Content-Aware IBM Storage Scale

Generate better AI answers by unlocking the semantic meaning hidden inside unstructured enterprise data

Highlights

Content-aware IBM Storage Scale uses natural language processing to extract meaning from unstructured data

Leverages NVIDIA NeMo Retriever, NVIDIA NIM Microservices, and the NVIDIA AI Data Platform

RAG databases are updated automatically as your data changes, so AI tools generate more accurate answers

Content-aware capabilities extend to data stored on third-party systems without data migration or duplication

As generative AI and agentic AI become a routine part of every enterprise, IT resources are shifting from training large language models to running applications on top of them. This process, known as inferencing, is how an AI assistant or chatbot can ingest a question and use the model to infer the best answer.

But the available large language models (LLMs) were mostly trained on public data – very little enterprise data was used to train their underlying models, potentially limiting their value.

To generate high-quality answers, AI tools must have access to information beyond the data used in their initial training sets. Retrieval augmented generation (RAG) improves inferencing by incorporating new information beyond the original machine learning model. This allows AI assistants and other chatbots to generate more accurate, timely, and relevant results, rather than relying only on incomplete or outdated training data.

Preparing and ingesting corporate data for RAG pipelines is a costly and time-consuming process, so organizations typically ingest only a portion of their data, refreshing it in batch mode from time to time.

That approach presents major challenges:

- Inaccurate answers – AI tools aren't magic; if they're fed incomplete or outdated information, they'll generate low-quality responses.
- High cost – RAG workflows typically involve transferring and copying data multiple times, increasing storage, networking, and compute costs.
- Data security – The proliferation of data copies increases security risks.
- Operational challenges – It takes considerable expertise to architect and deploy GPU-accelerated compute, storage, and networking infrastructure.

To meet these challenges, IT leaders today are moving toward a new paradigm that leverages the significant advantages of bringing intelligent data processing, such as vector processing, closer to the storage layer.

Instead of the storage system being a passive “black box” that just holds the 0s and 1s, it becomes an active participant in data preparation tasks – accelerating data processing pipelines and AI models.

Content-aware storage (CAS) is based on major innovations from IBM Research that use natural language processing to extract the semantic meaning hidden inside unstructured enterprise data. The document data extraction workflow also leverages NVIDIA NeMo Retriever microservices, built with NVIDIA NIM, part of NVIDIA AI Enterprise.

Content-aware storage improves RAG workflows by ensuring that responses from AI assistants and agents are informed by the latest relevant, contextual information. It enhances the value of AI applications through faster time to insights, reduced costs, improved performance, better security, and streamlined operations.

Content-Aware Storage

Enterprises today create massive and ever-increasing volumes of data, and much of it is unstructured – PDFs, chats and emails, audio and video files, social media posts, presentations, and legal and financial documents.

IBM Storage Scale provides organizations with a global data platform which creates a single namespace across multiple sites, across clouds, between data centers, and out to the edge. Storage Scale also provides a unique abstraction service to extend the global namespace to third-party data stores, providing access to existing legacy data stores and breaking down silos to provide access to all your data. It’s based on a massively parallel file system and can be deployed on multiple platforms including x86, IBM Power, IBM zSystem mainframes, Arm-based POSIX client, virtual machines, and containers (i .e.: Kubernetes).

Content-aware storage is a new AI-based capability in Storage Scale, designed to help organizations derive greater business value from their existing data stores.

Retrieval Augmented Generation
Retrieval augmented generation and many other AI processes rely on natural language processing (NLP) algorithms to parse text, break it into discrete chunks, and extract each chunk’s semantic meaning – distinguishing the “bank” that’s a financial institution from the “bank” on the edge of a river, or a “bank shot” in billiards.

A separate AI process called embedding converts these semantics into vectors (a series of numbers) using AI models like those in NeMo Retriever. To understand vectors, think of them as a way to turn words and sentences into coordinates in a multi-dimensional space. Similar ideas are placed closer together, while unrelated ones are farther apart – the financial “bank” vector is near the vectors for “mortgage provider” and “savings and loan” but far away from the vector for “broccoli.”

It’s this property that allows generative AI applications to compare vectors and retrieve information that’s not just keyword-matching but actually relevant in meaning.

