



The Storage Squeeze: How Smart Organizations Are Funding Their AI Future

Data Optimization, Cost Economics, and Infrastructure Strategies for AI Success



Executive Summary

The End of the "Just Store Everything" Habit

Treating enterprise storage as an infinite, cheap commodity is no longer financially viable. For over a decade, organizations hoarded unstructured data under the assumption that hardware capacity would remain perpetually inexpensive. Today, aggressive artificial intelligence initiatives demand massive computational power. Paired with stretching hardware procurement cycles and soaring infrastructure costs, unmanaged data accumulation has become a direct financial drag.

With nearly 66% of organizations actively increasing storage budgets or adopting new storage solutions, the influx of unstructured datasets for AI is straining traditional infrastructure.

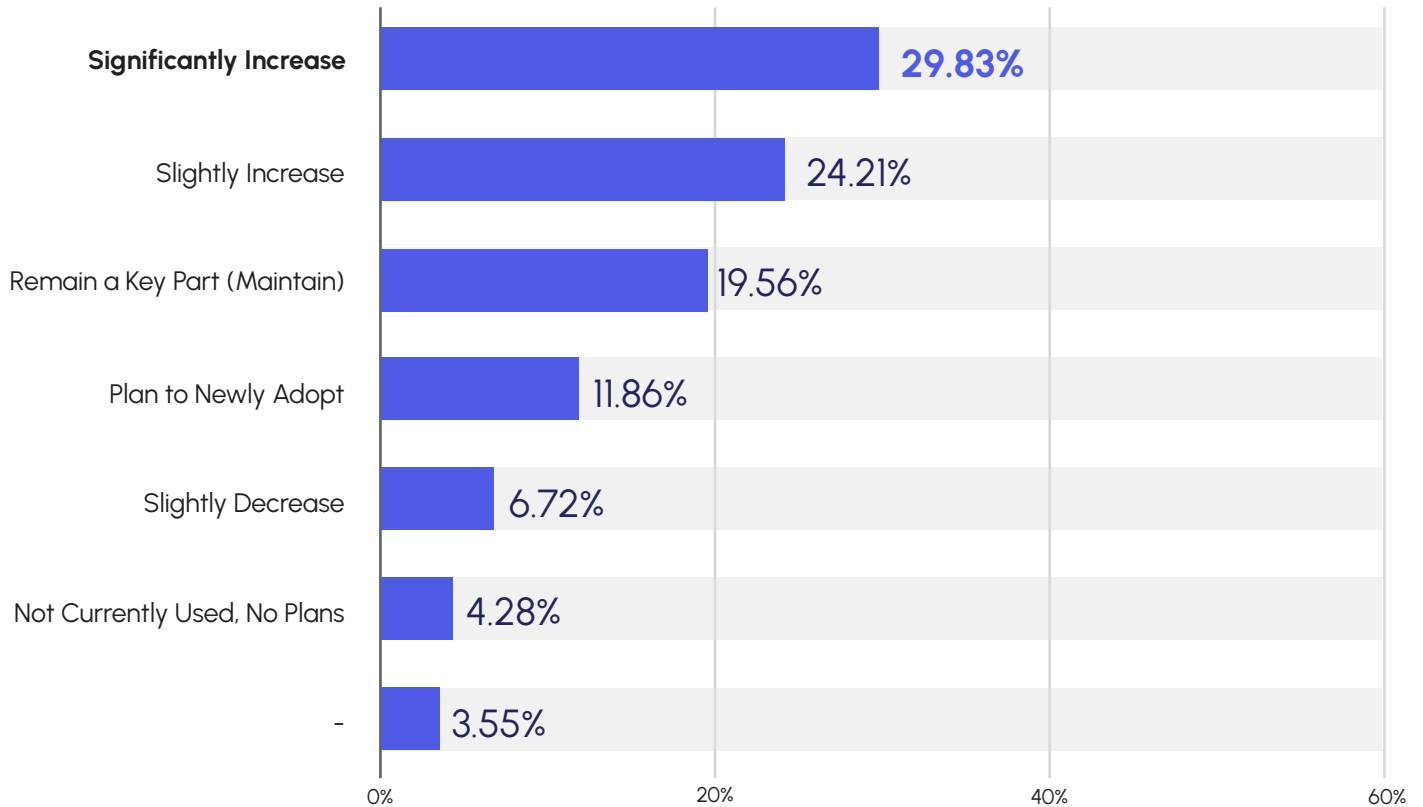
Source: Futurum Research, IH 2026 Data Management Decision Makers Survey (n=818).

