Enterprise Strategy Group™

BRIEF

MAY 2025

IBM Aims to Boost AI Inferencing With 'Content Aware' Storage

Simon Robinson, Principal Analyst

Overview

IBM made a number of announcements recently that highlight its ongoing commitment to help customers accelerate their AI journeys. Of particular note is a pending capability developed by IBM Research that adds 'content awareness' to both its own and third-party storage, which when integrated with the NVIDIA Data Platform will bring vector processing closer to the storage layer to substantially improve the effectiveness of RAG-based inferencing.

Analysis

The potential of generative AI continues to take hold across organizations globally: According to Enterprise Strategy Group, over half of respondents (53%) to a recent survey said that their most significant AI investment over the next 12 months would be in generative AI (GenAI).[1] For many organizations, their GenAI implementation will be hybrid, typically spanning a combination of on-premises data with public cloud-based large language models (LLMs).

However, the devil is in the details, and for many organizations there's a considerable gap between their high levels of anticipation around AI, and the realities of an effective implementation. For example, Enterprise Strategy Group research found that 62% of organizations had experienced 'extensive' or 'moderate' challenges moving AI models from development into production.[2]

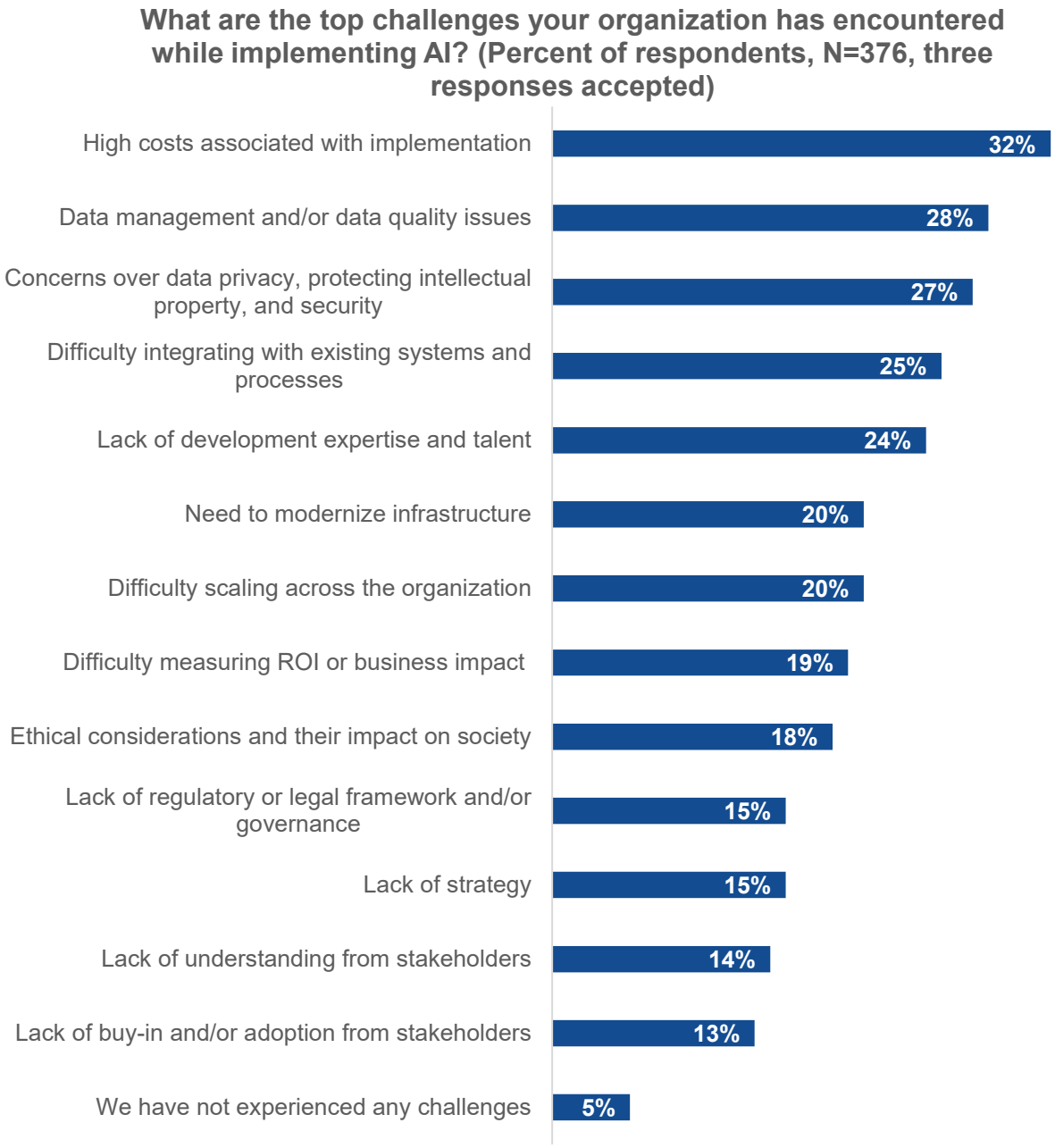
Enterprise challenges around implementing AI are broad and vary widely. But while much of the initial focus was focused on the compute environment, many challenges are now emerging at the data level.

