. Previously, NetApp architectures relied heavily on SnapMirror replication streams from ONTAP® data sources to specialized analytics layers, functioning as an active data-movement engine. Because these mechanisms were closely tied to ONTAP sources, multi-vendor environments often required structured data-placement workflows.

With NetApp's recent acquisition of DataPelago and the introduction of the NetApp Platform, enterprise architects must evaluate whether new product additions, governed by standard forward-looking product disclaimers regarding unreleased roadmap features, truly deliver in-place processing capabilities or maintain existing data pipeline dependencies.

Similarly, VAST Data deploys specialized synchronization engines, such as VAST SyncEngine, to discover, catalog, and bridge data born outside the VAST namespace into modern AI pipelines. While these engines do bridge external data sources, they also require additional data-mobility and metadata-management infrastructure to operationalize non-native data sources.

Everpure addresses data synchronization by utilizing an intelligent discovery layer across heterogeneous storage and public clouds. While this capability discovers and maps external assets in place without moving them, it leaves the burden on the customer to decide how to reconcile, ingest, or stage those assets for compute-intensive workloads, adding ongoing management complexity.

As enterprise data volumes scale into multiple petabytes across distributed estates, managing continuous background synchronization workflows adds persistent network overhead and consumes cluster compute. Synchronization also increases the risk of metadata misalignment and structural divergence. When enterprise AI workflows operate exclusively within a closed namespace, every piece of data outside that boundary becomes a continual operational obligation that an engineering team must own indefinitely.

Relying on permanent background sync mechanisms requires ongoing software licensing and infrastructure overhead, including fees for replication tools, sync engines, and management tiers. None of these come with a definitive project cutover date. Instead of advancing core business goals, organizations could find their budgets tied up in recurring software costs and administrative maintenance paid for over the platform's lifetime.

The Dell AI Data Platform helps eliminate continuous background replication and permanent sync relationships by querying distributed file systems, object stores, and clouds in place. By federating access instead of copying data, Dell can help spare organizations the permanent sync-engine overhead that quietly inflates the cost of running enterprise AI at scale.

# Pillar 3: Resolving the GPU Bottleneck Through In-Place Pipeline Acceleration

Everpure and VAST put up impressive headline speed metrics. Everpure achieved the official SPECstorage® AI_Image record with FlashBlade//EXA™ at 6,300 concurrent AI jobs,[7] and VAST reported a 9x time-to-first-token (TTFT) speedup using AMD Instinct™ MI355X hardware for key value (KV) cache offloading.[2]

These benchmarks capture immediate industry attention. But this focus on raw speed is often misplaced because benchmark leadership can change as hardware evolves, while architectural design dictates whether a system introduces permanent bottlenecks. Ultimately, marketing milestones fail to reflect ongoing production realities.

We include these metrics to illustrate industry marketing claims rather than to endorse them. Because both figures are vendor-reported and bound to specific hardware generations, performance leadership shifts continuously with each product cycle. Ultimately, benchmark numbers are transient. The underlying platform architecture is what determines long-term operational viability.

Performance acceleration matters most where the pipeline is slowest, which is rarely the final leg to the GPU. In most enterprise estates, the true bottleneck is upstream: CPU-bound preparation, including the decompression, format parsing, and metadata translation that must happen before data ever reaches the storage network or the GPU. If an infrastructure design forces data through those intermediate translation steps before it reaches compute, optimizing the final leg yields diminishing returns.

NetApp addresses performance through its alignment with the NVIDIA AI Data Platform reference design. However, customers must evaluate how these acceleration frameworks perform in production given ongoing shifts in NetApp's product roadmap and the integration of acquired data-processing capabilities.

To address upstream bottlenecks before data reaches compute layers, Dell applies NVIDIA CUDA-X AI® libraries, cuDF, and cuVS inside the data engines themselves, reporting up to 3x faster data processing, up to 12x faster vector indexing, and a 19x reduction in TTFT versus CPU-based workflows.[8] Dell embeds acceleration directly into the storage and retrieval engine, removing the traditional software-stack friction that slows down enterprise pipelines.

