CLOUDIAN®

AI Data Platform Buyer's Guide

The technical and business benefits of
on-premises AI infrastructure and guide to solutions
for your enterprise environment.

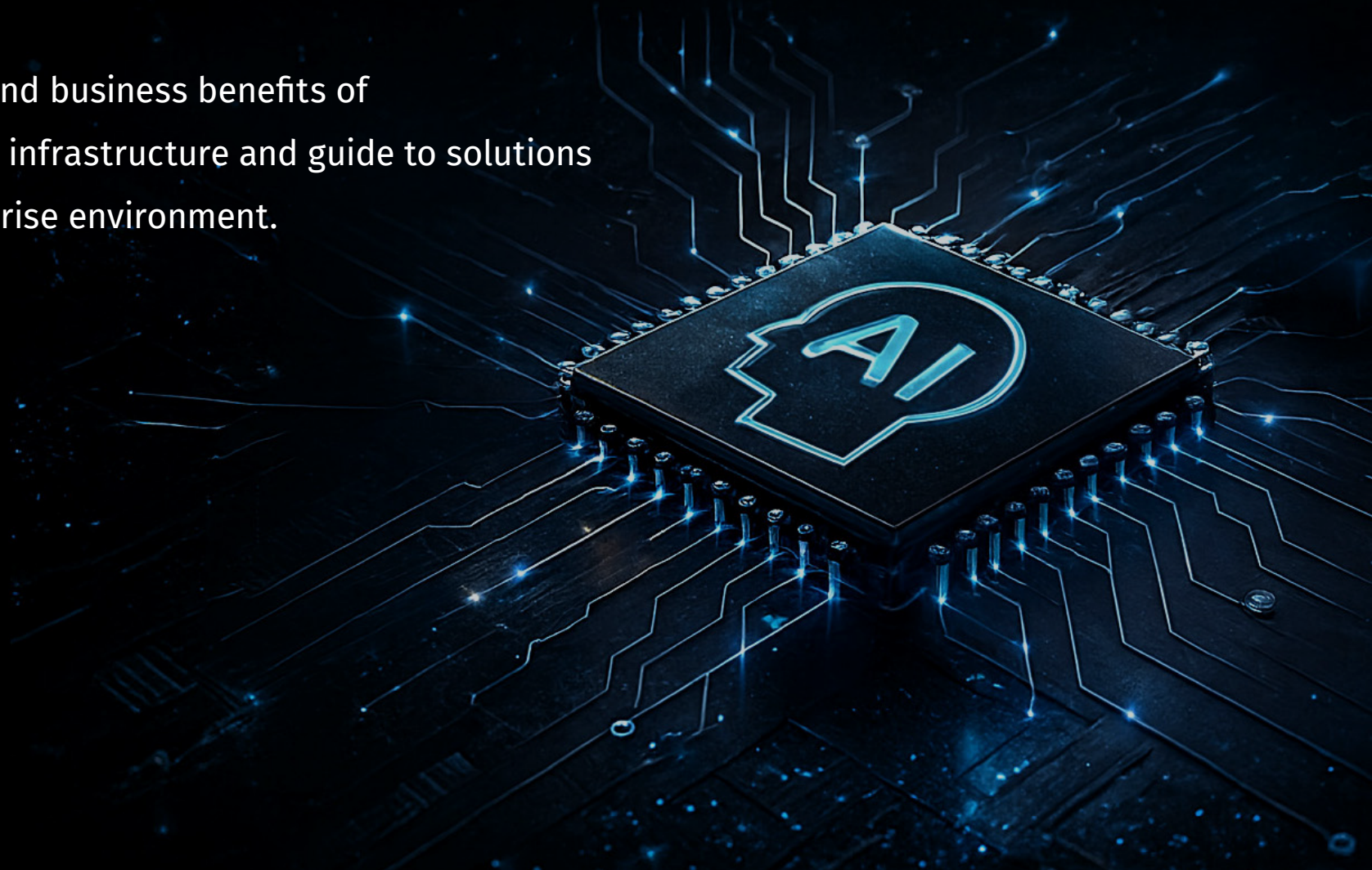


Table of Contents

Introduction

Artificial intelligence is no longer emerging technology—it is reshaping how enterprises operate, compete, and create value. According to McKinsey's 2025 State of AI survey, 78% of organizations now deploy AI in at least one business function, while 92% plan to increase AI investment over the next three years. The enterprise AI market, valued at $97 billion in 2025, is projected to reach $229 billion by 2030.

