Enhancing AI Results with Content-Aware Storage

Client L1

Vincent Hsu – IBM Fellow, VP and Chief Technology Officer, IBM Storage
Patrick Fay – Product manager, Content-aware Storage
Mike Kieran – Marketing manager, IBM Storage Scale

May 15th, 2025

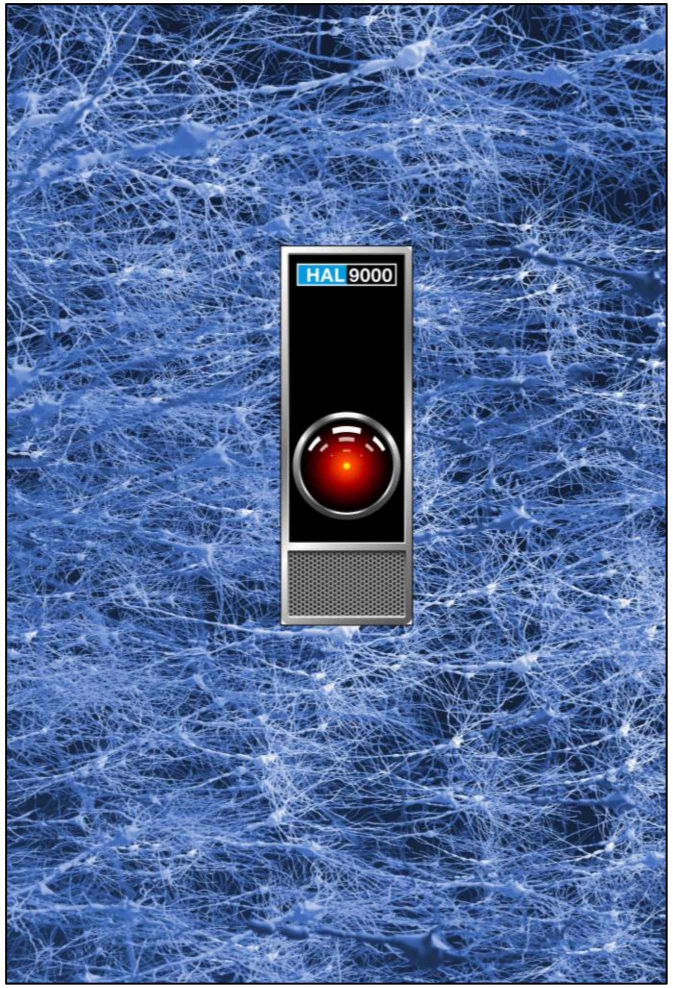
Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

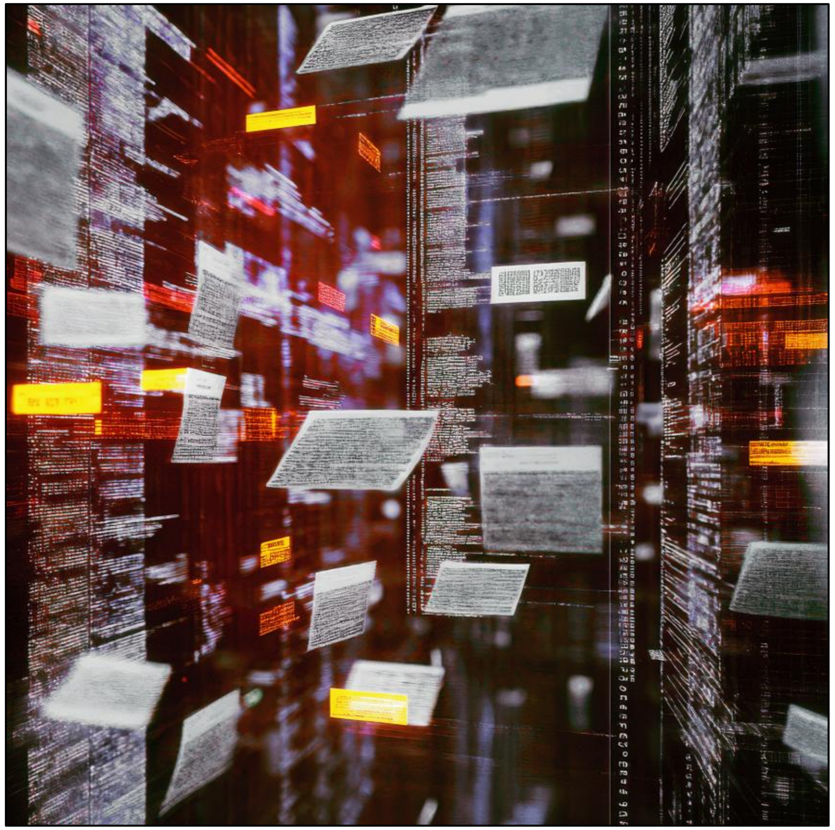
Trustworthy AI

- Chatbots and other AI assistants have become crucial tools throughout the enterprise.
- These are *inferencing* apps, which process the user's query against a large language model (LLM) to *infer* the best answer.
- But for AI systems to generate trustworthy answers, they must have access to complete and accurate real-time information beyond the original training set.



