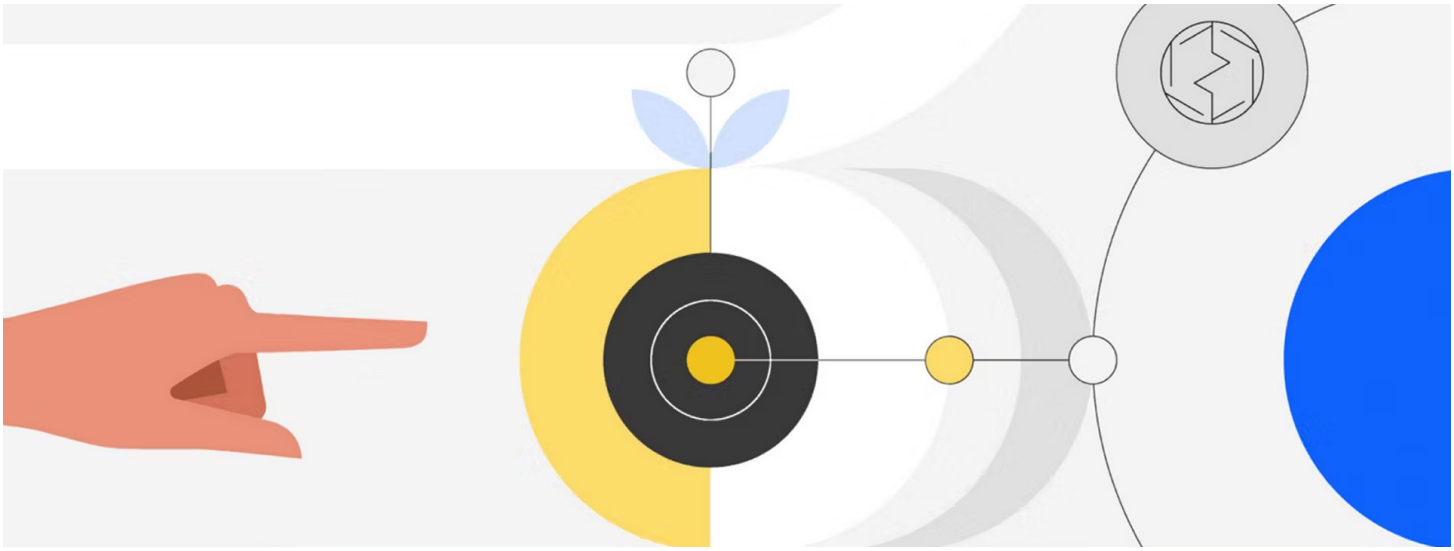
IBM Newsroom

IBM Taps NVIDIA AI Data Platform Technologies to Accelerate AI at Scale

Introducing New Storage Capabilities for Unstructured Data, Planned Integrations with watsonx and IBM Consulting Capabilities for Agentic Reasoning and other AI Workloads



ARMONK, N.Y., March 18, 2025 /PRNewswire/ -- IBM (NYSE:IBM) today announced new collaborations with NVIDIA (NASDAQ: NVDA), including planned new integrations based on the NVIDIA AI Data Platform reference design to help enterprises more effectively put their data to work to help build, scale and manage generative AI workloads and agentic AI applications. As part of today's news, IBM is planning to launch a content-aware storage capability for its hybrid cloud infrastructure offering, IBM Fusion; intends to expand its watsonx integrations; and is introducing new IBM Consulting capabilities with NVIDIA to help drive AI innovation across the enterprise.

A 2024 IBM report found that more than three in four executives surveyed (77 percent) say generative AI is market-ready, up from just 36 percent in 2023. With this push to put AI into production comes an increased need for compute and data-intensive technologies. The collaboration between IBM and NVIDIA will enable IBM to provide hybrid AI solutions that take advantage of open technologies and platforms while also supporting data management, performance, security, and governance.