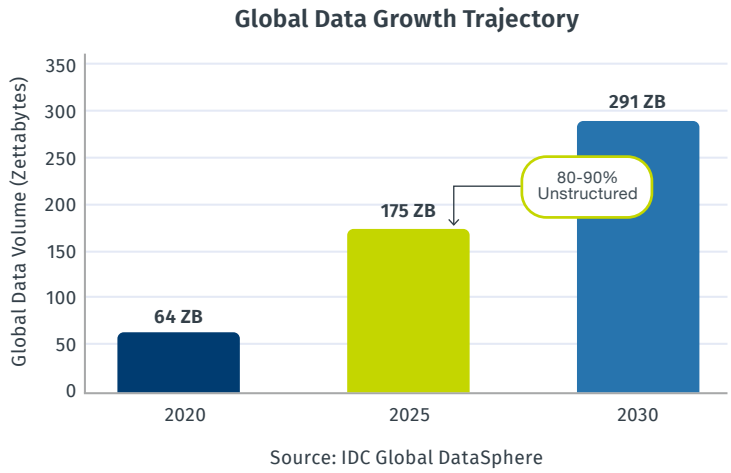
Yet despite this investment, most enterprises find themselves caught between two imperfect choices: cloud-based AI services that compromise data control and incur unpredictable costs, or complex on-premises solutions that require scarce AI expertise and carry high implementation risk.

The stakes are enormous. Global data volumes reached 175 zettabytes in 2025, with 80-90% of all new enterprise data being unstructured—documents, videos, images, and operational records that represent decades of institutional knowledge. This unstructured data is growing at 55-65% annually, yet only 18% of organizations report being able to leverage it effectively. Traditional storage architectures simply lack the semantic understanding, vector indexing capabilities, and GPU-optimized I/O paths that modern AI workloads demand.

That's where integrated AI data platforms come in.

These purpose-built solutions combine GPU compute, enterprise storage, AI software, and pre-validated applications into turnkey systems that deliver business value from day one—without requiring extensive AI expertise or compromising data sovereignty.

80-90% of enterprise data is unstructured—and only 18% of organizations can leverage it effectively.



In this buyer's guide, you'll learn about today's enterprise AI challenges and how both cloud and traditional on-premises approaches fall short. You'll discover how integrated AI data platforms provide unique capabilities to address these challenges without compromise. And you'll understand how Cloudian HyperScale AI Data Platform can fundamentally transform your organization's ability to leverage AI while maintaining complete control over your data.

1: The Enterprise AI Imperative

Public AI systems like ChatGPT and Claude have demonstrated remarkable capabilities, but they're trained on public internet data. While powerful for general knowledge queries, they know nothing about your specific business operations, proprietary processes, or accumulated institutional knowledge.

Enterprise Data Delivers Enterprise AI Value

The true value of AI for enterprises lies in leveraging proprietary data: customer records, operational insights, product documentation, financial analyses, and institutional knowledge accumulated over years or decades. This internal data holds the key to unlocking productivity gains that generic AI cannot deliver.

Consider what becomes possible when AI can access your enterprise data: employees stop spending hours searching for information across disparate systems. Customer service representatives instantly retrieve relevant product documentation and support history. Engineers access decades of design decisions and lessons learned. Finance teams synthesize insights from thousands of contracts and agreements.

Retrieval-Augmented Generation (RAG) makes this possible by grounding AI responses in your actual enterprise content. Rather than relying solely on pre-trained knowledge, RAG systems retrieve relevant documents, policies, and records at query time—then synthesize accurate, context-aware responses backed by your authoritative sources. This approach dramatically reduces AI hallucinations while ensuring responses reflect your current information, not outdated training data.

RAG grounds AI in your actual enterprise content—dramatically reducing hallucinations while ensuring responses reflect current information

The Data Sovereignty Imperative

The challenge is that enterprise data is often highly sensitive. Making it accessible to public cloud-based AI creates both internal and external obstacles. Within the organization, data governance rules and management approval chains can block progress at any step.

Outside the organization, data privacy regulations impose strict requirements. More than 150 countries now have data privacy regulations, including GDPR in Europe, CCPA in the United States (CA), PIPL in China, and the EU AI Act taking effect in 2025. Industries from healthcare (HIPAA) to financial services (SOX, DORA) face compliance requirements that often prohibit sending sensitive data to external cloud services.

> **"Data privacy (57%) and trust concerns (43%) are the biggest inhibitors of generative AI adoption, outranking skills gaps and implementation costs."**
>
> - IBM

The Implementation Reality

Even when organizations understand the value of enterprise AI on-prem, implementation remains extraordinarily difficult. According to RAND Corporation research, more than 80% of AI projects fail—double the failure rate of IT projects that don't involve AI. A 2025 S&P Global study found that 42% of companies abandoned most of their AI initiatives, up sharply from 17% the prior year, with the average organization scrapping 46% of AI proofs-of-concept before reaching production.

McKinsey research confirms this pattern: while 78% of organizations are using AI, only 1% of executives consider their AI rollouts mature. AI talent is scarce and expensive—data scientist roles are projected to grow 34% through 2034. Custom AI development projects frequently exceed budgets, and when key personnel leave, institutional knowledge about complex AI implementations often departs with them.

***Key Insight:** Pre-validated solutions eliminate the complexity that causes most AI projects to fail.*

The AI Implementation Challenge

80%	42%	Only 1%
of AI projects fail	abandoned AI initiatives in 2025	consider AI rollouts mature
— RAND Corporation	— S&P Global	— McKinsey

2: Why Public AI Falls Short

Public cloud-based AI services offer convenience and rapid deployment, but they introduce significant challenges that often prove insurmountable for regulated industries and security-conscious organizations.