This friction is widespread. According to the IH 2026 DIAI Market Sizing & Forecast Report, 18% of organizations rate keeping pace with rapid data growth as their single biggest internal challenge, while 44% cite raw capacity growth and complexity as the primary driver forcing net-new infrastructure purchases.

Enterprises face a funding paradox. Executive boards demand immediate AI progress – with 51% prioritizing increased investment in generative and agentic tools for 2026¹– yet IT budgets buckle under the weight of sprawling legacy footprints. Storage maintenance threatens to consume available capital before businesses can demonstrate a return on investment (ROI) from scaled production GenAI deployments. As unstructured datasets for training, retrieval-augmented generation (RAG), and vector representation expand, enterprises must reallocate infrastructure spend (see Figure 1).

¹ Futurum Research, IH 2026 Artificial Intelligence Decision Makers Survey

Figure 1. Data Storage Spend Growth: Surviving the AI Storage Squeeze



With nearly 66% of organizations actively increasing storage budgets or adopting new storage solutions, the influx of unstructured datasets for AI training and vector representation is straining traditional infrastructure, underscoring the need for automated tiering and cost optimization.

Source: Futurum Research, IH 2026 Data Management Decision Makers Survey (n=818).

As organizations scale AI deployments, they quickly learn that the foundational data layer dictates the economic viability of the entire project. Stale, unmanaged data degrades model context and retrieval quality, while bloated storage consumes the capital needed for GPU compute. High-performing enterprises avoid this budgetary gridlock by shifting tactics: rather than requesting net-new funding, IT leaders audit their environments to harvest savings from inactive "cold" data, actively funding AI ambitions directly out of optimized infrastructure.



Where Data Meets Dollars

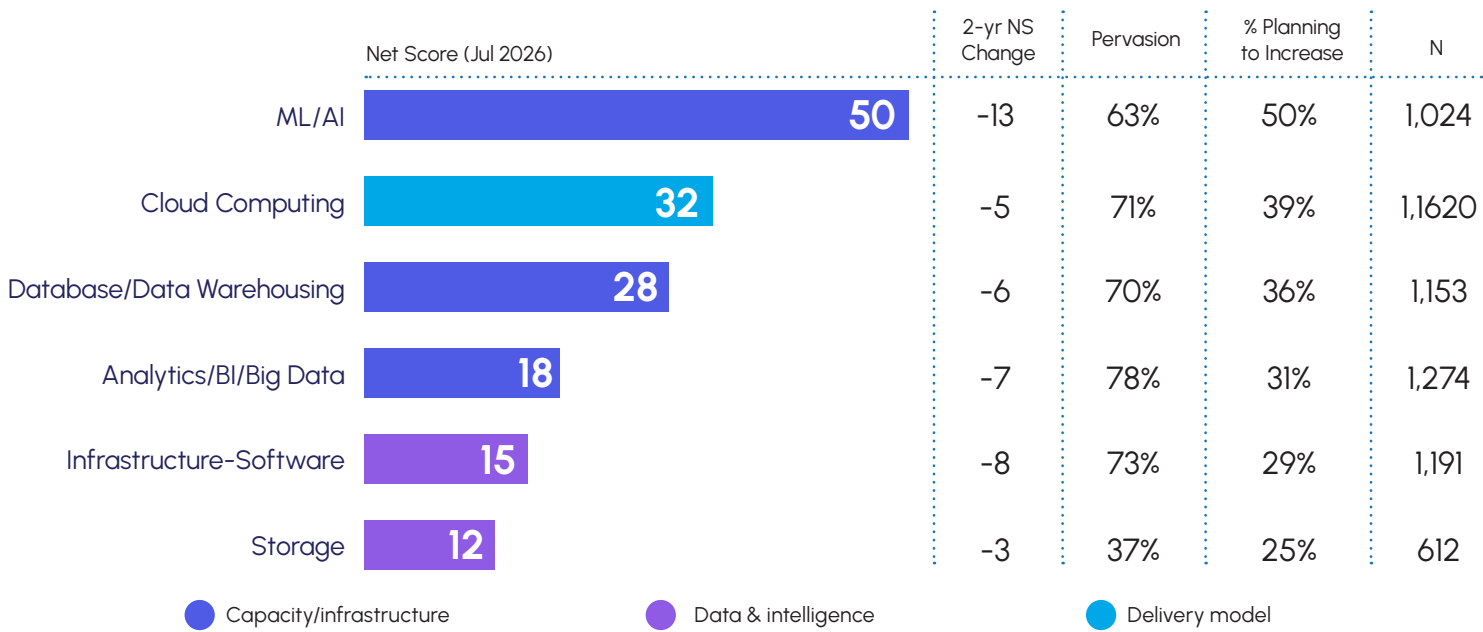
An analysis of 16 petabytes of telemetry data gathered by CTERA across enterprise NAS environments reveals a clear pattern: primary storage arrays overwhelmingly act as overpriced long-term archives.

