Redbooks
ibm.com/redbooks

Enabling AI Inference at Scale: IBM Storage Scale ECE and IBM Fusion CAS with NVIDIA AI Data Platform

Ka Leung
Joseph Dain
Dr. Amrin Maria Khan Adawadkar
Saif Adil
Rodrigo Gutiérrez Vázquez
Uman Ahmed Mohammed

IBM

NVIDIA

Transforming Enterprise Data for AI at Scale

The rise of artificial intelligence agents promises to revolutionize enterprise operations by automating complex workflows and delivering unprecedented business insights. However, a critical bottleneck threatens this transformation: according to Gartner research, approximately 40% of AI prototypes fail to reach production, with data availability and quality cited as the primary barriers to adoption. The challenge lies not in the sophistication of AI models themselves, but in preparing the vast repositories of enterprise data—roughly 80% to 90% of which exists as unstructured documents, multimedia files, and other non-standardized formats—for consumption by AI systems. This gap between raw enterprise data and what the industry now calls "AI-ready data" represents one of the most significant obstacles facing organizations seeking to operationalize AI agents at scale. The NVIDIA AI Data Platform (AIDP) reference architecture addresses this challenge through an innovative approach that transforms traditional storage infrastructure from passive data containers into active AI preparation engines. By including GPU acceleration directly into the data path, NVIDIA AIDP performs the complex operations required to make unstructured data AI-ready (including semantic chunking, vectorization, and continuous synchronization) as background processes invisible to end users. This minimizes the GPU resource required for data processing on an AI Factory. NVIDIA AIDP also eliminates the data drift and security vulnerabilities associated with traditional extract-transform-load pipelines, while dramatically reducing the time data scientists spend on data preparation tasks. Built on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs and NVIDIA BlueField-3 DPUs, and supported by IBM, NVIDIA AIDP represents a fundamental reimagining of enterprise storage for the generative AI era. This publication introduces an IBM software-defined solution architecture featuring IBM Storage Scale® Erasure-Coding-Edition (ECE), IBM Fusion, and IBM Content-Aware Storage (CAS) with NVIDIA AIDP to enable enterprise-scale AI deployments. The solution provides these benefits:

- Seamless Connection to Existing Enterprise Data: Eliminates need for data replication and supports a range of multimodal data types including text, tables, audio, and video.
- Keeping data always up-to-date and ready for AI applications to perform retrieval augmented generation (RAG) without manual intervention.
- Optimized for Scaling: Leverages NVIDIA Blackwell GPU, NVIDIA Bluefield and NVIDIA Spectrum™-X technologies along with the IBM Storage Scale parallel file system and NVIDIA NeMo™ Retriever to enable complex unstructured data (video, charts, documents, logs) at the speed of business.

Authors and Contributors

Ka Leung is a Solution Product Manager for IBM Storage, focusing on AI solutions and leading the IBM Fusion ISV Alliance Program.

Joseph Dain is an AI storage solutions technical leader within IBM's Infrastructure organization, designing next-generation data platforms that power AI training, inferencing, and agentic systems. His work drives innovation in IBM Content Aware Storage and the NVIDIA AI Data Platform, kv-cache reuse, and high-performance architectures integrating NVIDIA NIM™ and NVIDIA AI Enterprise. With more than 125 patents across AI, systems, and storage, he bridges advanced AI models with the scalable storage technologies that enable enterprise intelligence.

Dr. Amrin Maria Khan Adawadkar is a Software Developer at IBM with expertise in cybersecurity, Identity and Access Management (IAM), AI applications, and IBM Maximo Application Suite (MAS), IBM's enterprise asset management solution. Since joining IBM in 2015, she has contributed to IBM Security through technical enablement, including authoring learning content and videos. Amrin actively engages with the broader technology community through speaking, mentoring, and judging at national and international hackathons. She holds a PhD in Identity and Access Management and Reinforcement Learning, with research interests in reinforcement learning, secure AI systems, advanced encryption techniques, and AI-driven storage solutions.

Saif Adil is a Technical Product Manager at IBM focused on Storage Solutions Engineering, where he connects customer outcomes with solutions strategy and ecosystem innovation. He specializes in bridging the gap between technical infrastructure and business value across IBM's storage and AI platform portfolio.

Rodrigo Gutiérrez Vázquez is an AI/Cloud Software Engineer for IBM Fusion Content-Aware Storage product. He has over 8 years of industry experience, with an extensive track record leading backend software development teams in the storage sector at the enterprise level, collaborating on design, planning, and implementation phases. He has also contributed to different teams across the IBM Fusion organization, as well as public research and open-source projects. His areas of expertise include distributed application development on hybrid-cloud platforms, AI-based application development, high-performance computing, and parallel programming.

Uman Ahmed Mohammed is an Enterprise Infrastructure Architect with over 20 years of experience designing and managing secure, scalable platforms across private, hybrid, and multi-cloud environments. He is an IBM Consulting Expert, IBM Open Innovation Community (OIC) Rockstar Contributor, IBM Co-Inventor, IBM Gold Champion Learner, IEEE Senior Member, and IEEE 2025 Regional Outstanding Professional Awardee. A graduate of BITS Pilani, he combines deep technical expertise with recognized thought leadership, publishing research through IEEE conferences and journals, IBM Redbooks, and the IBM Think Blog. His current focus areas include AI-ready platforms, AI factories, digital twins, high-performance computing (HPC), and sovereign cloud solutions for next-generation enterprises.

Publication Guide and Next Steps

This publication provides comprehensive guidance for deploying NVIDIA AIDP using the integrated solution architecture of IBM and NVIDIA. The publication targets AI architects, solution architects, and technical decision-makers responsible for enterprise AI strategy and implementation. Business stakeholders find value in the architectural guidance and business case development, while technical teams can leverage the deployment architecture and reference implementations to accelerate their AI initiatives.

Introducing the NVIDIA AI Data Platform

The NVIDIA AI Data Platform (AIDP) is a customizable reference architecture for next-generation AI infrastructure. It delivers real-time inference and enhanced business intelligence by integrating enterprise-grade data storage with NVIDIA accelerated computing. It enables AI agents to access near real-time operational insights. The platform transforms unstructured enterprise data into AI-ready intelligence by embedding GPU acceleration in the data path, enabling in place processing that eliminates unnecessary copies and related security risks. By integrating GPU-accelerated compute with data storage infrastructure, the platform continuously ingests data to perform indexing, vectorization, and semantic embedding generation as background operations, making data immediately consumable by AI training, finetuning, and retrieval augmented generation (RAG) pipelines without additional preparation. This capability improves AI agents' operational efficiency by providing immediate access to contextual information and strengthening complex reasoning and multistep problem solving. The platform indexes stored data and exposes it to knowledge retrieval systems in near real time. This approach ensures AI applications can access the organization's knowledge through a search interface or an API call. LLM inference applications deliver more accurate, timely answers. This capability lets enterprises unlock value from unstructured data while keeping data secure and enforcing enterprise-grade access control across the entire AI pipeline.

Core architectural components of the NVIDIA AIDP reference design

1. GPU-accelerated compute layer: This layer delivers the computational foundation for extracting semantic meaning from data to make it AI-ready. It includes compute servers equipped with NVIDIA GPUs and NVIDIA BlueField DPUs to accelerate AI model training and inference. It also supports NVIDIA NIM, which simplify deployment of AI services like RAG and agentic workflows.
2. Storage layer: This layer optimizes high-throughput, low-latency access to large datasets. It uses NVIDIA ConnectX SuperNICs to enable fast data movement between storage and compute. This layer ensures:
   - low latency data retrieval for training and inference.
   - scalable throughput to support parallel workloads.
   - tight integration with the compute layer through high-speed networking interfaces.
3. Networking layer: NVIDIA Spectrum-X Ethernet This layer provides the high-bandwidth, low-latency backbone that connects compute and storage, ensuring consistent performance across distributed AI workloads. This layer ensures:
   - predictable performance across distributed AI clusters
   - efficient data movement to minimize bottlenecks
   - support for Ethernet based topologies
4. AI software layer: NVIDIA AI Enterprise and NVIDIA AI Blueprints This layer includes NVIDIA NIM microservices and libraries from NVIDIA AI Enterprise that simplify running and scaling NVIDIA AIDP in production. It uses the NeMo Retriever NIM to provide high accuracy when retrieving content across modalities and high throughput for extracting, embedding, indexing, and retrieving documents at scale.
5. NVIDIA AI-Q Blueprint: a key design reference for NVIDIA AI. This blueprint provides AI-ready infrastructure that uses local enterprise data and web search to create deep research assistants. Built on NVIDIA NIM microservices, NeMo Retriever, and NVIDIA Nemotron reasoning models, the AI-Q Blueprint serves as a framework for constructing AI agents that can access diverse data sources, use various tools, and perform complex reasoning at enterprise scale.