Data Residency and Security Risks

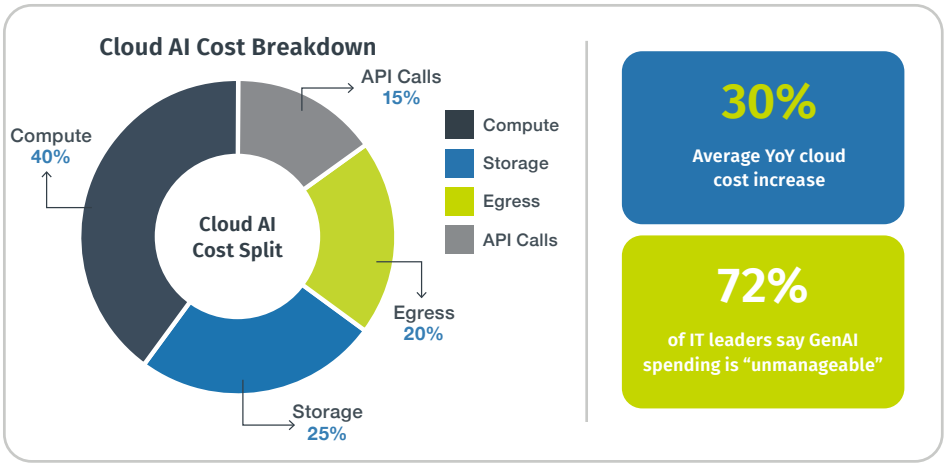
When you use public cloud AI services, your data must traverse external networks to reach the cloud providers' infrastructure. Once in the public cloud, that data resides on shared infrastructure that may be accessible to third parties. While cloud security is generally robust, misconfigured storage buckets and leaked credentials cause frequent breaches. A 2025 study found that 69% of organizations cite AI-powered data leaks as their top security concern, yet nearly half have implemented no AI-specific security controls.

> "My financial customers cannot take ANY customer data to the cloud."
>
> - SERVICE PROVIDER CEO

For enterprises in financial services, healthcare, government, and defense, these factors create unacceptable compliance and security risks. As one service provider CEO noted: "My financial customers cannot take ANY customer data to the cloud." The shift toward sovereign AI, where organizations maintain complete control over data, infrastructure, and AI models, reflects this growing concern.

Escalating Costs and Unpredictable ROI

Cloud AI pricing has become a critical concern for enterprises. A 2024 Tangoe study found that enterprise cloud costs rose an average of 30% year-over-year, with AI and generative AI cited as top drivers. More troubling, 72% of IT and financial leaders said GenAI-led cloud spending had become "unmanageable."



Cloud AI pricing compounds unpredictably: charges accumulate across data storage, data transfer, API calls, compute time, and premium features. Organizations frequently report costs increasing 5-10x within months of moving from pilot to production. Data egress fees alone—$0.09/GB for cross-regional transfers on major platforms—escalate quickly for AI workloads that require frequent access to large datasets.

72% of IT leaders say GenAI-led cloud spending has become "unmanageable."

Privacy and Transparency Concerns

Public cloud AI services often use customer data to improve their models unless explicitly opted out. While "enterprise plans" promise enhanced confidentiality, even then audit trails and privacy verification can be challenging. For organizations handling sensitive customer information, intellectual property, or classified data, this opacity creates unacceptable risk. The EU AI Act and similar regulations increasingly require organizations to demonstrate exactly how their AI systems process data and make decisions.

Vendor Lock-In

Public AI services typically use proprietary APIs and data formats that make migration between providers difficult and expensive. Once you've invested in training models and building integrations with a specific cloud provider, switching costs become prohibitive—creating long-term dependency on a single vendor's pricing decisions and service availability. This dependency becomes particularly acute as AI becomes embedded in critical business processes.

The Shadow AI Threat

Finally, when organizations lack a clear AI strategy, another threat emerges: employees don't wait—they find their own solutions. About 38% of employees share confidential data with AI platforms without approval, according to late 2024 research by CybSafe and the National Cybersecurity Alliance. This phenomenon, known as "shadow AI," occurs when workers turn to consumer-grade tools like ChatGPT to boost productivity, often without understanding the consequences.

Many generative AI tools store user inputs on platform memory to improve their models, meaning the AI provider could retain and access sensitive corporate data. Once proprietary information—source code, customer data, strategic plans, financial projections—is submitted to a public AI service, it may be used to train future models and could surface in responses to other users.

According to IBM's 2025 Cost of a Data Breach Report, breaches involving shadow AI cost organizations $4.63 million on average—$670,000 more than standard incidents. The risk compounds because 83% of organizations lack technical controls to detect or prevent employees from uploading confidential data to AI platforms. Without sanctioned, secure AI alternatives, well-intentioned employees become unwitting vectors for data exposure.