- 83.1% of total capacity (and 97.5% of individual files) has not been modified in over a year.
- 88.3% of files have sat entirely untouched for more than 12 months.
- The active working set (data modified within the last 90 days) accounts for just 4.4% of total enterprise capacity.

Paying Tier 1 primary storage prices, power, and cooling for dormant data drains capital directly from inference engines and GPU clusters.

Enterprise AI strategy is now an exercise in infrastructure economics. Futurum Research indicates that 54% of organizations plan to increase data management purchases², pivoting toward optimizing costs and inference performance rather than buying raw hardware capacity. Spending patterns across the ecosystem mirror this shift. ETR's July 2026 Technology Spending Intentions Survey (TSIS) shows enterprise spend moving up the stack toward higher-value intelligence layers (see Figure 2).

Figure 2. Enterprise Spend Intention Rises Up the Stack

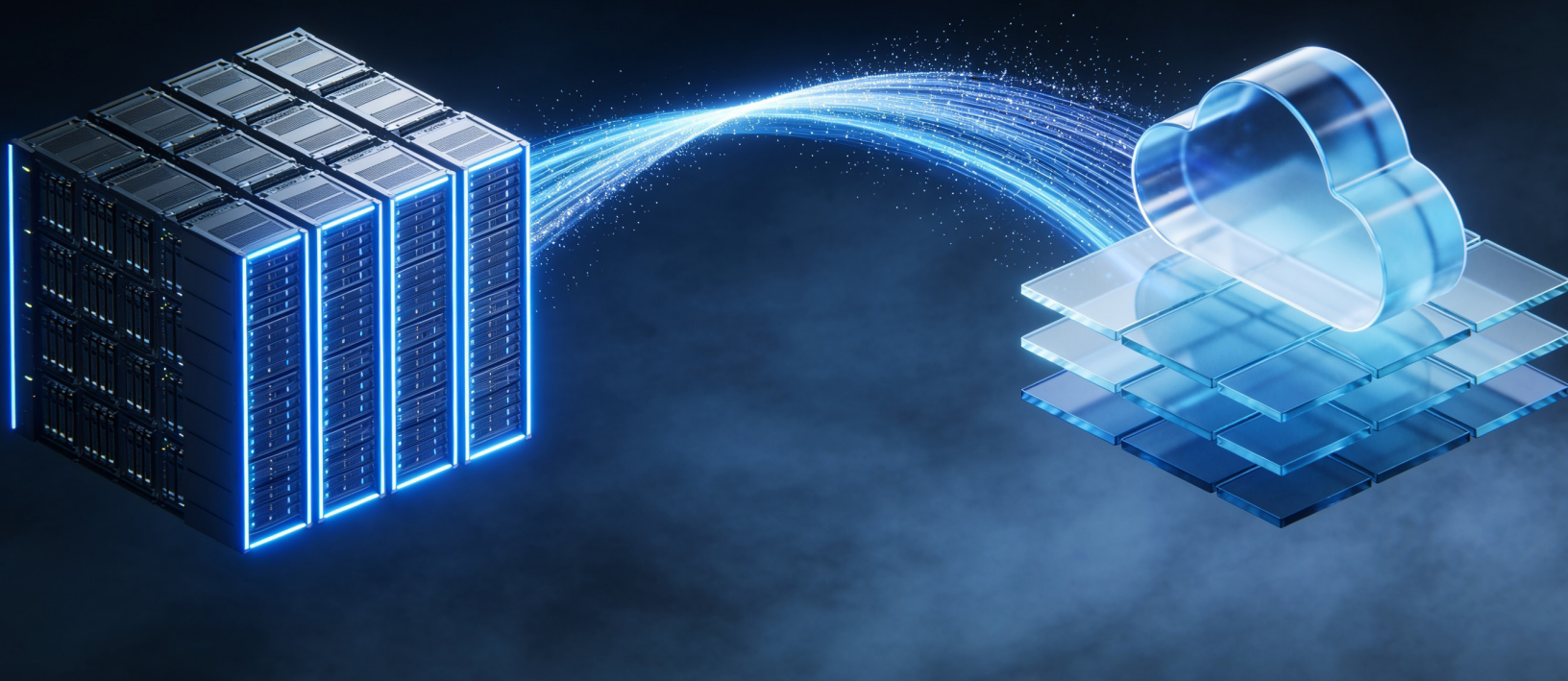


TSIS Net Scores place the commodity storage capacity layer at the bottom (Net Score 12), while the data-and-intelligence layer commands the top (ML/AI at 50). Enterprises are holding raw capacity spend flat and steering dollars toward AI and data platforms. "Pervasion" reflects vendor account penetration within the sample.

Source: ETR Technology Spending Intentions Survey (TSIS), July 2026 wave. Net Score = (Adopting + Increasing) – (Decreasing + Replacing).

Optimizing storage yields a critical secondary benefit: data hygiene. Cross-cut survey analysis shows that organizations investing in data observability, FinOps, and lifecycle management are **28% more effective** at keeping AI projects on schedule and avoiding bottlenecks. Infrastructure cost optimization self-funds the underlying storage layer while simultaneously cleaning the data supply chain feeding the AI models.

2 ETR Insights 466: Hardware Purchasing Drill Down; Futurum Research, 1H 2026 Data Management Decision Makers Survey



Avoiding the Architectural Squeeze

Current infrastructure squeezes enterprise IT from two directions:



Legacy hardware lock-in:

On-premises storage vendors push vertically integrated arrays that promise raw performance but tie up capital in depreciating assets.



Cloud lock-in & egress fees:

Proprietary cloud platforms capture the metadata layer, trapping customers in consumption pricing and heavy network egress fees when moving data to external AI compute engines.