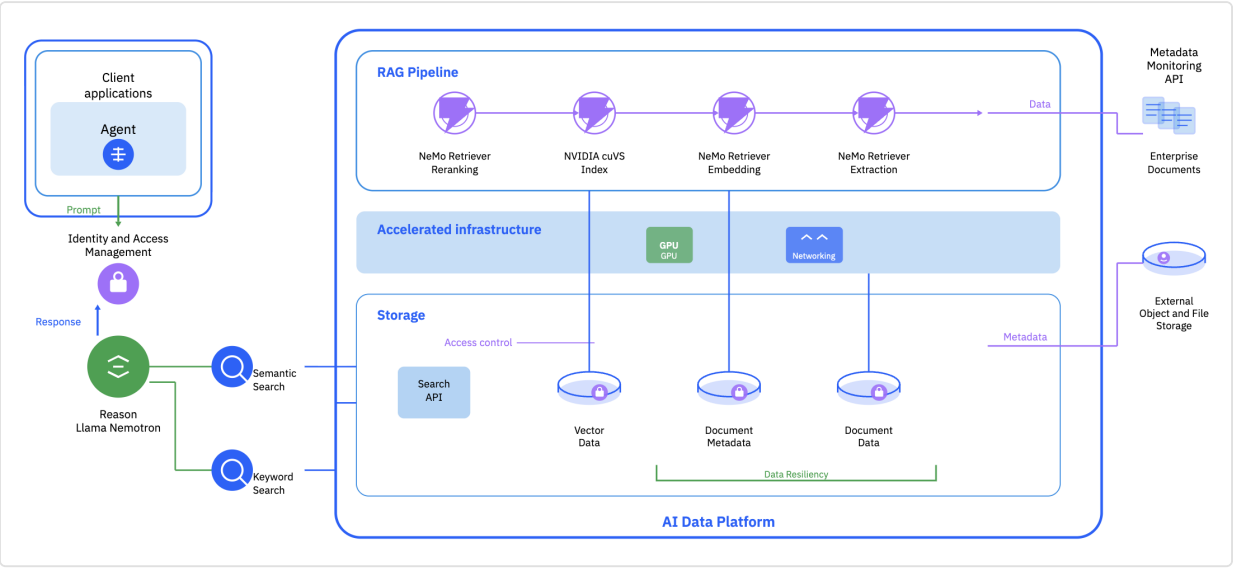

Figure 1. End-to-end RAG on an AI data platform

Optimizing NVIDIA AIDP on IBM

IBM is a key design partner for NVIDIA AIDP, focusing on optimizing performance and fidelity for agentic AI workflows. The core components of this IBM reference design include:

1. x86 compute servers with NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPUs, NVIDIA BlueField-3 DPUs (data processing units), and NVIDIA ConnectX®-7 or ConnectX®-8 SuperNICs
2. IBM Fusion software: provides the OpenShift-based Kubernetes foundation for building the NVIDIA AIDP
3. IBM Storage Scale ECE system: enables concurrent multi-GPU data access, eliminating I/O bottlenecks for real-time NVIDIA AIDP updates and retrieval
4. IBM Fusion Content Aware Storage (CAS): provides real-time data updates to NVIDIA AIDP, eliminates silos for faster GPU inference, and enforces existing data source access controls in the RAG pipeline
5. NVIDIA Spectrum™-X: provides high-bandwidth, low-latency connectivity for rapid data transfer between IBM Storage Scale and IBM Fusion systems

IBM on NVIDIA AI Data Deployment Topology

The storage layer is the critical layer in the NVIDIA AIDP. It supports high-throughput, low-latency data access for real-time AI workloads. The IBM components of this solution use the software-defined version of IBM Storage Scale and IBM Fusion CAS, running on x86 CPU- and GPU-based servers. This disaggregated solution lets you scale capacity and performance independently.

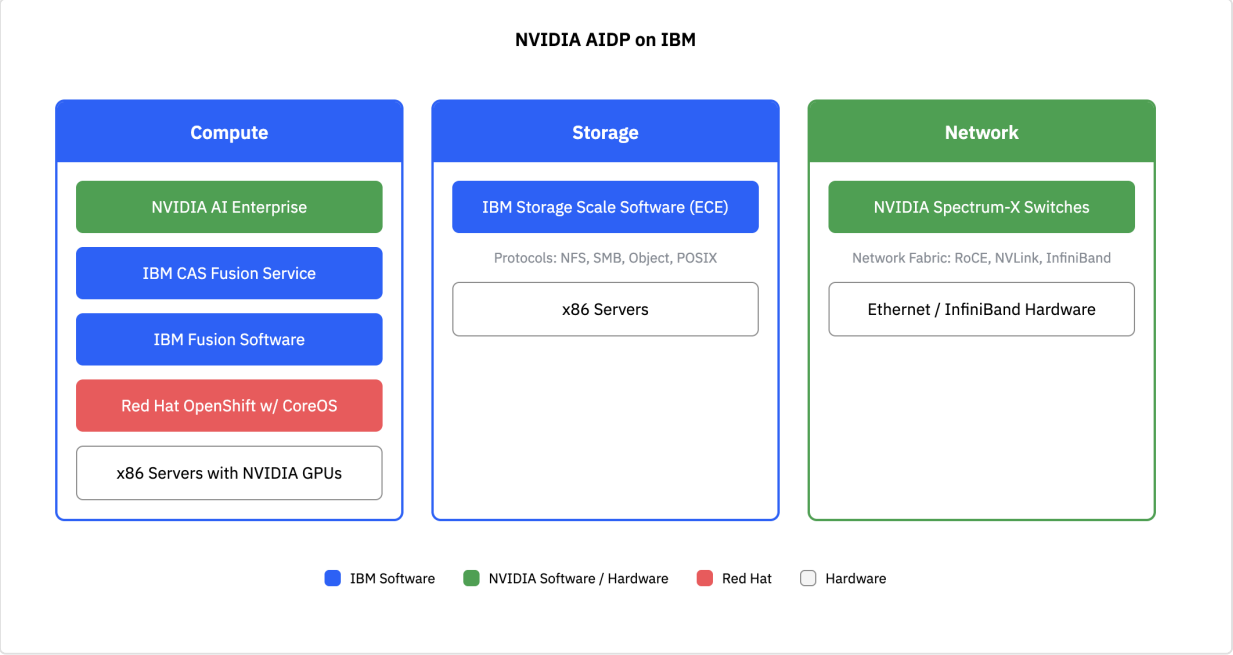

Figure 2. NVIDIA AIDP on IBM—reference stack

Enabling Real-Time Intelligence on NVIDIA AIDP with IBM Storage Scale ECE and CAS

This chapter explains the role and benefits of IBM Storage Scale ECE and Content-Aware Storage (CAS) in the NVIDIA AIDP for AI and agentic AI deployments.

IBM Storage Scale ECE Overview - Optimized for NVIDIA AIDP

IBM Storage Scale ECE is software-defined storage featuring a high-performance parallel file system designed to eliminate data I/O bottlenecks. It addresses a core challenge in enterprise AI: keeping GPUs continuously fed with data for real-time inference, RAG workflows, and agentic AI applications.