> ***38% of employees share confidential data with AI platforms without approval. Shadow AI breaches cost $670,000 more than standard incidents.***

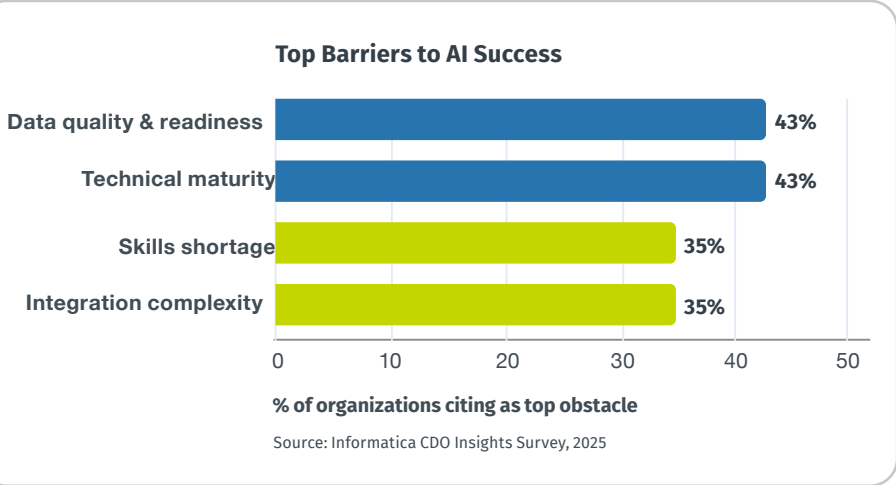
 |

3: The On-Premises Challenge

While on-premises deployment addresses data residency concerns, traditional approaches to building enterprise AI infrastructure present their own significant obstacles.

Implementation Complexity

Building enterprise AI infrastructure from scratch requires expertise across multiple domains: GPU computing, vector databases, embedding models, retrieval systems, and application development. These skills are scarce, expensive, and in high demand. Informatica's 2025 CDO Insights survey identified the top obstacles to AI success: data quality and readiness (43%), lack of technical maturity (43%), and shortage of skills (35%). Many organizations simply cannot attract or retain the specialized talent needed to build and maintain custom AI systems.



File storage does not scale well for very large volumes of unstructured data and is often difficult to manage for massive AI repositories.

Project Risk

Custom AI development projects carry substantial risk. Beyond the 80%+ failure rate, an MIT study of generative AI pilots found that only 5% achieved rapid revenue acceleration, with the vast majority stalling and delivering little measurable impact. When projects fail or key personnel depart, investments in custom development may be lost entirely. The complexity of maintaining compatibility across GPUs, storage, networking, and AI software frameworks creates ongoing operational burden that diverts resources from core business objectives.

Storage Architecture Limitations

Enterprises traditionally add an unnecessary file layer for AI, in addition to a datalake layer that acts as a long-term data repository. This multi-tiered approach makes scaling difficult due to its inherent complexity and the increasing need to migrate data among various stores.

Furthermore, traditional file storage systems were not designed for AI workloads. They lack native support for vector indexing, struggle to scale to the massive datasets AI requires, and can introduce security vulnerabilities through kernel modifications. File-based storage also requires translation layers to work with modern AI frameworks that are increasingly S3-native, adding complexity and reducing performance.

Integration Burden

Assembling best-of-breed components—GPUs, storage, networking, AI software—requires significant integration effort. Ensuring optimal interoperability, maintaining compatibility through upgrades, and troubleshooting across multiple vendors creates ongoing operational burden. When problems arise, each vendor points to others, leaving IT teams caught in the middle.

4: The Solution: An Integrated AI Data Platform

Integrated AI data platforms offer a fundamentally different approach to enterprise AI: turnkey solutions that combine all hardware, software, and support into a unified stack. These platforms integrate NVIDIA-accelerated computing with enterprise storage to centralize intelligent data handling and deliver AI-ready data—while reducing latency, enhancing security, and maximizing performance.

Real-Time and Accurate Insights

AI data platforms transform enterprise storage from passive repositories into intelligent systems. Continuous indexing of multimodal data enables hybrid search and retrieval, ensuring AI applications operate on the most recent and relevant information. GPU-accelerated vector embeddings enable natural language queries across all content types—documents, charts, PDFs, videos—providing context-aware responses rather than simple file retrieval.

All Data, AI-Ready

Previously untapped unstructured data becomes AI-ready through intelligent data handling that discovers, orchestrates, and connects information across agentic applications. Rather than requiring data scientists to manually prepare datasets, integrated platforms automate the ingestion, embedding, and indexing processes that make enterprise content accessible to AI workflows.

Turnkey Deployment

Leading enterprise storage vendors now integrate NVIDIA accelerated computing, networking, and AI software into turnkey AI data platforms. Pre-validated configurations eliminate the integration complexity that derails most custom AI initiatives, and production-proven applications deliver value from day one.

From passive repository to intelligent system that accelerates AI workflows.