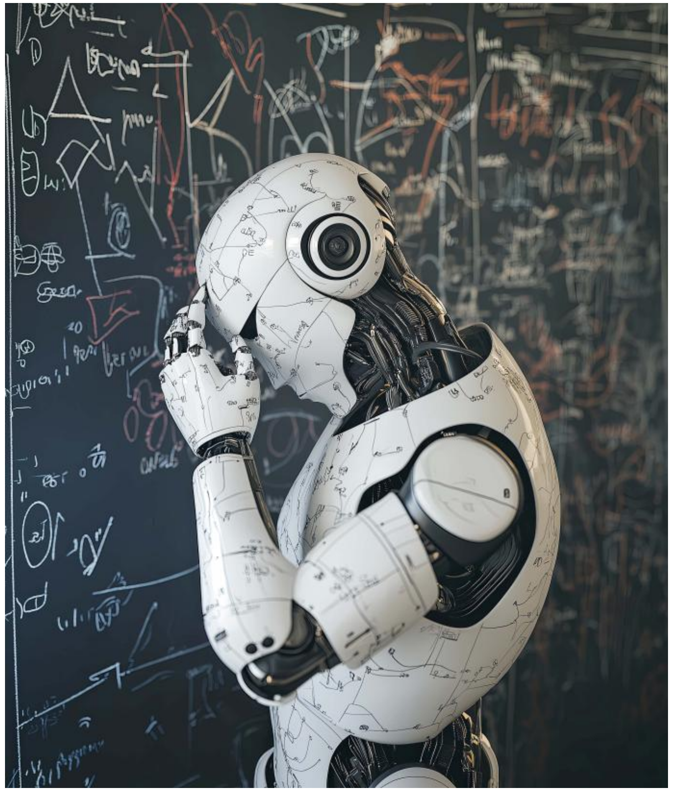
The 1% problem

- Organizations are swamped with unstructured data.
- But less than 1% of all enterprise data was used to train major large language models.
- Retrieval augmented generation (RAG) improves inferencing by incorporating near real-time data.



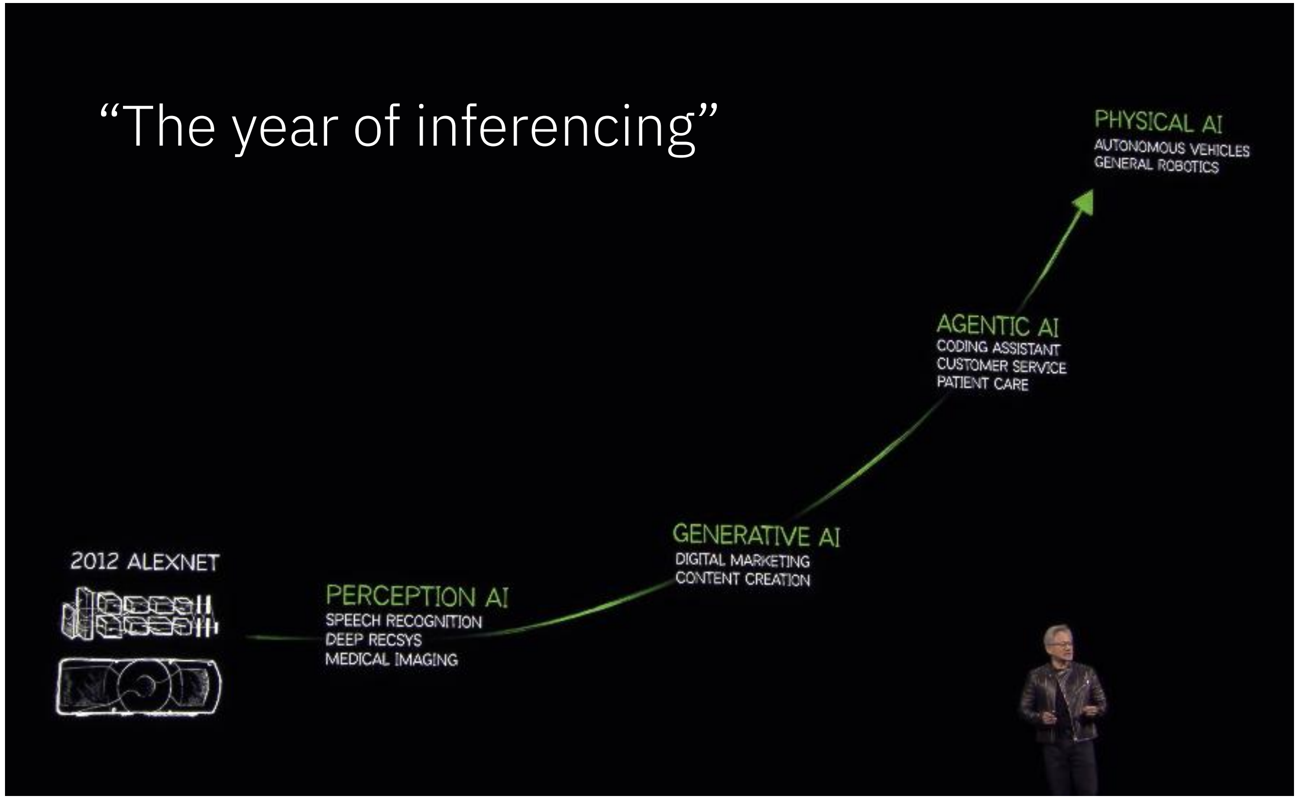
Significant challenges for the AI enterprise

- **Inaccurate answers**
- High costs
- Data security
- Operational challenges



From training AI models to inferencing

- AI infrastructure requirements are shifting from *training* AI models to *running* them – inferencing.
- At NVIDIA GTC in March, Jensen Huang called 2025 “the year of inferencing.”