Flexible Deployment for NVIDIA AIDP Architectures

IBM Storage Scale offers deployment flexibility aligned with NVIDIA AIDP reference designs through a software-defined storage (SDS) implementation on x86 servers with RHEL. The SDS approach lets IBM Storage Scale use the latest NVIDIA GPU and networking technologies in the underlying platform. It also enables cost-effective NVIDIA AIDP builds while maintaining the parallel I/O performance required for multi GPU environments.

Unified Data Platform for AI Agents

NVIDIA AIDP deployments face a challenge: AI agents must quickly access and process enterprise data at scale, but the data is siloed across cloud object storage, on-premises file systems, databases, and legacy repositories. Storage Scale's Global Data Platform architecture virtualizes dispersed data sources into a single namespace for NVIDIA GPUs, whether data resides natively in Storage Scale or is accessed from external storage through Active File Management (AFM).

Performance Architecture for GPU-Accelerated Workloads

Storage Scale's parallel file system delivers concurrent, high-bandwidth data access for NVIDIA AIDP workloads. Multiple GPUs simultaneously retrieve vector embeddings, document chunks, and knowledge base content without I/O contention, enabling RAG pipelines to scale linearly as compute resources grow.

Data Access Services

With a rich set of data protocols, IBM Storage Scale Data Access Services provide unified, shared file and object access to unstructured data across the organization. These services are “multilingual”: some applications create and access data with one protocol, while others access the same data with a different protocol concurrently.

Data Management Services

IBM Storage Scale provides comprehensive lifecycle management services, including a flexible policy engine that lets customers to define rules to optimize storage of unstructured data. These services transparently move data to the appropriate storage tier, optimizing cost and performance based on an organization's retention, archiving, and data governance policies.

Data Resiliency Services

Data Resiliency Services provide comprehensive tools to identify threats, protect an organization's data, and support response and recovery when security breaches occur. These services align with the NIST security framework, from practicing cyber hygiene before an event through detection, response, and recovery.

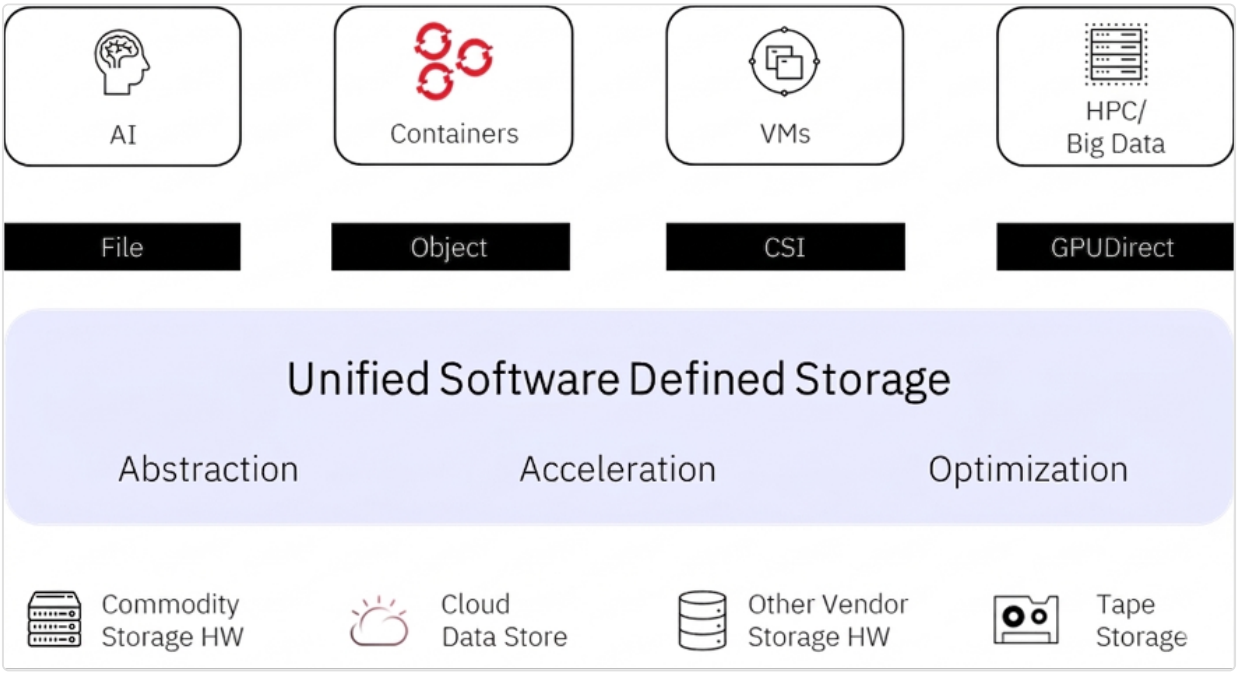

Figure 3. Unified software-defined storage for mixed AI and enterprise workloads

Comprehensive Multi-Protocol Support

IBM Storage Scale ECE provides robust multiprotocol data access for diverse workloads and environments. It supports a wide range of protocols, including:

- POSIX for traditional file system access
- S3 for object storage compatibility
- NFS and SMB for network file sharing
- CSI (Container Storage Interface) for Kubernetes-based container environments

This multiprotocol capability lets organizations consolidate data access across platforms and applications, enhancing flexibility and simplifying data management. IBM Storage Scale ECE creates a single global namespace, eliminating the need to juggle multiple storage systems. This unified view lets you access and manage data regardless of where it resides—whether in the cloud, on-premises NAS or object stores, or on traditional tape systems. Storage Scale brings it together under one logical structure, streamlining workflows and reducing complexity so you can scale, govern, and optimize your storage environment.

Accelerating AI Data Access with Storage Scale AFM

IBM Storage Scale Active File Management (AFM) is a distributed caching feature that establishes a global, unified data platform and namespace across geographically dispersed IBM Storage Scale clusters and third-party NFS and cloud object storage sources. By allowing a local cluster (the "cache") to fetch and store copies of remote data (the "home"), AFM masks network latency and delivers high-speed, local access to data for remote users and AI applications. Key features include partial file caching (fetching only the needed blocks), parallel data transfers, and asynchronous replication for data mobility and remote collaboration. AFM addresses data gravity in AI workloads by providing local, high-performance access to large datasets needed by distributed compute resources. It accelerates iterative data ingestion for model training, fine-tuning, and inference, reducing time-to-insight. You can use AFM as a high-speed cache in front of slower object storage to keep GPUs busy and achieve fast LLM training checkpointing and load times. This optimization is critical for maximizing GPU utilization and ensuring data consistency for collaborative, multi-site AI development.

