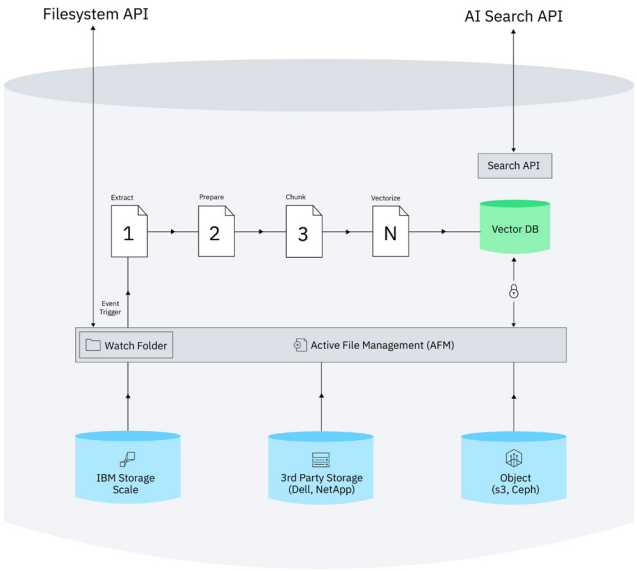
Transforming AI with content-aware storage for data-driven insights

Get smarter answers from AI assistants and agents by uncovering insights hidden in your data



Generative AI (gen AI) has transformed how businesses interact with data, but a major limitation remains—most organizations haven't leveraged their data to train their large language models (LLMs). As a result, the AI model's ability to provide accurate, up-to-date and contextually relevant responses to business-critical questions might be severely restricted.

For example, if a team member asked an AI assistant what happened in their market or competitive space in the past week, the response might be incomplete, inaccurate or outdated because the AI model lacked access to the enterprise's proprietary data.

Organizations can attempt to bridge this gap using retrieval-augmented generation (RAG). However, traditional approaches can be inefficient and costly. Downsides include:

- **Outdated data:** Vector databases don't track real-time changes and costly revectorization happens infrequently.
- **High costs:** Data is copied multiple times, driving up storage, networking and GPU expenses.
- **Security risks:** Data duplication increases exposure to breaches and complicates access control, requiring fragile sync processes.
- **Operational complexity:** Deploying and managing GPU-accelerated compute, storage and networking demands specialized expertise, adding to the operational burden.

Meet IBM Storage Scale

IBM® Storage Scale is a content-aware solution designed to enhance AI by embedding intelligence directly into the storage layer, providing fast, efficient and secured AI workflows. IBM Storage Scale software provides global data abstraction services that seamlessly connect multiple data sources across multiple locations, including non-IBM storage environments. It's designed to help your organization achieve:

- **Faster time to insights:** Track data changes in near real-time,
- resulting in rapid updates to vector databases without costly full rebuilds.
- **Reduced costs:** Minimize data replication and GPU usage with incremental updates and semantic optimizations.
- **Improved performance:** Leverage the power of NVIDIA GPUDirect for accelerated data transfers and streamline the embedding of model upgrades.
- **Enhanced security:** Align vector queries with role-based access control (RBAC) policies, eliminating the complexity of access control synchronization.
- **Streamlined operations:** Integrate your organization's vector database into storage, streamlining deployment and reducing management complexity.

Content-aware storage, driven by IBM Storage Scale, can bring intelligence closer to enterprise data, helping keep AI models current, efficient and secured. By transforming how AI interacts with data, organizations can gain data-driven insights, better performance and reduced costs, unlocking a true competitive edge.

Learn more about how content-aware storage transforms AI.
Read the brief →