Data Protection and Compliance

Zero-trust architecture and granular security policies safeguard enterprise data across every node. Encryption is accelerated for data at rest and in transit. Threats are detected in real time. Full audit trails support regulatory compliance across GDPR, HIPAA, SOX, and industry-specific frameworks. Air-gapped deployment capability addresses the most security-sensitive environments.

 |

5: Cloudian HyperScale® AI Data Platform

Cloudian's HyperScale AI Data Platform delivers fully integrated, on-premises AI infrastructure that transforms storage from passive repositories into intelligent systems capable of semantic understanding and natural language interaction. This complete solution is production-ready out of the box—no extensive AI expertise required, and no compromise on data control.

Storage that understands your data. Security that protects it. AI infrastructure you control completely.

Storage That Understands Your Data™

HyperScale AIDP represents a fundamental shift: from systems that merely store and retrieve data to infrastructure that comprehends it. Ask questions about your documents through a familiar chatbot interface. The system learns from your documents, charts, PDFs, and multimedia content to provide intelligent, context-aware responses rather than simple file listings.

Complete AI Infrastructure

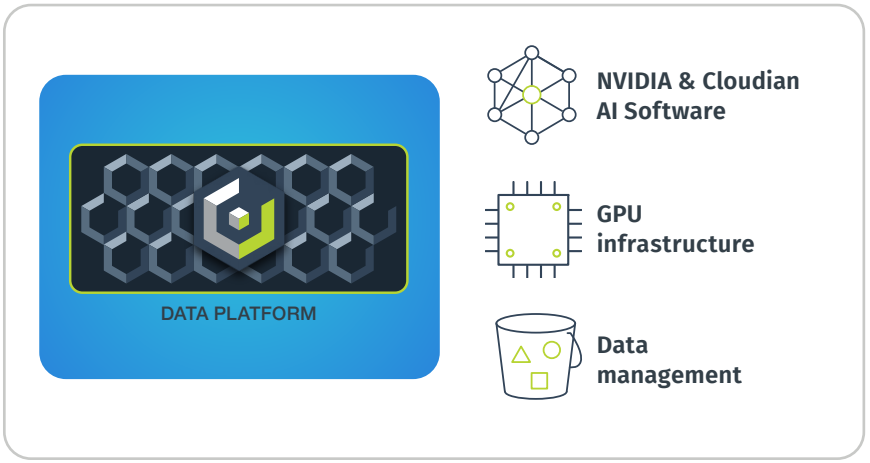
The platform combines GPU compute, enterprise storage, NVIDIA AI Enterprise software, and production-validated AI blueprints in one integrated appliance. Built on NVIDIA's enterprise AI platform with RTX PRO 6000 Blackwell GPUs and Cloudian's proven HyperStore architecture, HyperScale AIDP eliminates the integration complexity that derails most enterprise AI initiatives.

Sovereign AI: Complete Control

Your data never traverses external networks. Your RAG relies exclusively on your data. Your intellectual property remains under your complete administrative control. The platform operates entirely within your perimeter with zero external API dependencies—enabling air-gapped deployment for the most security-sensitive environments.

Proven Partnership with NVIDIA

HyperScale AIDP is the result of two years of joint engineering between Cloudian and NVIDIA. Cloudian delivers the S3-compatible object storage that AWS trusts, along with management tools, integration, testing, and white glove support, while NVIDIA provides AI leadership. Together, they deliver a solution validated for enterprise production workloads. NVIDIA recently announced general availability of RDMA for S3-compatible storage, with Cloudian among the first partners to integrate this breakthrough capability.



6: Architecture and Components

Cloudian HyperScale AIDP combines multiple technologies in a single, integrated system designed for enterprise AI workloads.

Integrated Solution Stack

GPU Server

4x or 8x NVIDIA RTX PRO 6000 Blackwell GPUs with dual Intel Xeon 6952P processors provide massive parallel processing power for AI workloads including vector embedding generation, model inference, and real-time video analysis.

High-Performance Networking

NVIDIA Spectrum switches with RDMA support deliver ultra-low latency data transfer between GPUs and storage, eliminating bottlenecks that would otherwise limit AI performance.

Cloudian HyperStore Storage

Exabyte-scalable, S3-compatible storage with NVMe media and S3-over-RDMA acceleration provides the high-throughput, low-latency data access that AI workloads demand. Cloudian pioneered S3 RDMA with NVIDIA to bypass CPU overhead, achieving 35GB/s reads per node with linear scalability across clusters.

AI Software Stack

NVIDIA AI Enterprise software, Cloudian AI data management tools, GPU-accelerated vector database, pre-configured AI application blueprints, semantic search, data ingestion tools, and an intuitive user interface.

HyperCare Support

Every software subscription includes 24x7 fully managed service with remote monitoring and expert guidance, backed by Cloudian's NVIDIA enterprise support agreement.

 |

NVIDIA AI Enterprise Blueprints

Rather than experimental development, HyperScale AIDP integrates NVIDIA's production-validated blueprints—pre-configured workflows representing extensive engineering and validation. You're not buying experimental AI; you're getting applications that NVIDIA engineers have tested and Cloudian engineers have integrated to create a highly-capable platform.