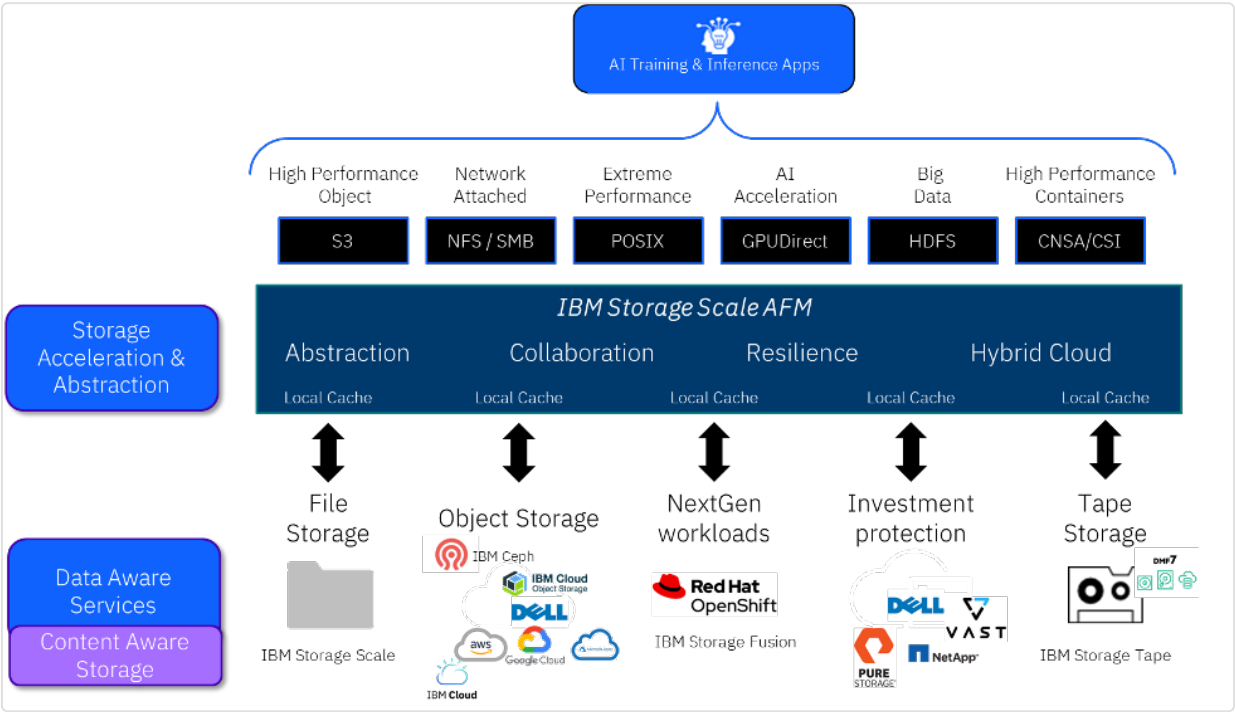

Figure 4. IBM Storage Scale AFM provides a local-cache, data-virtualization layer that accelerates and abstracts access for AI training and inference across S3, NFS/SMB, POSIX, GPUDirect, HDFS, and CNSA/CSI.

Accelerating Data Transfers with NVIDIA GDS on IBM Storage Scale ECE

IBM Storage Scale ECE supports NVIDIA GPUDirect Storage (GDS), which reduces I/O bottlenecks that can cause GPUs to idle while waiting for data. GDS creates a direct memory access (DMA) path between the Storage Scale file system and GPU memory, bypassing CPU system memory and the kernel I/O stack. This direct path provides three benefits:

- Maximizes Bandwidth: Eliminates intermediate data copies to achieve near-native wire speeds for GPU data ingestion
- Minimizes Latency: Reduces hop count to prevent GPU I/O starvation
- Reduces CPU Overhead: Frees CPU cycles for application logic by offloading dataset transfer management

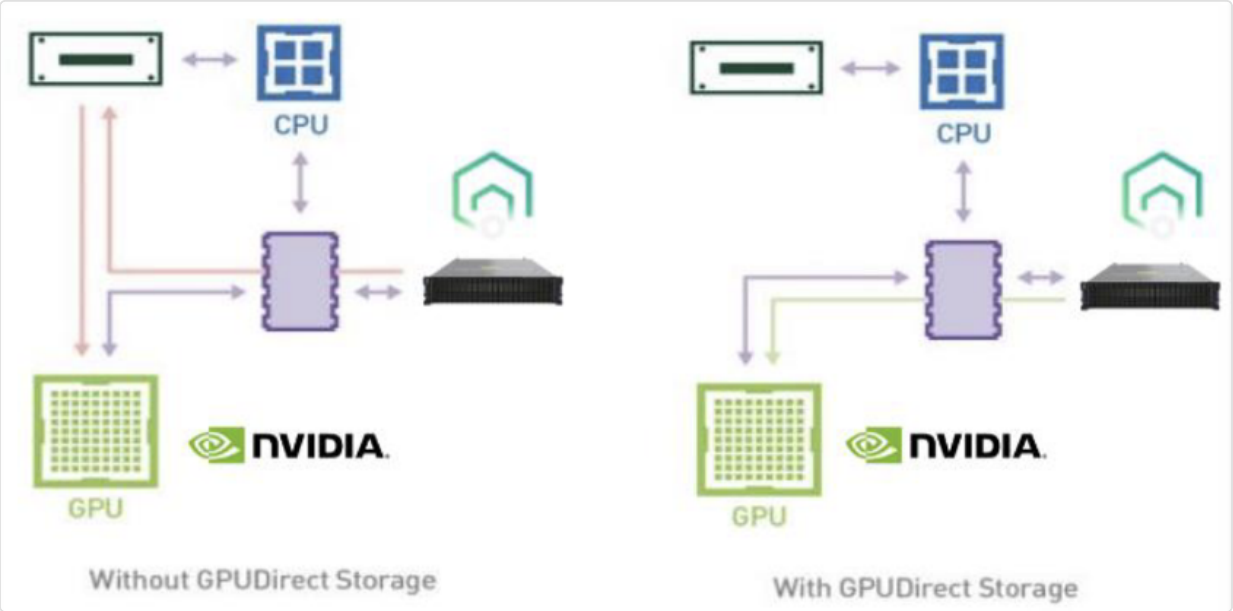

Figure 5. With NVIDIA GPUDirect Storage, data moves directly between storage and GPU memory, bypassing the CPU and extra host-memory copies. The streamlined path reduces hops and CPU involvement, improving throughput and latency for AI and data-intensive workloads.

Optimizing RAG Data retrieval in NVIDIA AIDP with IBM Fusion CAS

IBM Content-Aware Storage (CAS) solves data gravity in an NVIDIA AIDP design by embedding AI processing in the storage layer and eliminating the costly, slow practice of moving massive enterprise data to compute. IBM Fusion CAS enables zero-copy ingestion and incremental vector-database updates from unstructured sources, reducing data movement, latency, and operational complexity. Enterprises struggle with stale data in RAG pipelines because traditional methods require frequent full vector-database rebuilds—a process IBM Fusion CAS replaces with real-time, event-driven incremental updates. IBM Fusion CAS improves RAG workflows in NVIDIA AIDP by helping ensure that responses from AI assistants and agents use the most recent, relevant context. It also addresses security and compliance risks by ensuring that vectors inherit and enforce file-level access controls (ACLs) from source documents. By integrating vectorization and compute close to data sources, IBM Fusion CAS can improve RAG pipeline performance and reduce the operational burden of managing separate data and vector-database systems. In an NVIDIA AIDP implementation, IBM Fusion CAS rearchitects the retrieval-augmented generation (RAG) pipeline by executing critical data-preparation steps in the storage layer, which minimizes AI inference latency and resource costs. The IBM Fusion CAS RAG workflow in NVIDIA AIDP operates in two phases: the Data Ingestion/Preparation Flow and the Retrieval/Query Flow.

Making data AI-ready with data ingestion and preparation flow

IBM Fusion CAS enhances the NVIDIA AIDP, by providing a continuous and automated process to index and vectorize data as it arrives:

- Data ingestion and change detection: Ingests text and multimodal unstructured data sources from IBM Storage or third-party storage providers. When source data changes, IBM Fusion CAS, using Active File Management (AFM), detects the change, retrieves the data, and incrementally updates the RAG vector database.
- Incremental processing: Instead of reprocessing the entire dataset, IBM Fusion CAS triggers processing only for new or modified files. This approach reduces time and cost.

- Semantic extraction: The system passes the file to an embedded, accelerated data-processing pipeline that uses the NVIDIA NeMo Retriever microservice to perform deep content extraction. It goes beyond text to extract semantic meaning from tables, charts, images, and infographics in unstructured documents.
- Vectorization and storage: Extracted content is chunked and converted into high-dimensional vectors (embeddings) using an NVIDIA embedding NIM. The system stores these vectors in an IBM Fusion CAS-managed vector database on IBM Storage Scale ECE.
- Security inheritance: Vectors inherit role-based access control (RBAC) policies and access control lists (ACLs) from their source files, maintaining governance and security at the vector level.