These cost pressures are forcing strategic adjustments. According to Futurum Research's 2026 Key Issues & Predictions report, 71% of CIOs are reevaluating cloud workload placement due to high AI processing costs and data gravity constraints.

To maintain operational control, IT organizations need agility in the middle tier. They must manage unstructured data across core, edge, and cloud without locked-in pricing models or degraded performance. With 51% of data solutions now deployed in public cloud compared to 30% on-premises³, hybrid flexibility is mandatory. Enterprises require a unified data fabric spanning these environments, keeping files governed, cached, and accessible regardless of where compute runs.

³ Futurum Research, 2026 Data Management Decision Makers Survey



The Cost-Optimization Playbook

To break free from the storage squeeze and fund AI initiatives at scale, IT organizations must move away from reactive hardware purchasing and execute a definitive modernization playbook. Modernizing the storage tier strips out operational complexity, consolidates fragmented file silos, and frees up capital immediately.

Play 1: Consolidate & Tier with Intent

Legacy environments are often fragmented across isolated arrays, branch file servers, and cloud buckets – each adding licensing and management overhead.

Modern enterprise architecture unifies these silos into a single global data fabric. Implementing a global file system creates a centralized management layer across edge sites, core data centers, and public clouds. This allows IT to automatically tier inactive cold data (over 80% of capacity) to low-cost object storage. Because the global namespace remains unified, users and AI agents retain transparent file access without breaking application paths or requiring manual restores.

Play 2: Immunize the Data Fabric & Build Data Resilience

Granting AI models and autonomous agents read/write access to corporate file systems changes the risk profile. Perimeter security alone is insufficient when automated agents crawl enterprise data shares. With 80% of CIOs citing data security and privacy as primary AI concerns, storage must shift from passive retention to active defense – a model known as Data Resilience.

Securing the data fabric requires embedding defenses directly into the storage tier. Solutions such as the CTERA Intelligent Data Platform incorporate real-time behavioral analytics to block ransomware, paired with immutable Write-Once-Read-Many (WORM) retention policies and zero-trust access controls. This creates a secure environment where organizations can safely expose corporate files to AI models without risking unauthorized alterations or exposure.