“The year of inferencing”

Two key elements to inferencing

1. Data ingestion process

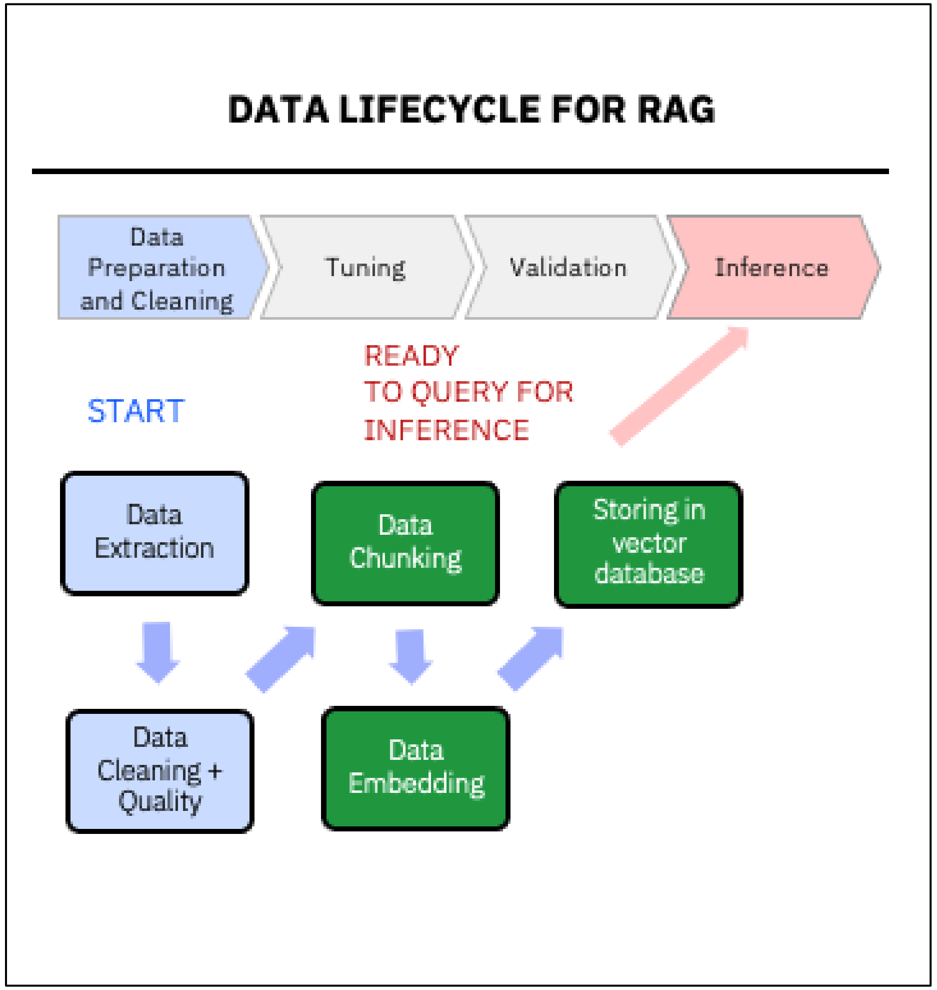
   – Natural language processing is used to parse, chunk, and embed the *meanings* of text.

2. Inferencing operations

   – Queries are processed by a large language model, enhanced via retrieval augmented generation.