Initial blueprints include Enterprise Document RAG and Video Search and Summarization. Future releases will include Real-Time Streaming Data Processing, Healthcare Diagnostic Agents with HIPAA compliance, Financial Fraud Detection, and 30+ additional industry-specific workflows.

Scalable Architecture

HyperScale AIDP features a disaggregated hardware design that allows independent scaling of GPU processing and storage capacity. Start small with a starter configuration, then expand as your AI initiatives evolve.

Config	Data Ingest	Query Rate	GPU Servers	GPUs	Storage
Starter	Up to 50GB/day	Up to 5 QPS	1	4	Single node
Medium	50-200GB/day	5-10 QPS	1	8	6-node cluster
Large	200-300GB/day	10-20 QPS	3	12	6-node cluster
Custom	TBs/day	1000s QPS	Custom	100s+	100s nodes

Cloudian HyperCare Benefits

Every HyperScale AIDP solution includes these Cloudian HyperCare support benefits:

- **Fully managed service** - Remote 24×7 management of Cloudian storage by dedicated experts, eliminating day-to-day operational tasks
- **On-premises control** - Provides cloud-like consumption experience while keeping data behind your firewall and under your control
- **Complete lifecycle management** - Includes monitoring, upgrades, expansions, incident and change management, and best practices optimization
- **Proactive oversight** - Weekly performance reporting and quarterly reviews with Cloudian Trusted Advisor to assess capacity trends and future planning
- **Always current** - Ensures systems stay up-to-date with latest features, patches, and software updates
- **Skills shortage solution** - Addresses IT staffing challenges by leveraging Cloudian's storage expertise instead of hiring internal specialists

Start small, scale smart: From pilot to petabytes—grow seamlessly on one platform.

7: Why Object Storage for AI?

Cloudian is a recognized leader in S3-compatible object storage—the de facto standard for large-scale storage in the public cloud. Since 2011, Cloudian has equipped enterprises with this technology on-premises, building the ideal foundation for enterprise AI.

Massive Scale Without Complexity

Cloudian object storage scales to exabytes with a unified namespace, eliminating the complexity of managing multiple file systems or storage silos. AI workloads can access all data through a single, consistent interface regardless of scale—critical when unstructured data is growing at 55-65% annually.

Native S3 Compatibility

The most popular vector databases and AI frameworks—including TensorFlow, PyTorch, Apache Spark, and Milvus—are S3-native. With the industry's highest level of S3 API compliance, Cloudian delivers seamless integration without the translation layers that file-based solutions require.

TensorFlow, PyTorch, Spark, Milvus—the most popular AI frameworks are S3-native. Cloudian delivers the industry's highest S3 API compliance.

Embedded Metadata

Rich metadata is stored and indexed with the object itself, simplifying data management and enabling sophisticated search and retrieval capabilities that AI workloads depend on. This metadata architecture supports the semantic understanding that makes RAG applications effective.

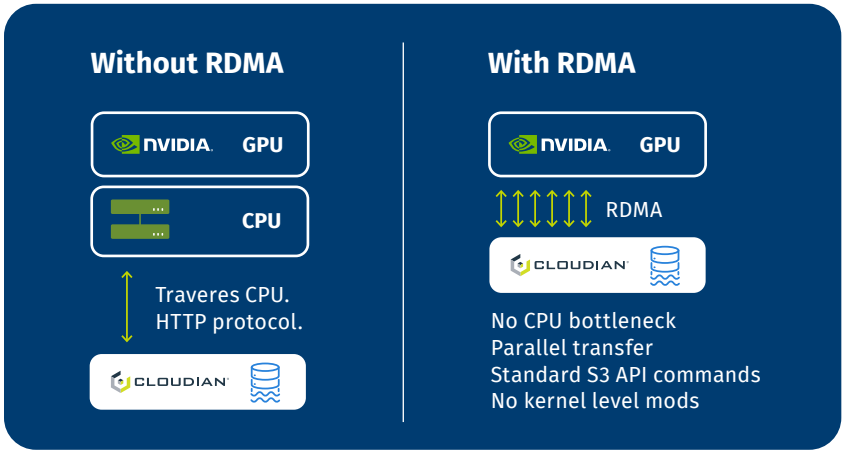
Secure by Design

Object storage requires no kernel modifications, eliminating security vulnerabilities that file-based storage solutions can introduce. This is critical for enterprises in regulated industries where security audits scrutinize every component of the infrastructure stack.

 |

RDMA for S3 Compatible Storage: Transformational AI Storage Performance

Traditional storage architectures rely on TCP-based data transfer, which creates a significant bottleneck for AI workloads. Every data request must pass through the host CPU and system memory before reaching the GPU—adding latency and consuming compute resources that should be dedicated to AI processing.