Retrieval and Query Flow (Serving the AI Context)

When a user or AI agent asks a question, IBM Fusion CAS accelerates and secures context retrieval:

- Query vectorization: The system converts the user's natural-language query into a vector with the same embedding model used for the data.
- Vector search at storage: The query vector is sent to the IBM Fusion CAS Search API, which executes a high-speed search (supporting semantic, keyword (BM25), and hybrid methods) against the integrated vector database.
- Access control enforcement: Before returning results, IBM Fusion CAS enforces inherited file ACLs, ensuring returned chunks are authorized for the requesting user. Generation: The augmented prompt is sent to the LLM for final response generation, producing more accurate, grounded answers based on real-time, proprietary enterprise data.

IBM Fusion CAS provides the following advantages for agentic AI:

- Metadata intelligence: CAS indexes data at the file and object level, capturing semantic information for fast, accurate search and retrieval—essential for real-time access.
- Real-time data access and classification: CAS allows dynamic classification and retrieval based on content and context, supporting constant ingestion and analysis. Smooth interaction between LLMs and data: CAS enhances insight and recommendation capabilities that facilitate interaction between LLMs and large volumes of structured, semi-structured, and unstructured data.

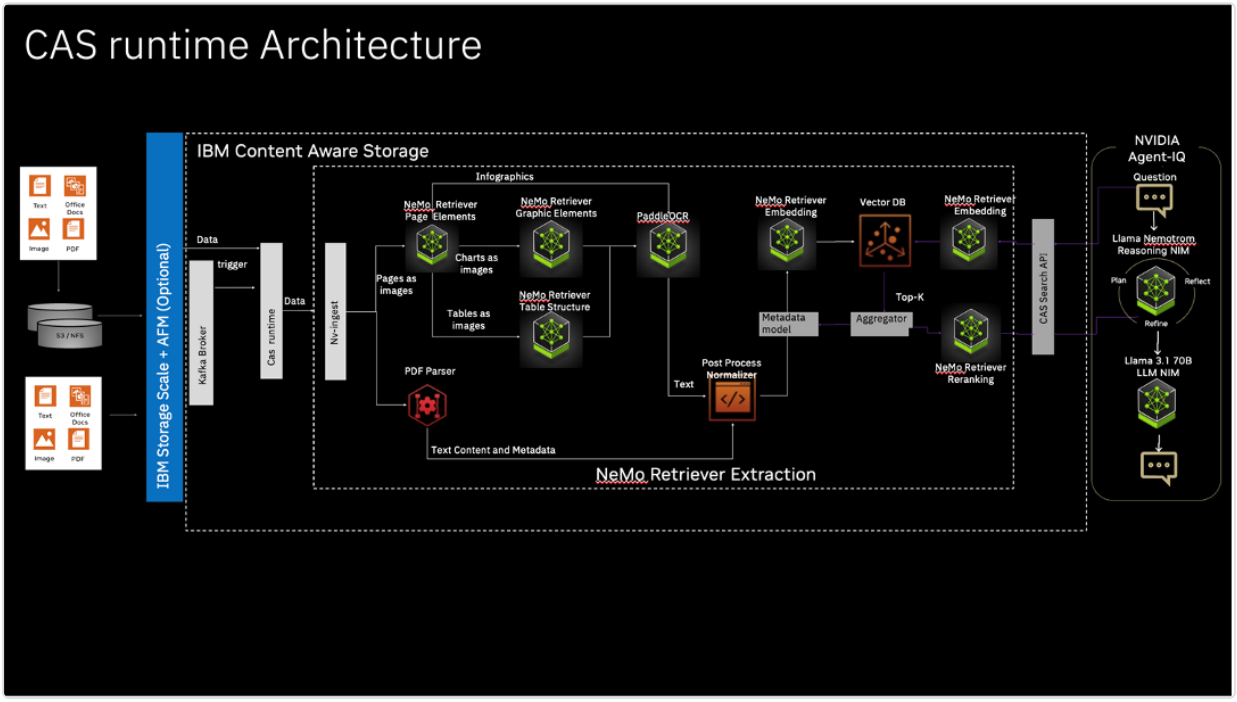

Figure 6. CAS runtime: IBM Content Aware Storage ingests heterogeneous files, parses and extracts page/graphic/table elements, generates NeMo Retriever embeddings, and stores them in a vector database.

Why IBM Fusion CAS Over Traditional Pipelines?

IBM Fusion Content-Aware Storage (CAS) provides high-performance advantages for agentic AI compared to other RAG pipelines. Its integrated architecture embeds compute and vector processing in the storage layer, yielding improvements in key areas. Specifically, IBM Fusion CAS hardware-accelerated metadata processing and real-time incremental updates translate to lower latency for AI inference. This design scales to handle massive enterprise data volumes and improves accuracy by grounding agents in the freshest, most contextually rich data available.

IBM Fusion CAS: Advancing RAG Accuracy with Smarter Semantic Search

IBM Fusion Content-Aware Storage (CAS) improves the accuracy of Retrieval-Augmented Generation (RAG) systems by optimizing how relevant information is retrieved and ranked before being provided to large language models. Rather than focusing solely on vector-search performance, CAS emphasizes end-to-end question-answering accuracy, using evaluation metrics such as Recall@K and nDCG@K across BEIR benchmark datasets. The platform integrates high-dimensional embedding models, efficient DiskANN vector indexing, and hybrid retrieval techniques that combine semantic vector search with traditional lexical scoring. These capabilities can be enhanced with cross-encoder re-ranking to improve the ordering of retrieved results. Together, these optimizations enable IBM Fusion CAS to significantly improve retrieval precision and relevance, surpassing results from traditional RAG designs based on vector-database searches alone.

IBM Fusion CAS Value in NVIDIA AIDP Data Pipeline Processing

Customer Value	Processing RAG Pipeline Closer to Storage Enables
Faster Time to Insights	• Fine-grained knowledge of data changes allows for rapid incremental updates• Avoids the need for complete rebuilding of Vector Database with incremental data updates
Reduced Cost	• Reduces the number of data replicas• Optimizations that leverage the semantic understanding of the data (such as deduplication and decryption avoidance) are implemented
Improved Performance & Accuracy	• Uses NVIDIA GDS protocol to improve data loading bandwidth (GB/s) by up to 8.4× in data loading bandwidth (GB/s)• Improve RAG accuracy through advanced semantic retrieval, hybrid search, and intelligent re-ranking• Storage tiering to preserve intermediate outputs cuts the cost of upgrading the embedding model by up to 90%
Better Security	• Reduces the threat surface area's security exposure• Maps vector queries directly to RBAC (Role-Based Access Control) policies enforced by the file system, eliminating the need to synchronize ACLs (Access Control Lists)
Simplified Operations	• Simplifies overall RAG architecture for Agentic AI• Reduces both skill requirements and Day-0 through Day-2 operational burden
Unleashing Legacy Data	• Connects seamlessly to existing data stored on other IBM and broad range of third-party storage• No data duplication or movements is required

Solution Architecture and Deployment Configurations

IBM Fusion for NVIDIA AIDP Architecture

IBM implements the NVIDIA AIDP reference architecture using its IBM Storage Scale ECE and IBM Fusion CAS. The following figure shows the details of a starter NVIDIA AIDP configuration.