Data is the lifeblood of any AI initiative—the one aspect that can make the difference between success and failure. Getting the data aspect right at scale presents numerous substantial challenges across the broader data environment. A recent Enterprise Strategy Group research study found that data management and/or data quality issues were the second most frequently cited challenge associated with implementing AI, behind overall cost issues (see Figure 1).[3] Concerns over data privacy, protecting intellectual property, and security were also frequently cited, along with integration issues and the need to modernize infrastructure.

Many organizations on their AI journeys are, therefore, concluding that modernizing the infrastructure—right down to the storage environment—might be a necessary step to fully take advantage of AI, especially as they move from model training to the inference phase.

[1] Source: Enterprise Strategy Group Research Report, *2025 Technology Spending Intentions Survey*, December 2024.
[2] Source: Enterprise Strategy Group Research Report, *Navigating Build-versus-buy Dynamics for Enterprise-ready AI*, January 2025.
[3] Ibid.

Figure 1. Top Challenges With Implementing AI



Source: Enterprise Strategy Group, now part of Omdia

IBM Unveils Expanded AI Focus With CAS

IBM has a broad, multifaceted approach to AI that encompasses platforms and models (e.g. watsonx, Granite), open-source initiatives, a range of education, training and consulting capabilities, and a comprehensive partnership strategy with key ecosystems players such as NVIDIA.

In-house technology development is a core pillar of IBM's approach to AI, and its efforts in the storage and data realm are a prime example. On March 27, the company launched a "content aware storage" (CAS) capability that will form part of its hybrid cloud infrastructure offering, IBM Fusion. It will be offered as part of IBM Storage Scale (formerly known as General Parallel File System, or GPFS), IBM's high performance, scale-out file system, and will

also leverage a range of NVIDIA capabilities, including the NVIDIA Data Platform, NVIDIA Spectrum-X networking and NVIDIA NIM.

The aim of CAS is to enable enterprises to more fully take advantage of AI inference capabilities that sit on top of LLMs but leverage their own, proprietary data sets through capabilities such as retrieval-augmented generation (RAG).

Currently, the task of preparing and ingesting enterprise data for RAG is typically costly and time-consuming. Data must be copied from its original source into multiple other destinations—data lakes, cloud services and so on—for preparation and vectorization. This limits both the volume of data that enterprises are able to move into RAG, and the frequency by which they can refresh this data, which ultimately limits the value of the inferencing process, yielding low-quality results, high costs, increased risk (through multiple copies) and operational challenges.

By contrast, IBM believes that by "bringing the AI to the data" it can address all of these limitations in one fell swoop, drastically improving the quality of responses in the process. Its approach with CAS essentially transforms IBM Storage Scale from a passive storage system storing 1s and 0s into an intelligent infrastructure layer that's an integral part of the AI data preparation process.

Leveraging significant innovations from IBM Research around natural language processing, CAS extracts semantic meaning from unstructured data—such as PDFs, chats, emails, audio and video files, legal, financial, and other business documents—from within the storage infrastructure itself. This reduces the number of steps required for inferencing, meaning it can be done more quickly, more frequently, and with greater overall efficiency, by minimizing data movement and latency. A neat aspect is that users can set up "watch" folders to identify data changes as they occur, helping ensure that data is always current for AI applications. What's more, Storage Scale's support for third-party storage systems—such as from Dell, NetApp, and others—means that CAS will be able to support data stored in a wide range of corporate storage repositories without having to move or migrate this data.

The initial release of CAS will support a purpose-built AI pipeline running an IBM version of NVIDIA NIM and the NVIDIA multimodal PDF extraction blueprint, finely tuned to ensure that AI assistance and agents consistently provide enterprise-grade accuracy.

Conclusion

The promise of emerging technologies such as generative AI is vast, but the barriers to a successful implementation for the average enterprise can often appear daunting. Moreover, organizational appetite for longer-term investments in AI capabilities might diminish if initial returns around inferencing are underwhelming, or even plain wrong.

Hence, pending innovations such as IBM's content aware storage provide organizations with a pathway to inferencing that both increases the likelihood of driving strong, positive outcomes, while potentially significantly reducing infrastructure resource consumption at the compute/GPU, network, and storage layer. IBM already has a strong pedigree in HPC- and at-scale performance storage with Storage Scale/GPFS. For organizations looking to get the most out of their initial AI inferencing investments, IBM's evolving capabilities in this sphere are well worth a look.

About Enterprise Strategy Group
Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com