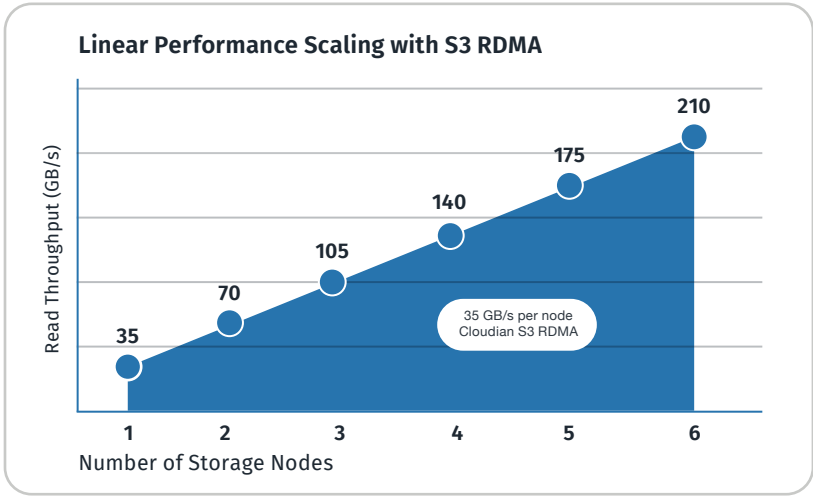


Cloudian collaborated with NVIDIA to pioneer RDMA (Remote Direct Memory Access) for S3-compatible storage, revolutionizing how data moves between storage and GPUs. With NVIDIA GPUDirect integration, data flows directly from Cloudian object storage to GPU memory—bypassing the CPU entirely and eliminating traditional storage I/O bottlenecks.

Real-world testing demonstrates dramatic performance gains:

- **35 GB/s** throughput per node (reads), with linear scalability across clusters
- **3-5x** throughput improvement compared to conventional TCP-based object storage
- Up to **90%** reduction in CPU utilization by establishing direct data pathways to GPUs
- **8x** performance boost in vector database operations (demonstrated with Milvus using NVIDIA cuVS)
- **Scalability to TBs/s** with Cloudian's parallel processing architecture

Cloudian HyperStore is the world's first production-ready storage platform to deliver RDMA for S3 over Converged Ethernet (RoCE). A 6-node HyperStore cluster delivers sustained data transfers of 210 GB/s, with linear performance scaling as additional nodes are added. Testing with Cloudian and NVIDIA reduced Milvus vector database indexing time from 2 hours to just 16 minutes—enabling real-time data ingestion and query readiness for large-scale AI applications.



Cloudian pioneered RDMA for S3 with NVIDIA—delivering 35GB/s reads per node, 90% CPU reduction, and 8x faster vector database operations.

 |

8: Business and Technical Benefits

HyperScale AIDP delivers compelling value through cost predictability, risk reduction, and enterprise-grade capabilities.

Financial Benefits

Predictable Costs / ROI Visibility: Capital expenditure model replaces unpredictable cloud fees. No data egress charges. No per-query fees. No surprise bills. Organizations deploying on-premises AI infrastructure achieve significant long-term savings compared to cloud AI—particularly as workloads scale beyond initial pilots.

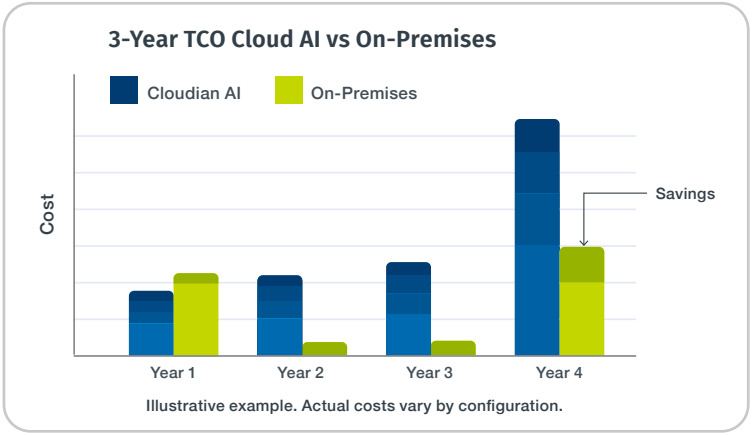
Reduced Implementation Risk: Pre-validated NVIDIA blueprints eliminate the integration complexity that causes 80%+ of custom AI projects to fail. Production-proven applications deliver value from day one, dramatically accelerating time-to-value compared to build-from-scratch approaches.

Accelerated Time-to-Value: Complete, integrated solutions deploy faster than custom-built alternatives. Begin realizing AI benefits immediately rather than waiting months or years for complex development cycles to complete.

Operational Efficiency: 24x7 HyperCare support with remote monitoring reduces burden on internal IT staff. Single-vendor support simplifies troubleshooting—no more finger-pointing between component vendors when issues arise.

Technical Benefits

Semantic Understanding: GPU-accelerated vector embeddings enable natural language queries across all content types—documents, charts, PDFs, videos—providing context-aware responses rather than simple file retrieval. Users interact with enterprise knowledge through conversational interfaces.



Defense-in-Depth Security: Comprehensive security including RBAC, multi-tenant isolation, AES-256 encryption, TLS 1.3, Object Lock for immutability, and SIEM integration. Air-gapped deployment capability addresses the most security-sensitive environments.