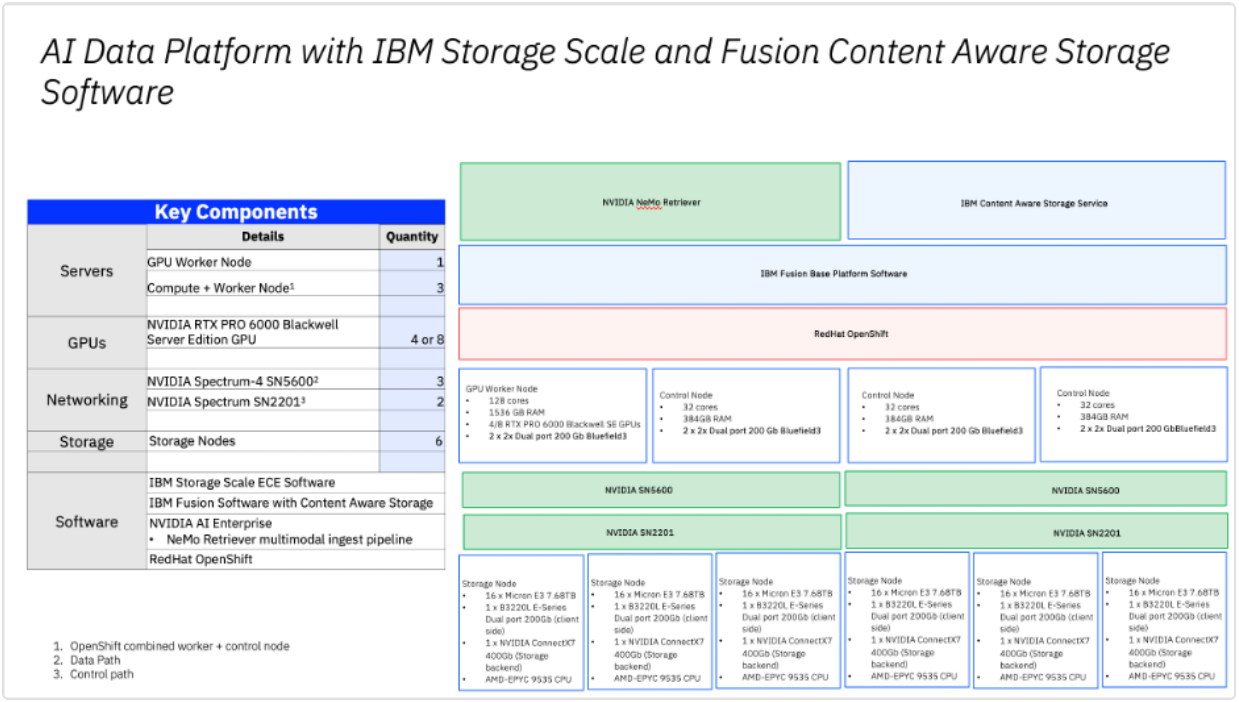

Figure 7. Example AI data platform: IBM Storage Scale ECE storage nodes provide the high-throughput data layer, while IBM Fusion Content Aware Storage and NVIDIA NeMo Retriever services run on Red Hat OpenShift.

IBM Fusion for NVIDIA AIDP Starter Rack with Optional Components

You can add the following optional components to the IBM Fusion for NVIDIA AIDP starter configuration:

- Storage Scale protocol node(s): provides S3, NFS, SMB protocol access services to data residing on the Storage Scale file system
- Storage Scale AFM node(s): transparently caches data from external data sources The following figure provides a rack level view of the starter configuration and optional components

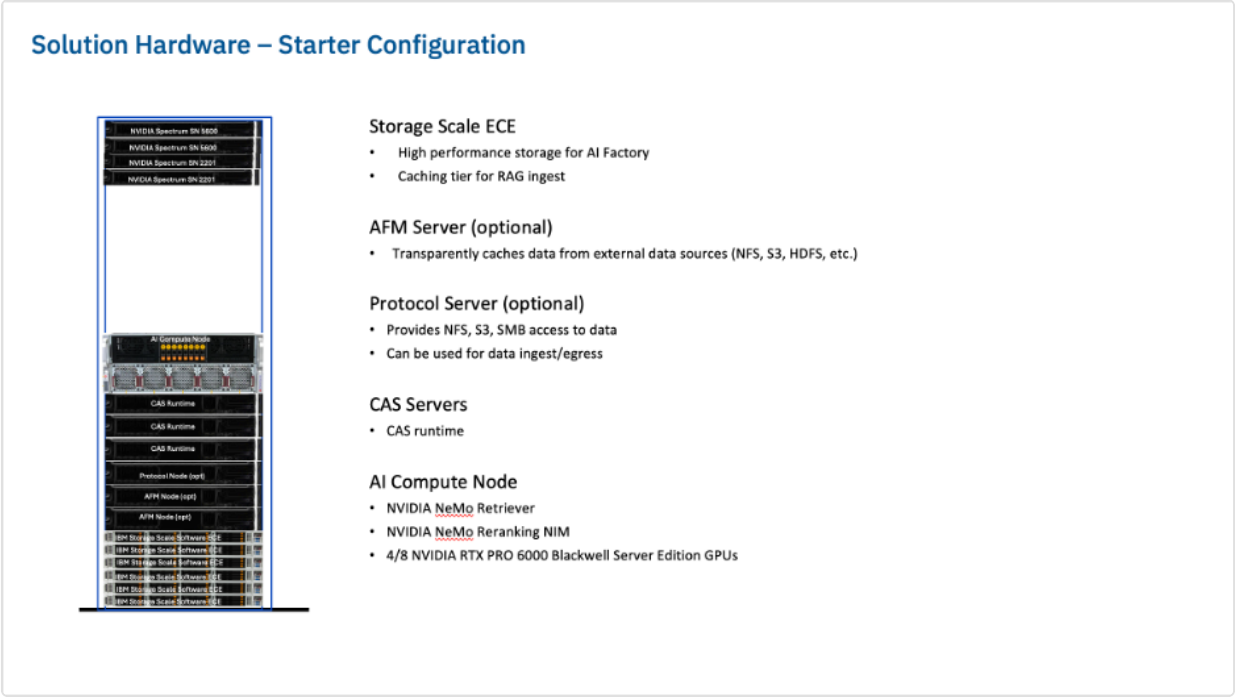

Figure 8. Starter rack configuration for an AI data platform: IBM Storage Scale ECE supplies high-throughput storage and a RAG caching tier; optional AFM and Protocol servers add transparent external-source caching and NFS/S3/SMB access.

IBM Fusion for NVIDIA AIDP Scaling

IBM Fusion CAS instances using NVIDIA NeMo Retriever can scale to meet RAG ingest and query SLAs. IBM Storage Scale file systems can scale independently by adding storage-rich servers running IBM Storage Scale. Each IBM Fusion CAS GPU server supports up to 8 NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. The reference architecture supports up to 8 GPU servers per rack—each with 8 NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs—for a total of 64 GPUs per expansion rack. The following figure shows the base rack with the expansion rack.