Play 3: Illuminate and Classify at the Source

Stale, redundant, or sensitive files locked in unmanaged arrays create operational and compliance risks if ingested into LLMs or RAG pipelines.

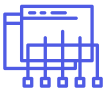
Leading organizations integrate classification and real-time discovery directly into the storage layer. Classifying unstructured files at the source isolates cold data for tiering while tagging sensitive information to exclude it from AI training sets. Enriching unstructured files with metadata tags and vector embeddings gives RAG pipelines precise business context, improving output accuracy and reducing hallucinations.

Play 4: Bring Compute to the Data

Shuttling petabytes of unstructured data across networks to cloud AI engines causes latency, operational friction, and high egress bills. System deployment complexity remains a primary reason enterprise AI projects stall.

Rather than moving datasets to remote compute engines, hybrid architectures bring compute closer to where data lives. Using an edge-to-cloud file caching architecture lets local processing engines and remote sites access cloud-resident data at LAN speeds. This eliminates wide-area network bottlenecks and avoids egress penalties, keeping agent feedback loops sub-second (see Table 1).

Table 1: Problem/Solution Matrix – The AI Storage Squeeze Playbook

Enterprise Pain Point	Architectural Solution
 High TCO of Siloed Storage & Inactive Data	Global File System & Tiering: Unifies data across core, edge, and cloud into a single namespace. Automatically stores data in low-cost object storage without altering file paths or changing protocols.
 Integration Latency & Network Egress Costs	Edge-to-Cloud Caching: Delivers LAN-speed access for local AI workloads via edge caching, eliminating cloud network egress fees.
 Data Privacy & AI Agent Governance	Data Classification: Embeds real-time file classification and metadata tagging at the source to discover, govern, and isolate sensitive data before AI ingestion.
 Ransomware & Pipeline Fragility	Cyber Protection & Data Resilience: Combines real-time behavioral threat analytics with immutable WORM snapshots to stop attacks instantly and ensure zero-loss pipeline recovery.

Aligning technical architecture with cost drivers allows organizations to fund AI innovation directly through infrastructure savings.



Funding the Future

Organizations are evaluating AI initiatives based on business outcomes rather than model novelty. Futurum Research shows that 55.1% of organizations measure AI success through productivity gains, while 50.6% target direct cost reductions.³ Yet, achieving these returns remains difficult: only 26.5% of enterprises report AI returns that exceed initial expectations.⁴

Operational efficiency is the primary lever for improving AI ROI. Optimizing the storage layer directly targets this gap. Moving cold data off expensive hardware frees up capital for compute. Simultaneously, edge caching and active governance ensure AI models ingest clean, secure, highly available data at network speeds.

IT leaders must look at their storage footprint through an economic lens. Auditing unstructured data estates, deploying a unified global file system, and enforcing proactive data resilience builds a predictable foundation that funds future AI initiatives.

Important Information About This Report

Authors

Brad Shimmin

Vice President & Practice Lead, Data Intelligence,
Analytics, & Infrastructure | The Futurum Group

Publisher

Futurum Research

Inquiries

Contact us if you would like to discuss this report and
The Futurum Group will respond promptly.

Citations

This paper can be cited by accredited press and analysts,
but must be cited in context, displaying author's name,
author's title, and "The Futurum Group." Non-press and
non-analysts must receive prior written permission by The
Futurum Group for any citations.

Licensing

This document, including any supporting materials, is
owned by The Futurum Group. This publication may not be
reproduced, distributed, or shared in any form without the
prior written permission of The Futurum Group.

Disclosures

The Futurum Group provides research, analysis, advising,
and consulting to many high-tech companies, including those
mentioned in this paper. No employees at the firm hold any
equity positions with any companies cited in this document.



ABOUT CTERA

CTERA empowers organizations to stop simply storing data and start activating its full value for analytics and AI. As unstructured data grows fragmented across edge locations, offices, and cloud, it becomes harder to govern, protect, and use. The CTERA Intelligent Data Platform unifies that data into a single data fabric spanning edge to cloud. This modernizes legacy infrastructure, reduces the operational footprint, and lowers data ownership costs, while keeping data resilient, sovereign, and secure. Consistent governance applies everywhere data lives, and AI and analytics engines access it without moving data, turning data from a cost center into an engine for innovation. To learn more visit www.ctera.com.

Futurum

About The Futurum Group

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum

CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337