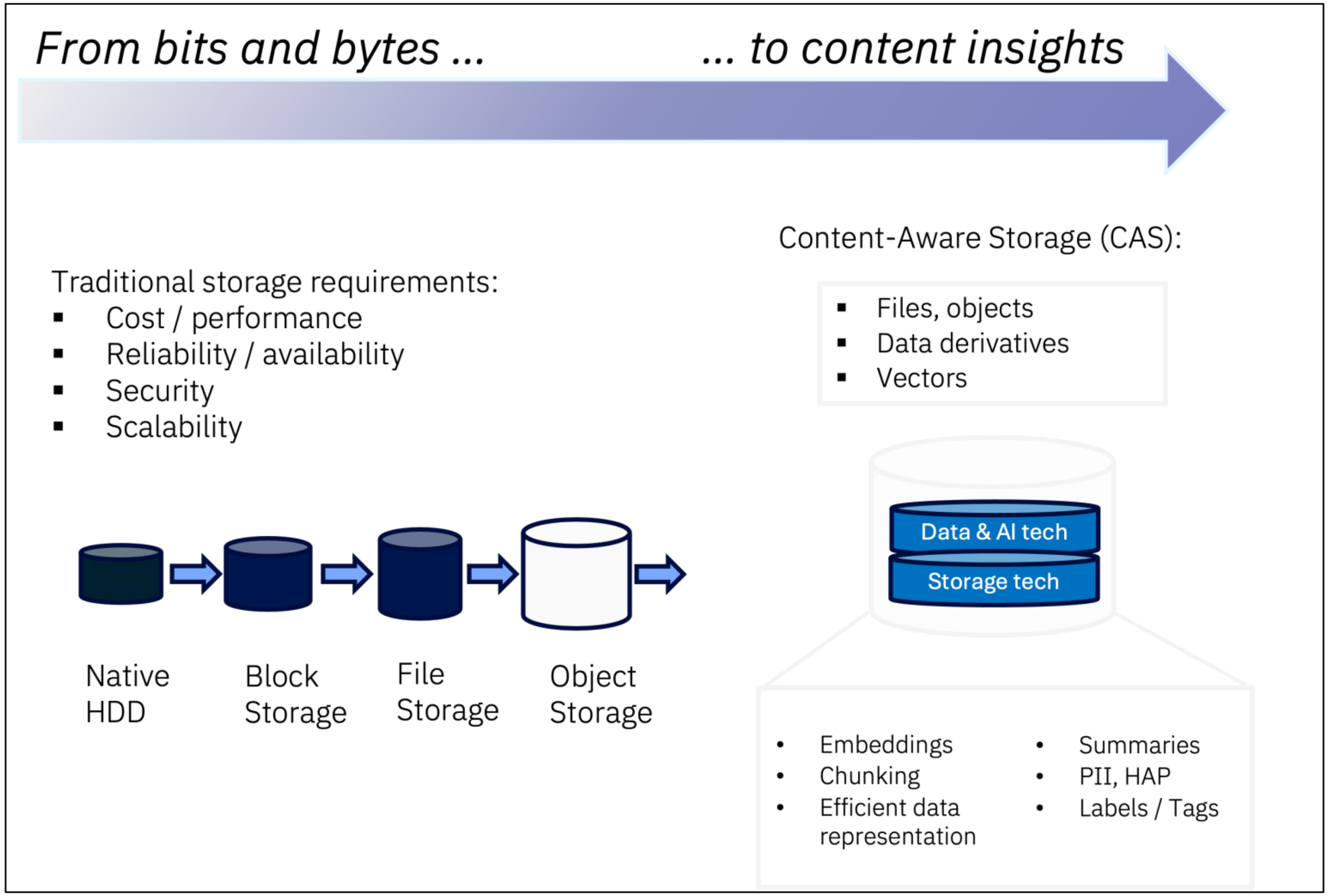
Limitations in existing RAG solutions

- **Only a fraction of enterprise data is indexed.**
- It’s typically indexed daily or weekly in batch mode, not real time.



The solution? Content-aware infrastructure

- A new paradigm that leverages bringing vector processing closer to the storage layer.
- From “bringing data to AI” to “bringing AI to data”.



From bits and bytes … *… to content insights*

Traditional storage requirements:

- Cost / performance
- Reliability / availability
- Security
- Scalability

Content-Aware Storage (CAS):

- Files, objects
- Data derivatives
- Vectors

- Embeddings
- Chunking
- Efficient data representation
- Summaries
- PII, HAP
- Labels / Tags

Announcing: Content-aware IBM Storage Scale

- Significantly improves inferencing.
- Helps organizations derive greater business value from PDFs, presentations, audio and video files, emails, social media posts.
- Applies natural language processing tools to extract the meaning inside these data sources.

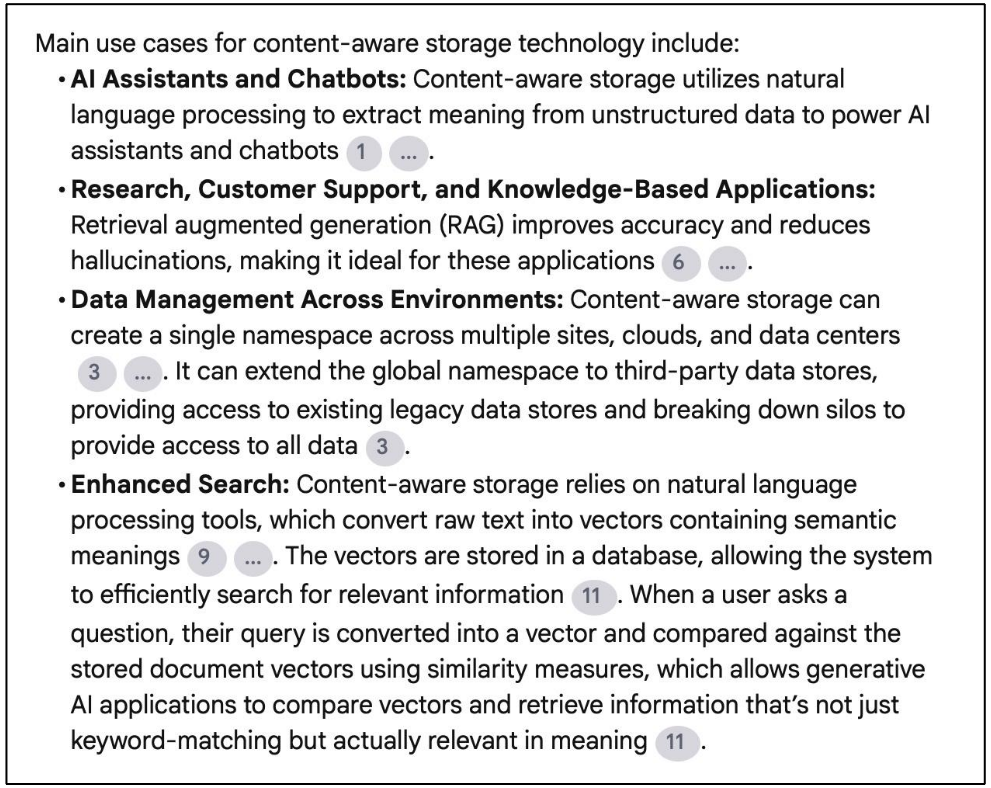
Look, Dick.
Look, Jane.
Look and see.
See Baby.
See, see.