Disaggregated Scale-Out: Independent scaling of GPU processing and storage capacity from terabytes to exabytes with linear performance. Non-disruptive expansion without downtime or data migration enables infrastructure to grow with AI initiatives.

Open Standards: S3 API compatibility, standard GPU interfaces, and open-source AI frameworks prevent vendor lock-in. Your data and models remain portable, protecting your investment regardless of how the AI landscape evolves.

Predictable capital expenditure replaces unpredictable cloud fees—no egress charges, no per-query fees, no surprise bills.

9: Use Cases

Enterprise RAG and video analytics now. 30+ use case blueprints coming.

Enterprise Knowledge Base

Deploy complete RAG pipelines with multimodal document ingestion (PDF, DOCX, video transcripts), GPU-accelerated vector embeddings, and semantic search. Employees ask questions in natural language and receive instant, accurate answers grounded in your actual documentation—eliminating hours spent searching for and synthesizing information. Real-time knowledge base updates ensure AI responses reflect current policies and procedures. Full audit trails maintain regulatory compliance.

Customer Support Automation

Deploy AI-powered chatbots trained on product documentation, support articles, and call transcripts. Add new knowledge at any time as products and policies evolve. Answer customer questions in real time with accuracy backed by your authoritative documentation. All customer interactions and data remain on-premises—critical for industries with strict data handling requirements.

Computer Vision for Operations

Process video streams through GPU-accelerated computer vision models with natural language rule definition. Monitor manufacturing lines, warehouses, data centers, or security perimeters 24x7 with automated alerting. Define conditions conversationally—"Alert if shelving units exceed 30 items"—and receive sub-second detection. On-premises processing eliminates cloud bandwidth costs and addresses privacy concerns inherent in streaming video to external services.

Healthcare Diagnostic Support

Deploy HIPAA-compliant AI agents for diagnostic assistance, medical record analysis, and clinical decision support. RAG systems retrieve current treatment guidelines and relevant research at query time, ensuring recommendations reflect the latest medical evidence. Patient data never leaves your infrastructure. Full audit trails support regulatory compliance and litigation defense.

Financial Fraud Detection

Process real-time transaction streams through AI models to identify fraudulent patterns. Keep sensitive financial data within your perimeter while leveraging advanced AI analytics. AI-powered fraud detection can evaluate over 1,000 data points per transaction in real time, identifying anomalies that rule-based systems miss.

Beyond these, the possibilities are limitless as new blueprints are developed within the AIDP ecosystem. Other use cases include compliance & risk management, regulatory monitoring and analysis, legal document review, technical documentation Q&A, and personalized marketing content.

 |

10: Target Environments

HyperScale AIDP is designed for organizations where data control, regulatory compliance, and infrastructure ownership are essential requirements.

Regulated Industries

Financial services, healthcare, and government agencies where regulatory frameworks mandate on-premises data processing, complete audit trails, and data residency controls. Organizations operating under HIPAA, GDPR, FedRAMP, SOX, DORA, and similar frameworks benefit from infrastructure that keeps all data processing within their administrative domain.

Built for organizations where data control isn't optional—it's mandatory.

Manufacturing Operations

Organizations requiring real-time production line monitoring, quality control automation, safety compliance verification, and process optimization. On-premises AI eliminates network latency and bandwidth constraints while keeping operational data secure from external access.

Enterprise IT Organizations

Companies deploying internal knowledge management systems, employee support automation, and documentation search without exposing proprietary information to external services or incurring ongoing cloud API costs.

Professional Services

Law firms, consultancies, and advisory organizations that must maintain strict client confidentiality while leveraging AI for contract analysis, proposal generation, due diligence, and accelerated decision support.

Research and Development

Teams seeking to unlock insights from accumulated institutional knowledge, technical documentation, experimental data, and scientific repositories without exposing intellectual property to external services.

 |

11: Conclusion

The fundamental question facing enterprise IT isn't whether AI delivers value—it's whether you can deploy AI capabilities while maintaining complete control over your data, meeting regulatory requirements, and achieving predictable costs.

Cloud AI services offer convenience but introduce unacceptable risks for many enterprises: unpredictable costs that escalate rapidly, data residency concerns that conflict with compliance requirements, and vendor dependencies that limit future flexibility. Traditional on-premises approaches address these concerns but require scarce expertise and carry high implementation risk—as evidenced by 80%+ project failure rates.

Cloudian HyperScale AI Data Platform resolves this dilemma through sovereign AI architecture: production-validated applications that understand your data, deployed entirely within your infrastructure, with zero external dependencies.

Complete AI Infrastructure

GPU compute, enterprise storage, NVIDIA AI Enterprise software, and pre-validated blueprints in one integrated solution that delivers business value from day one.

Data Control Without Compromise

Your data never traverses external networks. Your AI is powered by your data alone. Your intellectual property remains under your complete administrative control.

Proven Technology Partnership

Built on two years of joint engineering between Cloudian and NVIDIA, integrating software validated through extensive real-world deployment. HyperScale AIDP is storage that understands your data, security that protects it, and AI infrastructure you control completely.

Zero external dependencies. Complete administrative control. Production-ready from day one.