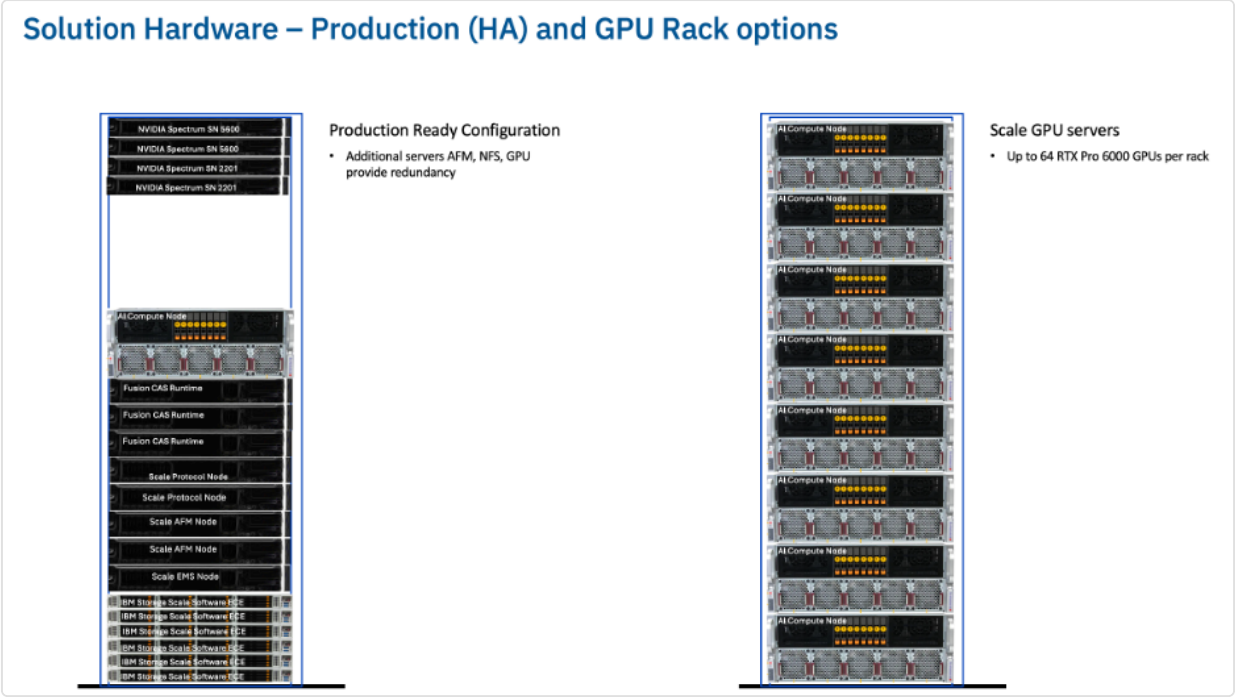

Figure 9. Starter rack for an AI data platform: IBM Storage Scale ECE delivers high-throughput storage and a RAG caching tier; optional AFM and Protocol servers add external-source caching and NFS/S3/SMB access.

IBM Fusion for NVIDIA AIDP Performance Scaling

Ingest and query performance of the IBM Fusion CAS instance that uses NVIDIA NeMo Retriever increases as more NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs are added. The following figure provides performance scaling estimates for the IBM Fusion NVIDIA AI Data Platform.

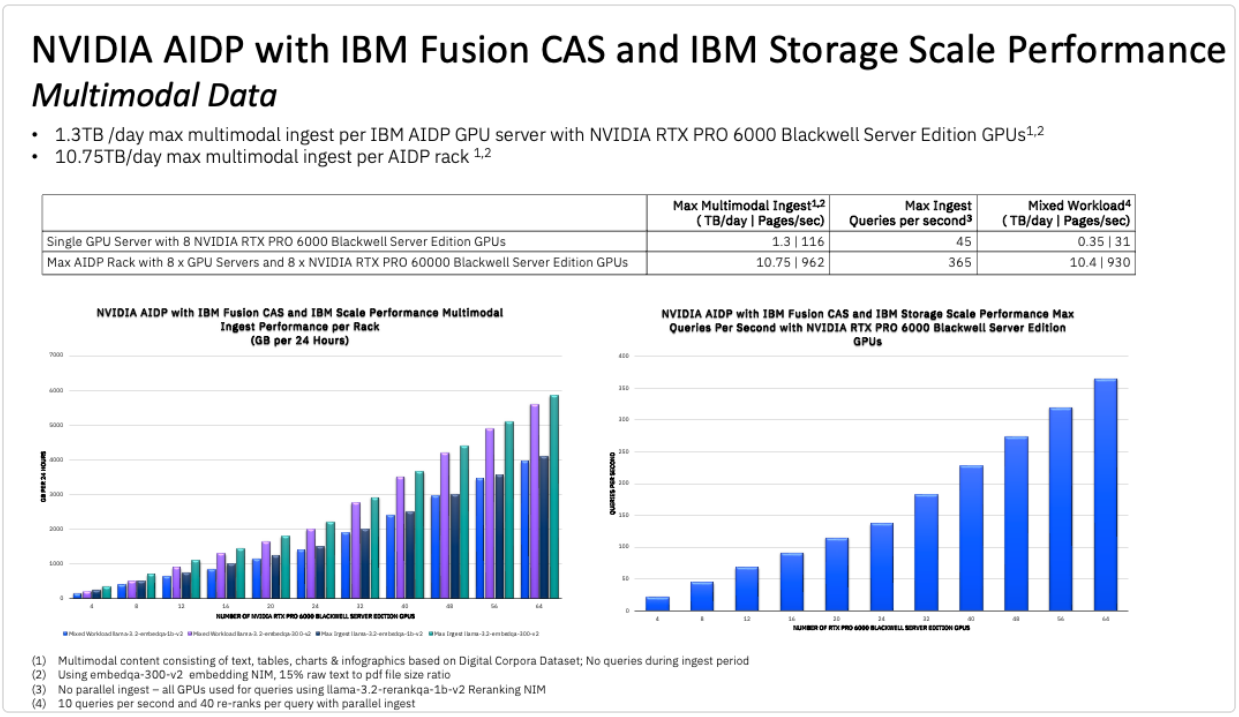

Figure 10. NVIDIA AIDP with Content Aware Storage (CAS) and IBM Storage Scale sustains high throughput for multimodal RAG ingest and serving.

Summary

Powering Enterprise AI with NVIDIA AIDP on IBM

The escalating demand for AI inference, particularly for large language models, necessitates a fundamental shift in enterprise data storage. NVIDIA and IBM are partnering to deliver AI-ready solutions that transform traditional storage into dynamic knowledge retrieval systems. These IBM solutions, built on NVIDIA's AI Data Platform enable businesses to unlock critical insights from their vast datasets, making them instantly accessible for high-performance AI applications. The IBM implementation of NVIDIA’s AI Data Platform ensures enterprise content is indexed and accessible in near real-time, dramatically improving the accuracy and timeliness of LLM inference and agentic AI applications. It allows enterprises to securely leverage all their data, maintaining robust access control and security throughout the AI pipeline. With IBM Storage Scale ECE providing efficient real-time access to diverse data and IBM Fusion Content-Aware Storage (CAS) ensuring real-time data updates and retrieval accuracy, businesses can rapidly develop and deploy high-performance agentic AI and RAG workloads in production.

Related Publications and Resources

NVIDIA AIDP

- NVIDIA AIDP
- NVIDIA AI-Q Research Assistant Blueprint
- NVIDIA Unveils AI-Q Blueprint to Connect AI Agents for the Future of Work
- NVIDIA Nemo Agent Toolkit
- Accelerate research using NVIDIA AI-Q and NVIDIA RAG Blueprint on IBM Fusion HCI — Technical guide for deploying NVIDIA RAG blueprint and AI-Q Research Agent Blueprint on IBM Fusion HCI platform

IBM Storage Scale

- Storage Scale System resources — IBM (Official resource page with technical documentation)
- [PDF] Introduction to Storage Scale (Architecture, deployment models, performance benchmarks)
- IBM Storage Community (Latest developer articles and resources)
- IBM Storage Scale delivers real-world performance: MLPerf benchmark results (Recent performance validation)
- IBM Redbooks (Published Redbooks for storage technologies, including Spectrum Scale)

IBM Content-Aware Storage

- [PDF] IBM Content-Aware Storage (Technical slide deck, usage for AI workflows and RAG integration)
- Content aware storage for the generative AI era - IBM Research (Blog detailing features and generative AI use cases)
- IBM boosts Storage Scale with content-aware AI integration (News coverage of AI features and enhancements)
- Next-Generation Search: IBM Fusion CAS Outshines State-of-the-Art RAG Benchmark Suite (IBM Blog)
- Enhancing AI Results with Content-Aware Storage Client L1 (Technical deep dive, architecture, integration with NVIDIA NIM and vector DB)
- IBM's Content-Aware – Storage Scale for AI Workloads - Futurum (Industry analyst take on market impact)
- AI Data Assistance with IBM Content Aware Storage (CAS) (Support and orchestration documentation)
- Content-Aware Storage (CAS) - IBM (Official IBM Fusion service documentation)
- Research Note: IBM Content-Aware Storage for RAG AI Workflows (Research note, NLP techniques in storage)
- [PDF] WHITE PAPER Infrastructure Simplification for Tiered Storage - IBM (White paper: tiered, content-aware storage)

Glossary of Acronyms

AI & Machine Learning

AI - Artificial Intelligence : Computer systems designed to perform tasks that typically require human intelligence, including learning, reasoning, and problem-solving. IBM Storage Scale optimizes data access for AI workloads. : *Context: Core technology driving the NVIDIA AIDP* : *Related terms: machine learning, deep learning, LLM*