Customer benefits

- Cognitive edge – faster time to insights
- Reduced costs
- Improved performance
- Simplified operations

Use cases for content-aware Storage Scale

- AI assistants and agents
- Real-time data sync
- Streamlined AI data pipelines
- Enhanced search

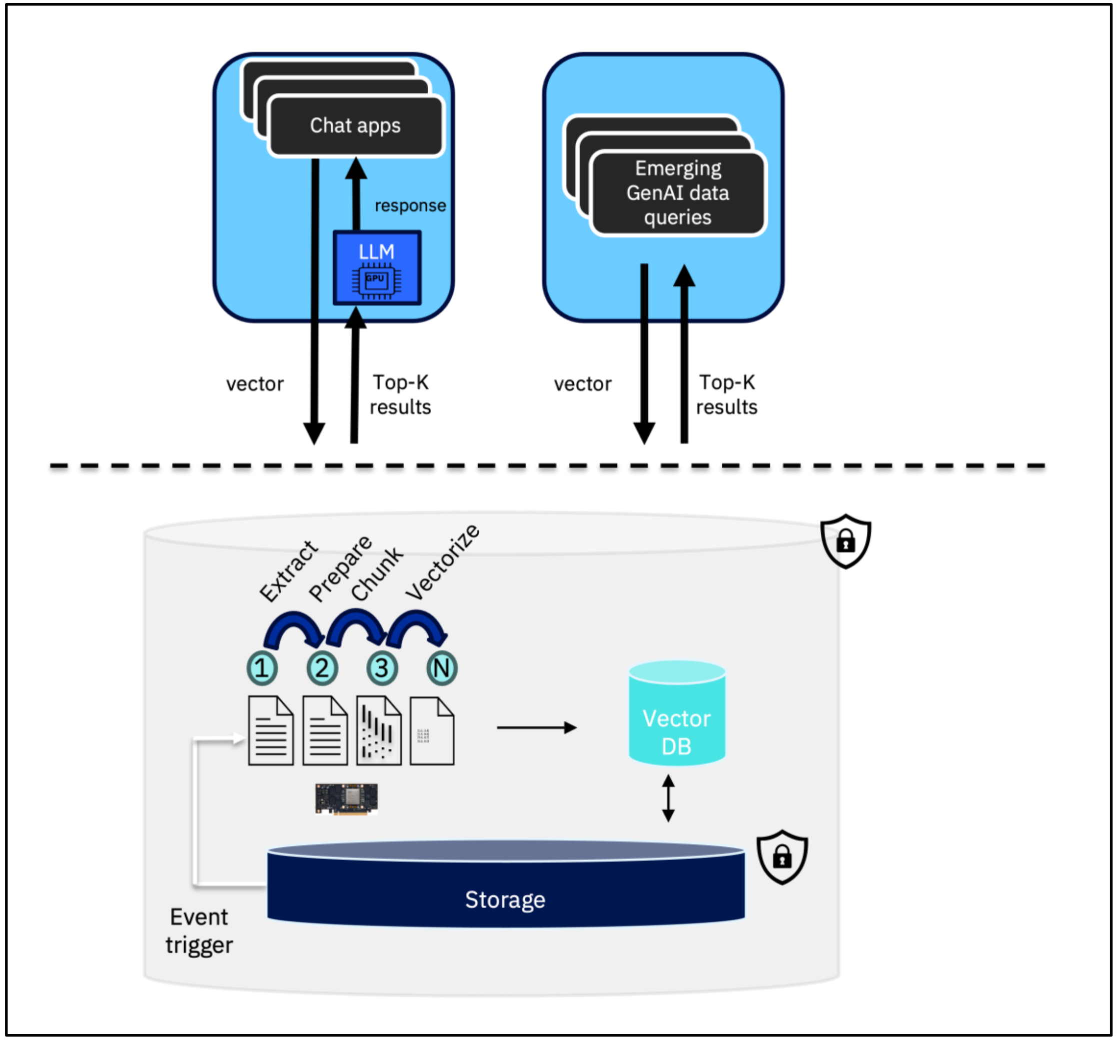


Main use cases for content-aware storage technology include:

- **AI Assistants and Chatbots:** Content-aware storage utilizes natural language processing to extract meaning from unstructured data to power AI assistants and chatbots [1] [...].
- **Research, Customer Support, and Knowledge-Based Applications:** Retrieval augmented generation (RAG) improves accuracy and reduces hallucinations, making it ideal for these applications [6] [...].
- **Data Management Across Environments:** Content-aware storage can create a single namespace across multiple sites, clouds, and data centers [3] [...]. It can 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 data [3].
- **Enhanced Search:** Content-aware storage relies on natural language processing tools, which convert raw text into vectors containing semantic meanings [9] [...]. The vectors are stored in a database, allowing the system to efficiently search for relevant information [11]. When a user asks a question, their query is converted into a vector and compared against the stored document vectors using similarity measures, which allows generative AI applications to compare vectors and retrieve information that’s not just keyword-matching but actually relevant in meaning [11].