This is the critical distinction: Dell accelerates data where it rests, not a staging copy of it. Because the data engines run NVIDIA acceleration in place, processing and vector indexing speed up exactly where the data already lives, with no preliminary ingestion or sync copy required. A last-mile network or GPU-cache benchmark only helps after data has been moved and ingested, so the competing architecture still pays the upstream data-movement tax that the headline figure never measures.The takeaway for buyers is that benchmarks can change, but single-vendor, in-place processing can remove the data-movement bottleneck and the cross-tier finger pointing that follows.

# Pillar 4: Unified Governance, Security, and Open Interoperability in the AI Era

As AI workloads increasingly require multimodal data discovery, real-time vector search, and complex metadata querying, an infrastructure platform's ability to integrate seamlessly with analytics and search engines becomes a core operational requirement. Modern enterprise data lakes rely heavily on open table formats and query engines such as Apache Iceberg®, Snowflake®, Databricks®, and Trino to unify structured and unstructured assets.

NetApp addresses analytics and search through its storage management and data services framework. This raises critical questions about support accountability when integrating with multi-vendor engines like Trino or external data platforms like Databricks and Snowflake, as these integrations frequently encounter operational friction or cross-stack failures.

VAST integrates search and analytics natively into its global namespace architecture using distributed metadata management, alongside capabilities designed to interoperate with open table formats like Apache Iceberg. However, because VAST depends heavily on external partner integrations, such as joint ecosystems with cloud analytics platforms, troubleshooting performance bottlenecks and query degradation can become complex.

Everpure™ Data Intelligence and Everpure Data Stream index files across hybrid environments, building new streaming, vector search, and data lakehouse functionality through hardware co-engineering partnerships and integrations with analytics platforms. This approach can also test cross-vendor accountability for support.

The Dell AI Data Platform avoids forcing data movement or relying on restrictive proprietary wrappers for analytics and search. It delivers analytics and search through open, named data engines rather than a proprietary core.

- The Dell Data Analytics Engine, powered by Starburst, provides open, distributed SQL analytics.
- The Dell Data Processing Engine, powered by Apache Spark™, handles large-scale data processing and preparation.
- The Dell Data Search Engine, powered by Elastic®, executes hybrid keyword and vector retrieval across enterprise content sources, leveraging GPU-accelerated index creation to achieve up to 12x faster vector indexing.[8]

Metadata is the data being searched, not the search engine itself. The Dell Data Search Engine, powered by Elastic, is what performs the search, including the vector search that powers retrieval-augmented generation (RAG). By combining PowerScale, ObjectScale, and the Lightning File System with this federated control tier, Dell allows organizations to query files, objects, cloud repositories, and open table formats like Apache Iceberg directly in place, keeping the engines a data team already chose, such as Trino, Databricks, and Snowflake, as first-class citizens rather than migration targets.

By delivering unified, single-vendor AI infrastructure accountability across servers, networking, and storage, Dell eliminates finger-pointing and streamlines support when resolving cross-stack analytics, vector indexing, or search performance issues. This keeps operations predictable and cost-effective as data estates scale.

# Pillar 5: Evaluating Operational Overhead and TCO in Enterprise AI

Initial infrastructure proposals often focus heavily on acquisition costs and raw peak performance numbers. But true platform evaluations must account for ongoing operational overhead, management complexity, and TCO. When hidden administrative and maintenance burdens compound over a multi-year deployment, initial hardware savings evaporate.

Administrative friction occurs when platform management is fragmented across multiple interfaces. For example, documentation for NetApp environments directs administrators across the AIDE Console, ONTAP System Manager, and the ONTAP command-line interface (CLI) for routine operational tasks.[9] Jumping across multiple disparate control planes increases the risk of configuration errors and the training costs for engineering teams.

NetApp's exclusive focus on the storage layer requires organizations to stitch together separate third-party compute, cloud, and management tools, introducing administrative overhead and toolchain fragmentation.

VAST delivers a globally unified namespace through its DataSpace architecture. However, enterprise deployments often extend beyond the core storage platform to include external analytics, AI, cloud, and data-mobility services. As these environments scale, organizations might need to coordinate upgrades, validate interoperability, and troubleshoot across multiple technology domains. This can increase operational overhead and lifecycle management costs that initial acquisition comparisons can often miss.