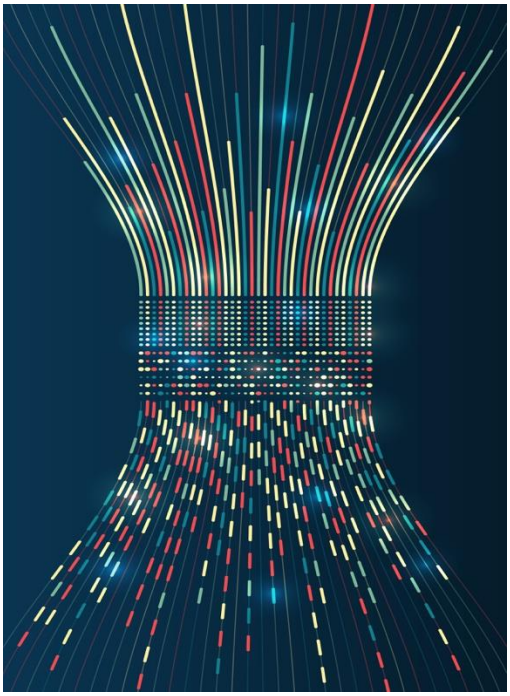


Figure 1. Content-aware storage relies on natural language processing tools, which convert raw text into vectors containing semantic meanings.

The vectors are stored in a database, allowing the system to efficiently search for relevant information. When a user asks a question, their query is converted into a vector and compared against the stored document vectors using similarity measures. The most relevant text chunks are retrieved and added to the query before being sent to a large language model, which generates a response based on both the retrieved information and its general knowledge.

By using this approach, RAG improves accuracy, reduces hallucinations, and allows AI models to stay updated with new information without retraining, making it ideal for research, customer support, and knowledge-based applications.

But there's no free lunch – to effectively deploy AI workloads, IT organizations must now create and regularly update databases of vector embeddings based on their enterprise data. Re-vectorization is expensive, so it's done infrequently. The result is vector databases that are out of date. For example, the RAG vector databases for many large enterprise websites are only rebuilt on a weekly basis.

Another problem is that RAG workflows often copy data from original sources to data lakes, then copy it again to one or more cloud services for data preparation and vectorization. And many of the copies must be vectorized, consuming scarce and expensive GPU resources, so organizations may only vectorize a fraction of their enterprise data, constraining the quality of AI responses.

IBM Storage has a better way.

Content-Aware IBM Storage Scale

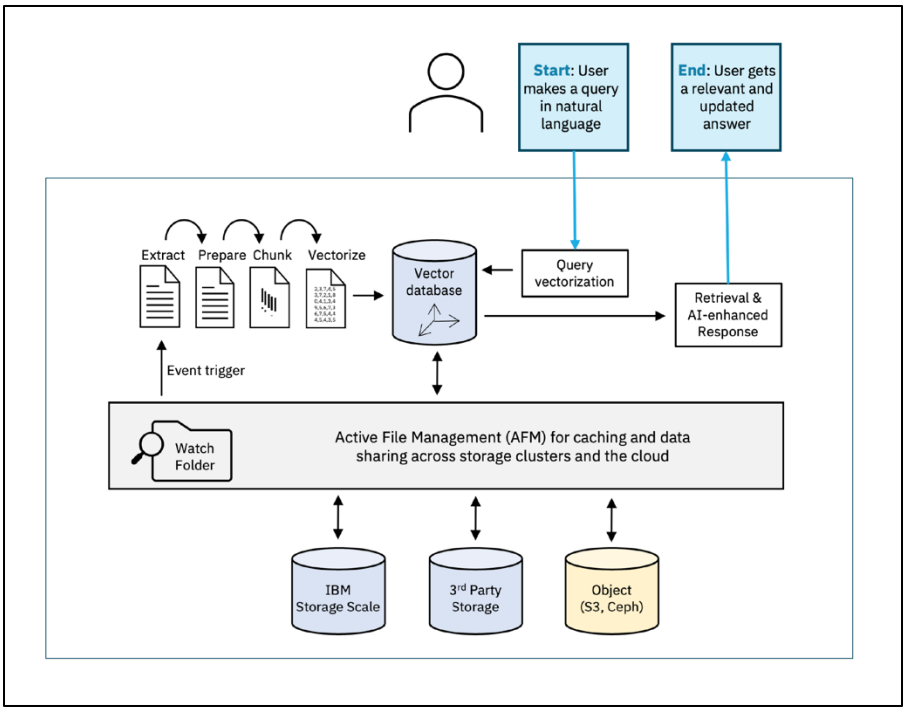
IBM has done innovative work on many aspects of AI data pipelines, including parsing, chunking, vector embedding models, and hardware acceleration. Indeed, some of these innovations are at the heart of IBM watsonx and are responsible for many of its differentiated capabilities for AI training and inferencing on premises or in the cloud.