Foundational technologies for content-aware storage

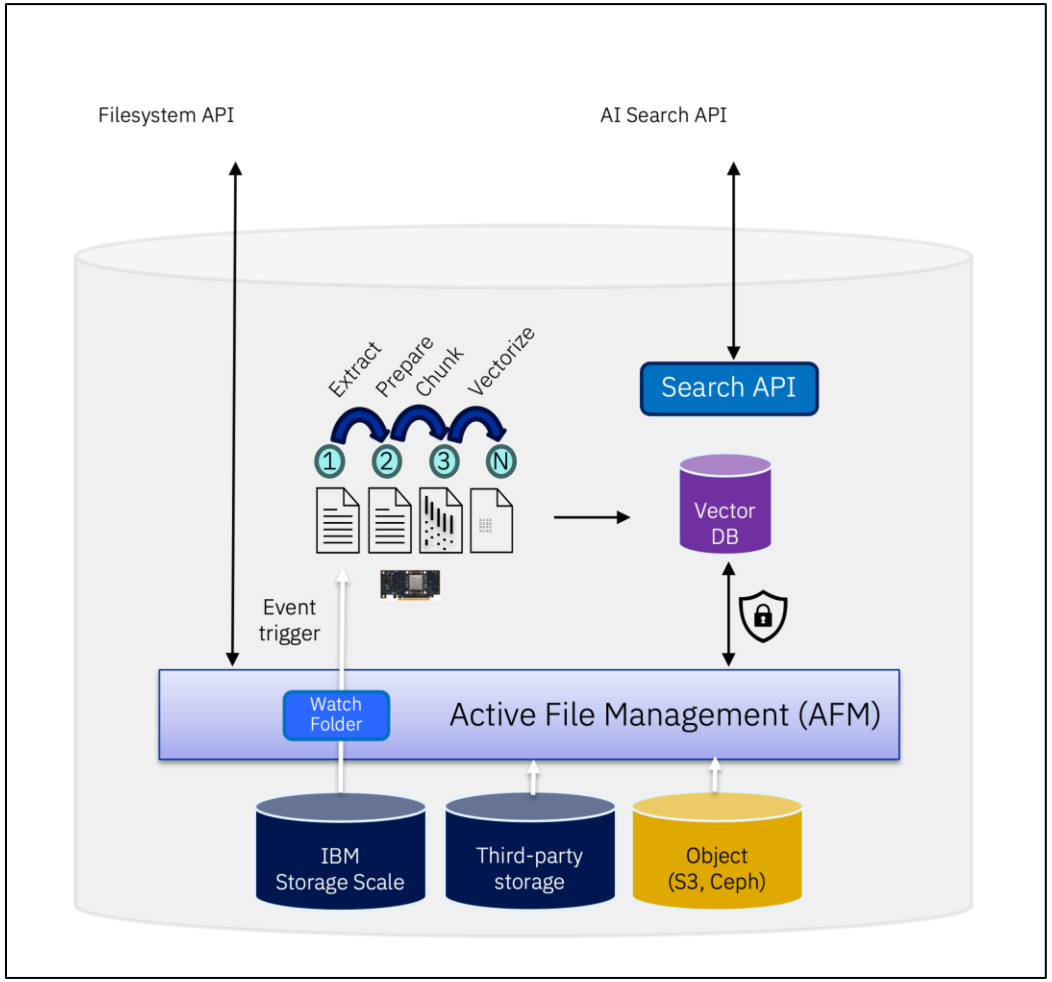
- Content-aware storage leverages:
  - AI optimized storage – IBM Storage Scale
  - AI data pipelines, such as NVIDIA NIMs
  - Vector databases and metadata model
  - Hardware accelerators (such as GPUs)