Everpure uses an intelligent discovery layer to map assets across heterogeneous storage and public clouds, requiring engineering teams to continuously reconcile underlying data planes and coordinate multi-vendor support pathways that drive up long-term management complexity. Relying entirely on all-flash arrays across every tier adds unnecessary infrastructure costs and fails to optimize the economics of warm or cold datasets that don't require high-speed flash performance.

The Dell AI Data Platform reduces administrative friction by unifying file, object, and cloud repositories under a federated control tier, eliminating the operational tax of disparate tools and ensuring predictable, single-vendor infrastructure management.

The most sustainable AI data platform minimizes administrative effort, consolidates management interfaces, and delivers fully realized capabilities today rather than future promises. By providing unified management and predictable economics, a mature platform protects IT budgets and helps ensure operational continuity.

# What to Ask Before Procurement

Evaluating modern AI data platforms requires separating proven architecture from unverified vendor claims by forcing teams to answer hard questions grounded strictly in shipping documentation. Before finalizing their selections, customers should explicitly pose these critical inquiries to every vendor regarding each of the pillars:

- **Data locality constraints:** For data that will remain outside the platform over a three-year horizon, what in-place access capabilities does the current shipping release support versus what remains dependent on future product roadmap promises?
- **Synchronization risks:** If data must move or synchronize across distributed storage tiers, which underlying replication engines or background synchronization mechanisms are triggered, and how can you avoid perpetual data movement dependencies?
- **Pipeline acceleration and inference performance:** What native mechanisms or hardware-accelerated libraries does the platform use to eliminate upstream bottlenecks and accelerate analytics and AI pipelines without relying solely on benchmark highlights?
- **Ecosystem integration and support accountability:** If an organization retains existing analytical tools like Databricks, Snowflake, or open formats like Apache Iceberg, what specific integration layers or proprietary database cores break, and who owns sync jobs when metadata structures change?
- **Operational overhead and stack complexity:** Which administrative console or interface does each specific operational task actually execute within, and what does the architectural exit strategy look like if enterprise AI strategies shift in year three?

Ultimately, architectural costs and complexities are never fully captured on day one. As environments scale, new AI workloads emerge, additional data sources come online, and continuous migration requirements or synchronization engines compound operational friction. As a result, early architectural shortcuts turn into heavy long-term burdens.

Shifting the focus from simple vendor or feature evaluation to comprehensive risk evaluation reveals that the fundamental question is no longer, "Which platform has the best list of features?" The definitive strategic choice facing enterprise architects is, "Which architectural assumptions create the least operational risk over the next three to five years?" Grounding decisions in open, in-place data access and single-vendor accountability helps ensure that infrastructure remains resilient, cost-effective, and fully aligned with long-term enterprise AI objectives.

## Endnotes


1 NetApp. "**NetApp AI Data Engine**." Accessed October 2026.

2 VAST. "**VAST Data Expands Collaboration with AMD to Advance AI Infrastructure for the Inference Era**." July 2026.

3 Blocks & Files. "**Everpure extends data and control planes for agentic AI**." June 2026.

4 Dell. "**Dell AI Data Platform: A Technical Perspective**." July 2026.

5 VAST. "**Massively Scalable AI Data Synchronization**." October 2026.

6 Digital Realty. "**Data Gravity Index (DGx)™ 2.0.**" 2023.

7 Blocks & Files. "**Everpure tops SPECstorage Solution 2020 AI IMAGE benchmark charts**." March 2026.

8 Dell. "**Dell AI Data Platform with NVIDIA Supercharges Enterprise AI with Breakthrough Data Orchestration and Storage Innovations**." March 2026.

9 NetApp. "**ONTAP Technical Reports | Management tools.**" August 2026.




## Legal Notices and Disclaimers

The analysis in this document was done by Prowess Consulting and commissioned by Dell Technologies.
Claims have been verified against publicly available information and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.
Prowess and the Prowess logo are trademarks of Prowess Consulting, LLC.
Copyright © 2026 Prowess Consulting, LLC. All rights reserved.
Other trademarks are the property of their respective owners.

1126/260102