Adding content-aware capabilities to Storage Scale enables enterprises to integrate compute, data pipelines, and a vector database with the storage system to provide efficiency gains, faster time-to-value, and access to real-time data. Content-aware storage also reduces data movement over the network and can orchestrate data pipelines efficiently by caching data locally and enabling data artifact persistence to relevant storage tiers, ensuring that data is refreshed as soon as possible.

The initial release prescribes a purpose-built AI pipeline that is an IBM curated version of NVIDIA NIM & the NVIDIA multi-modal PDF data extraction blueprint. This pipeline, built to tightly integrate with NVIDIA & IBM infrastructure technology, is finely tuned to ensure that AI assistants and agents consistently provide enterprise-grade accuracy

NVIDIA NIM is a key part of the architecture for content-aware Storage Scale. It builds on NVIDIA NeMo – an open-source AI framework for building, training, and deploying large-scale deep learning models, particularly for tasks like natural language processing. NVIDIA NIM helps optimize inferencing performance by providing pre-packaged, accelerated AI microservices that can run on NVIDIA accelerated computing across all environments, including enterprise data centers, private clouds, and hybrid infrastructures.

Figure 2. Technical architecture for Content-aware IBM Storage Scale



Content-aware Storage Scale provides customers with:

- Faster time to insights – The storage layer has fine-grained knowledge of all data changes, which enables rapid incremental updates to the vector database, avoiding the complete rebuilds that are common today. This is a huge difference; if you ask an enterprise chatbot for a summary of all your competitors' news releases this week it generates exactly the list you want rather than a "not in my model" response.
- Reduced costs – GPU requirements are significantly reduced because CAS performs incremental updates to the vector database when new there's activity in a watched folder, rather than rebuilding it from scratch each time. The CAS architecture also requires fewer replicas of data and includes optimizations that leverage semantic understanding of the data, i.e.: deduplication and decryption avoidance, further reducing the requirements for GPU resources.
- Improved performance – Storing vectors adjacent to their source data enables the use of NVIDIA GPUDirect Storage – a protocol for high-speed communication between storage systems and GPUs via RDMA. The result is a significant performance boost. In addition, CAS leverages storage tiering to preserve intermediate outputs in the vectorization pipeline, which can significantly reduce the cost of upgrading the embedding model.
- Simplified operations – Encapsulating the vector database and data ingest pipeline within storage simplifies the architecture and reduces skill requirements both for deployment and ongoing operations.

Automating AI Pipelines

Embedding compute, data pipelines, and vector database capabilities within the storage system (i.e.: doing the vectorization near where the data resides) minimizes data movement and latency, resulting in significant efficiency gains.

Content-aware Storage Scale can leverage watch folders to identify changes as they occur, automatically run pre-built pipelines, and update the vector database, helping to ensure that data is always current for AI applications.

Storage Scale also has a unique active file management capability that provides seamless access to enterprise data stored on existing third-party storage systems without migration or duplication of data, bringing the benefits of content-aware storage to non-IBM storage systems. Storage Scale provides shared multi-protocol data access – the ability to ingest data via one protocol and make it available to multiple workloads simultaneously, even workloads using different data protocols.

Storage Scale provides live notifications of changes in connected storage systems, enabling almost real-time synchronization between remote data sources and the vector database by only processing delta changes. An advanced caching and abstraction service for backend storage provides low-latency read and write access to remote data sources, avoiding the need to make full copies of data, and ensuring that updates are captured quickly and efficiently for downstream applications.

Storage Scale supports the new NVIDIA AI Data Platform, which brings enterprise storage into the era of agentic AI and helps IT leaders build distributed systems that unlock the full value of business data to fuel data-driven actions. Built on NVIDIA's expertise in AI workflow optimization, the platform is a customizable reference design for integrating NVIDIA accelerated computing, networking, and AI software with enterprise storage, transforming data into actionable intelligence.

For more information
To learn more about IBM Storage Scale, contact your IBM representative or IBM Business Partner, or visit https://www.ibm.com/products/storage-scale.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
March 2025