Content-aware storage leverages AI storage and data processing pipelines

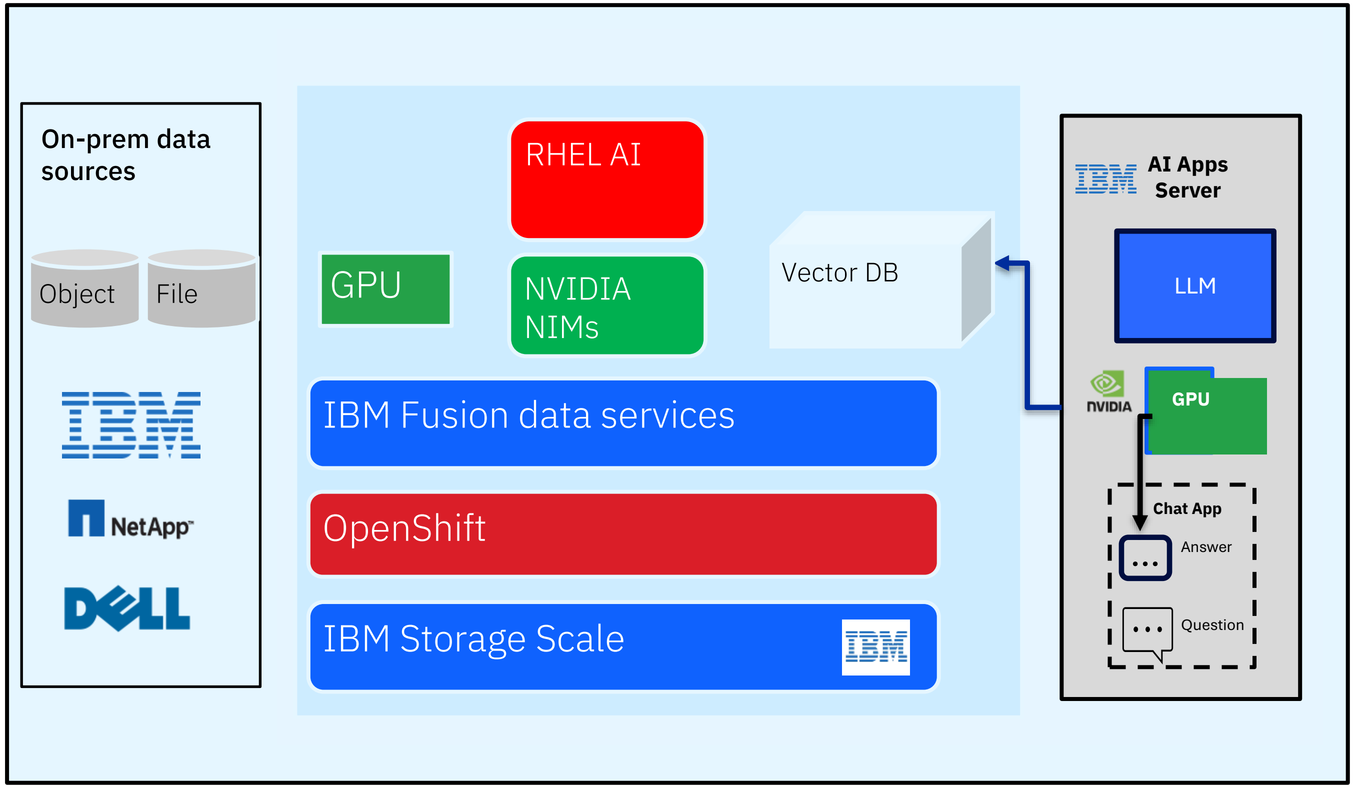
Content-aware legacy storage on-premises or in the cloud

- Storage Scale's active file management abstracts other storage systems, including legacy IBM and third-party systems
- Automatically detects data changes for incremental processing
- Supports NVIDIA GPUDirect acceleration.



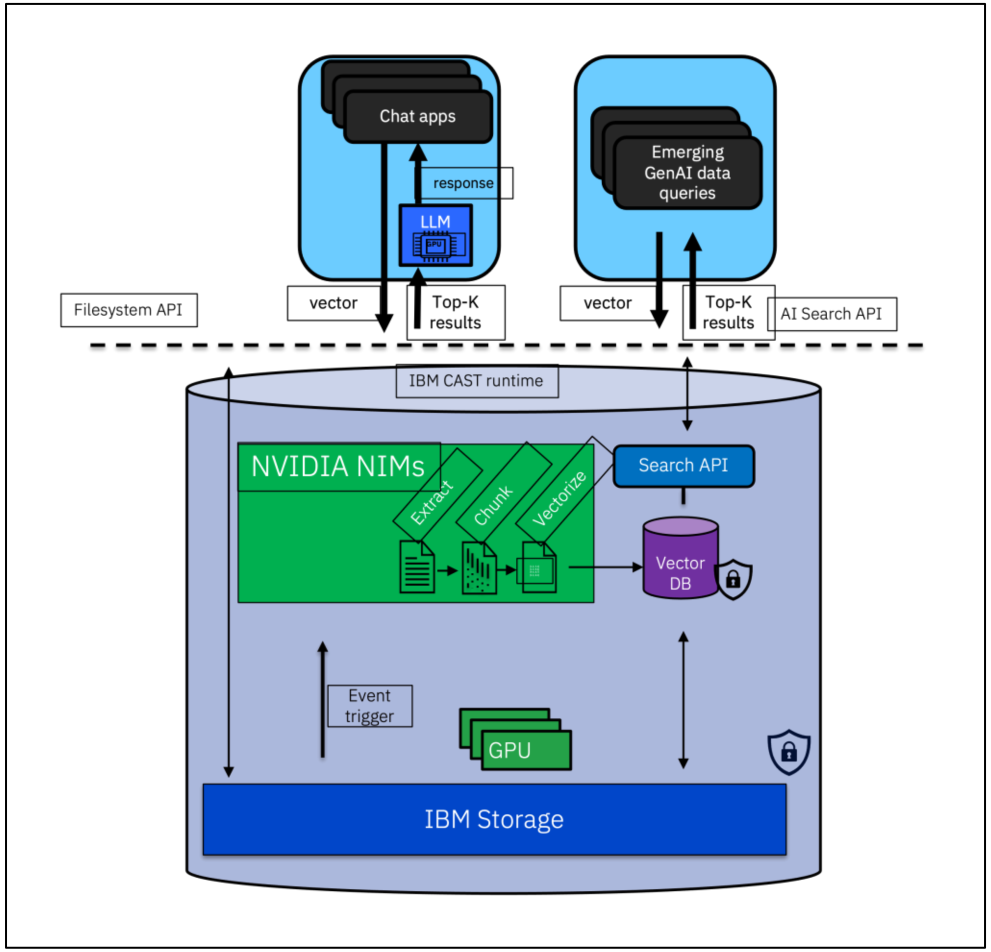
Content-aware storage software architecture