Leveraging the NVIDIA AI Data Platform reference architecture, these new solutions are the latest in the IBM and NVIDIA collaboration to build enterprise infrastructure for AI:

- **Augmenting Unstructured Data Processing for AI Performance**: With IBM's new content-aware storage (CAS) capability, enterprises will be able to extract the meaning hidden in their rapidly growing volumes of unstructured data for inferencing, without compromising trust and safety, to responsibly scale and enhance AI applications like retrieval-augmented generation (RAG) and AI reasoning. IBM Storage Scale will respond to queries using the extracted and augmented data, speeding up the communications between GPUs and storage using NVIDIA BlueField-3 DPUs and NVIDIA Spectrum-X networking. The multimodal document data extraction workflow will also leverage NVIDIA NeMo Retriever microservices, built with NVIDIA NIM. CAS will be embedded in the next update of IBM Fusion planned for the second quarter of this year.
- **Enabling More Accessible AI**: IBM plans to integrate its watsonx offerings with NVIDIA NIM as part of a larger effort to provide access to leading AI models across multiple cloud environments. This will allow organizations to leverage

watsonx.ai, IBM's enterprise-grade AI platform and developer studio, to develop and deploy AI models into their applications of choice while utilizing externally hosted models. IBM's watsonx.governance also allows enterprises to implement robust monitoring and governance of NVIDIA NIM microservices across any hosting environment. This type of interoperability is increasingly essential as organizations adopt agentic AI and other advanced applications that require AI model integration.

- **Increasing Support for Compute-Intensive Workloads**: With more enterprises embracing generative AI and high-performance computing (HPC), IBM Cloud expanded its NVIDIA accelerated computing portfolio by announcing the availability of NVIDIA H200 instances on IBM Cloud. With its large memory capacity and high bandwidth, the NVIDIA H200 Tensor Core GPU instances are engineered to meet the demands of modern AI workloads and larger foundation models.
- **Transforming Processes with Agentic AI and NVIDIA:** IBM Consulting is introducing AI Integration Services to help clients transform and govern end-to-end business processes with agentic AI using NVIDIA Blueprints, such as industry-specific workflows that require agentic AI at the edge. Example use cases include autonomous inspection and maintenance in the manufacturing industry or proactive video data analysis and anomaly response in the energy industry.
- **Optimizing Compute Intensive AI Workloads Across Hybrid Cloud Environments:** IBM Consulting helps clients build, modernize and manage compute-intensive AI workloads across hybrid cloud environments leveraging RedHat OpenShift and NVIDIA AI. This includes technologies like NVIDIA AI Foundry, NVIDIA NeMo, NVIDIA AI Enterprise, NVIDIA Blueprints, and NVIDIA Clara to accelerate high-compute, complex tasks, while managing AI governance, data security and compliance requirements.

"IBM is focused on helping enterprises build and deploy effective AI models and scale with speed," said Hillery Hunter, CTO and General Manager of Innovation, IBM Infrastructure. "Together, IBM and NVIDIA are collaborating to create and offer the solutions, services and technology to unlock, accelerate, and protect data – ultimately helping clients overcome AI's hidden costs and technical hurdles to monetize AI and drive real business outcomes."

"AI agents need to rapidly access, fetch and process data at scale, and today, these steps occur in separate silos," said Rob Davis, vice president, Storage Networking Technology, NVIDIA. "The integration of IBM's content-aware storage with NVIDIA AI orchestrates data and compute across an optimized network fabric to overcome silos with an intelligent, scalable system that drives near real-time inference for responsive AI reasoning."

To learn more about IBM's presence at GTC, please visit https://www.nvidia.com/gtc/session-catalog/?search.suggestedaudiencelevel=1732117107498003nOoA&search=ibm#/

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

About IBM

IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs and gain a competitive edge in their industries. Thousands of government and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and consulting deliver open and flexible options to our clients. All of this is backed by IBM's long-standing commitment to trust, transparency, responsibility, inclusivity and service. Visit www.ibm.com for more information.

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

other countries.

Media contact:

Alexandra Demetriades
IBM
Alexandra.Demetriades@ibm.com

SOURCE IBM

Additional assets available online: Photos

https://stage.mediaroom.com/ibmnewsroom/2025-03-18-ibm-taps-nvidia-ai-data-platform-technologies-to-accelerate-ai-at-scale