AIDP - AI Data Platform : NVIDIA's customizable reference architecture for next-generation AI infrastructure that delivers real-time inference and enhanced business intelligence by integrating enterprise-grade data storage with NVIDIA-accelerated computing. : *Context: Primary platform architecture discussed in this publication* : *Related terms: NVIDIA, GPU acceleration, RAG*

LLM - Large Language Model : Advanced AI models trained on vast amounts of text data to understand and generate human-like text. Used for inference, question-answering, and agentic AI applications. : *Context: AI models that consume data prepared by NVIDIA AIDP* : *Related terms: AI, inference, RAG*

NIM - NVIDIA Inference Microservices : Containerized AI services from NVIDIA AI Enterprise that simplify deployment and scaling of AI models in production environments. Includes NeMo Retriever for RAG workflows. : *Context: Software layer in NVIDIA AIDP architecture* : *Related terms: microservices, NeMo Retriever, RAG*

RAG - Retrieval-Augmented Generation : AI technique that enhances LLM responses by retrieving relevant information from external knowledge bases before generating answers, improving accuracy and grounding responses in enterprise data. : *Context: Primary AI workflow enabled by IBM Fusion CAS* : *Related terms: LLM, vector database, semantic search*

Storage Technologies

AFM - Active File Management : IBM Storage Scale feature that enables automated data movement and caching between file systems, supporting disaster recovery and multi-site data management. : *Context: Used for transparent caching of external data sources* : *Related terms: data management, replication, caching*

CAS - Content-Aware Storage : IBM Fusion feature that embeds AI processing in the storage layer, enabling zero-copy ingestion, incremental vector-database updates, and real-time data preparation for RAG pipelines. : *Context: Key IBM component for NVIDIA AIDP implementation* : *Related terms: IBM Fusion, vectorization, RAG*

CSI - Container Storage Interface : Standard interface for exposing storage systems to containerized workloads on Kubernetes and other container orchestration platforms. : *Context: Container storage integration for IBM Storage Scale* : *Related terms: Kubernetes, persistent volumes, containers*

ECE - Erasure-Coding-Edition : IBM Storage Scale deployment model featuring software-defined storage with erasure coding for data protection, optimized for high-performance parallel file system access. : *Context: IBM Storage Scale variant used in NVIDIA AIDP architecture* : *Related terms: IBM Storage Scale, software-defined storage, data protection*

GDS - GPUDirect Storage : NVIDIA technology that creates a direct memory access path between storage and GPU memory, bypassing CPU and system memory to maximize bandwidth and minimize latency. : *Context: Performance optimization for GPU data access* : *Related terms: GPU, DMA, NVIDIA*

NFS - Network File System : Distributed file system protocol that allows remote file access over a network. IBM Storage Scale supports NFS protocol for client access. : *Context: Protocol support in IBM Storage Scale* : *Related terms: file sharing, protocol, SMB*

POSIX - Portable Operating System Interface : Family of standards for maintaining compatibility between operating systems. IBM Storage Scale provides POSIX-compliant file system interface. : *Context: File system compatibility standard* : *Related terms: Unix, standards, file system*

S3 - Simple Storage Service : Object storage protocol originally developed by Amazon Web Services. IBM Storage Scale supports S3 protocol for object storage access. : *Context: Object storage protocol support* : *Related terms: object storage, cloud storage*

SDS - Software-Defined Storage : Storage architecture that separates storage software from hardware, enabling flexible deployment on commodity servers. IBM Storage Scale ECE uses SDS approach. : *Context: Deployment model for IBM Storage Scale* : *Related terms: virtualization, flexibility, x86 servers*

SMB - Server Message Block : Network file sharing protocol primarily used by Windows systems. IBM Storage Scale provides SMB protocol support for Windows client access. : *Context: Protocol support for Windows clients* : *Related terms: CIFS, Windows file sharing, NFS*

Computing & Hardware

DMA - Direct Memory Access : Technology that allows hardware subsystems to access system memory independently of the CPU, reducing CPU overhead and improving data transfer performance. : *Context: Underlying technology for GPUDirect Storage* : *Related terms: GDS, GPU, performance optimization*

DPU - Data Processing Unit : Specialized processor designed to offload and accelerate data-centric tasks from CPUs, including networking, storage, and security operations. NVIDIA BlueField-3 DPUs are used in NVIDIA AIDP. : *Context: Hardware component in NVIDIA AIDP architecture* : *Related terms: SmartNIC, CPU offload, BlueField*

GPU - Graphics Processing Unit : Specialized processor designed for parallel processing, widely used for AI/ML training and inference. IBM Storage Scale supports GPUDirect Storage for optimized data access. : *Context: Core accelerator hardware in NVIDIA AIDP* : *Related terms: CUDA, parallel processing, NVIDIA RTX*

NIC - Network Interface Card : Hardware component that connects a computer to a network. NVIDIA ConnectX NICs enable high-speed data movement in NVIDIA AIDP architecture. : *Context: Networking hardware for storage-compute connectivity* : *Related terms: networking, ConnectX, Spectrum-X*

Security & Access Control

ACL - Access Control List : List of permissions attached to an object that specifies which users or system processes can access the object and what operations they can perform. : *Context: File-level security inherited by vectors in CAS* : *Related terms: RBAC, security, permissions*

RBAC - Role-Based Access Control : Security approach that restricts system access based on user roles within an organization. IBM Fusion CAS enforces RBAC policies at the vector level. : *Context: Security model for RAG pipeline* : *Related terms: ACL, security, access control*

Search & Retrieval

BEIR - Benchmarking Information Retrieval : Standardized benchmark suite for evaluating information retrieval systems across diverse datasets and tasks. : *Context: Benchmark used to evaluate IBM Fusion CAS accuracy* : *Related terms: nDCG, evaluation metrics, search quality*

BM25 - Best Match 25 : Probabilistic ranking function used in information retrieval to estimate the relevance of documents to a search query based on term frequency and document length. : *Context: Keyword search algorithm in hybrid retrieval* : *Related terms: keyword search, ranking, information retrieval*

nDCG - Normalized Discounted Cumulative Gain : Evaluation metric for ranking quality in information retrieval that measures how well the ranking matches the ideal ordering of results. : *Context: Metric for evaluating RAG retrieval accuracy* : *Related terms: BEIR, evaluation, search quality*

Operating Systems & Platforms

API - Application Programming Interface : Set of protocols and tools for building software applications. Defines how software components should interact. : *Context: IBM Fusion CAS Search API for vector queries* : *Related terms: integration, programming, interface*

RHEL - Red Hat® Enterprise Linux® : Commercial Linux distribution developed by Red Hat. Supported platform for IBM Storage Scale and IBM Fusion deployments. : *Context: Operating system for software-defined storage* : *Related terms: Linux, enterprise OS, x86*

Notices

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Trademarks

IBM®, the IBM logo, ibm.com®, IBM Redbooks®, Db2®, IBM Storage Scale, IBM Fusion, IBM Content-Aware Storage, and Maximo® are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Red Hat®, OpenShift®, and Red Hat Enterprise Linux® are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

Kubernetes® is a registered trademark of The Linux Foundation in the United States and other countries.

NVIDIA, the NVIDIA logo, BlueField, ConnectX, GPUDirect, NeMo, RTX, and Spectrum-X are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries.

Other company, product, or service names may be trademarks or service marks of others.

Version History

- **Version 1.0** (2026-04-15): Initial publication
- **Version 1.1** (2026-04-22): Added link to "Accelerate research using NVIDIA AI-Q and NVIDIA RAG Blueprint on IBM Fusion HCI" white paper in Related Publications section

AI Attribution

This work was primarily human-created. AI was used to make stylistic edits, such as changes to structure, wording, and clarity. AI was used to make content edits, such as changes to scope, information, and ideas. AI was prompted for its contributions, or AI assistance was enabled. AI-generated content was reviewed and approved. The following model(s) or application(s) were used: IBM Bob, IBM Consulting Advantage.