- Provides flexibility to integrate data pipelines with IBM Fusion data services
- Integrates Storage Scale for data access and storage optimization
- Integrates advanced vector database and enables hardware acceleration



NVIDIA multimodal PDF data extraction

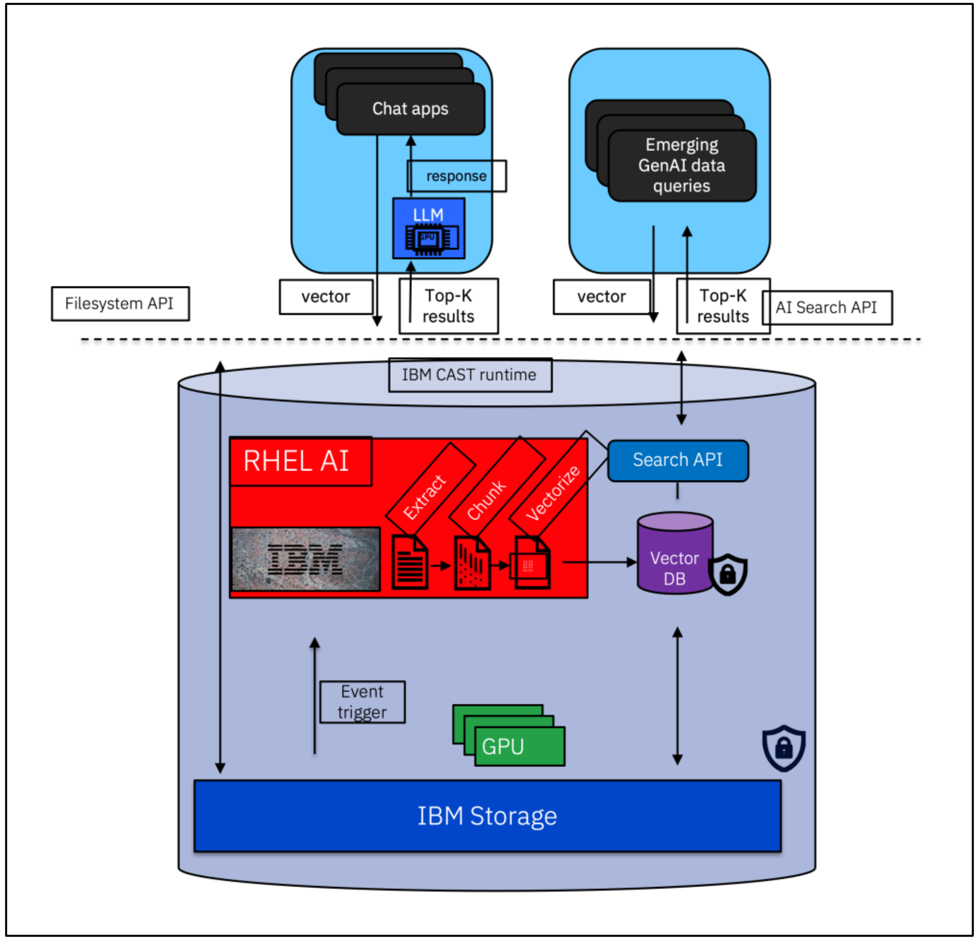
- Leverages NeMo Retriever with support for text extraction from charts and images.
- Includes a vector database that scales to 1B vectors and beyond while preserving data source ACL permissions.
- Supports RAG vectorization for enterprise data, with incremental data processing.
- Uses NVIDIA NIM for inferencing and NVIDIA L40S GPUs for hardware acceleration.



Content-aware IBM Storage Scale with NVIDIA NIM data pipeline

Integrating open-source data processing for RAG

- A widely used open-source data pipeline.
- Leverages IBM innovations in AI parsing and embedding.
- Supports 1B vector database while preserving data source ACL permissions and incremental RAG vectorization.
- Currently uses NVIDIA L40S GPUs.



Content-aware IBM Storage Scale with RHEL AI data pipeline

Unlocking the business value in unstructured data

- Content-aware Storage Scale can help organizations gain a cognitive edge – smarter answers, faster time to insights.
- It can help accelerate inferencing, simplify operations, and reduce TCO.
- The architecture is designed to enable content-aware capabilities across an enterprise’s entire data estates without copying or replicating data.
- Improves data security by retaining data source access controls in derivatives such as chatbot responses.

IBM