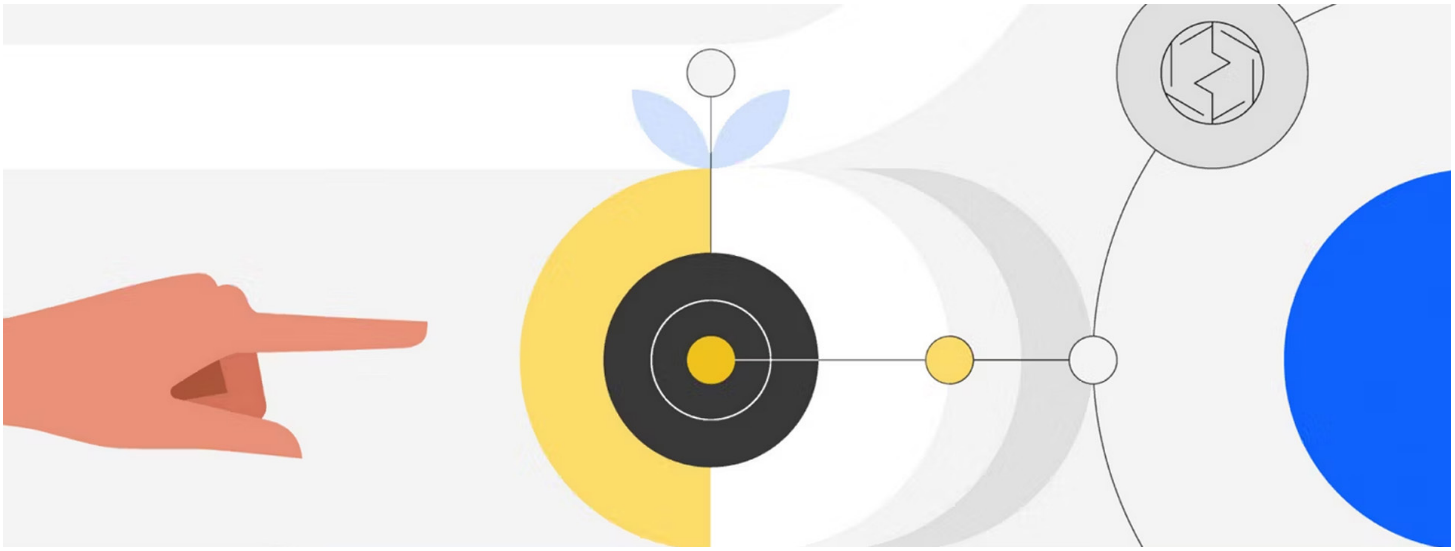
IBM Newsroom

IBM Announces Expanded Collaboration with NVIDIA to Advance AI for the Enterprise

Advancements across GPU-native data analytics, unstructured data extraction, on-premises and cloud infrastructure, Nestlé global supply chain decision speed, and consulting to mobilize enterprise AI at scale



ARMONK, N.Y., March 16, 2026 – IBM (NYSE: IBM) today announced at GTC 2026 an expanded collaboration with NVIDIA to help enterprises operationalize AI at scale. Advancing efforts across GPU-native data analytics, intelligent document processing, on-premises and regulated infrastructure deployments, cloud, and consulting, the collaboration aims to give enterprises the data foundation, infrastructure, and expertise to move AI from pilot to production.

Enterprises are making significant investments in AI, but too many remain stuck between experimentation and production at scale. The barriers are consistent: data is fragmented and difficult to access; infrastructure wasn't built for advanced AI workloads; AI deployments don't support the compliance and residency requirements of regulated industries; and many organizations still need the guided expertise to implement and deploy the technologies. Today's announcements from IBM and NVIDIA are designed to close these gaps.

"In the next wave of enterprise AI, the model layer will rely on the data, infrastructure, and orchestration layers – and on businesses that can bring all three together," said **Arvind Krishna, Chairman and CEO, IBM**. "Our partnership with NVIDIA goes to the heart of that challenge. Together, we're giving enterprises the solutions they need to stop experimenting with AI and start running on it."

"IBM pioneered enterprise computing and data processing six decades ago — and today they are redefining it for the AI era," said **Jensen Huang, founder and CEO of NVIDIA.** "Data is the ground truth that gives AI context and meaning. Together with IBM, we are bringing CUDA GPU acceleration directly into the data layer — turning analytics and document processing from bottlenecks into real-time intelligence engines."

Accelerating Structured Data Analytics with GPU-Native Computing

IBM and NVIDIA are collaborating on an open-source integration to increase performance and reduce costs around how enterprises extract intelligence from their massive datasets. IBM watsonx.data's SQL engine Presto is accelerated by NVIDIA cuDF to enable faster query execution on large datasets.

To validate in production, IBM and NVIDIA applied GPU-accelerated watsonx.data to Nestlé’s Order-to-Cash data mart. The data mart tracks every order, fulfillment, delivery, and invoice across 186 countries and processes terabytes across 44 tables. Nestlé was ideal for this proof of concept because of its strong digital backbone. With globally unified data models, a consolidated data foundation, and a single source of truth across markets, Nestlé already had timely, accurate, and trusted data at scale — the right foundation to put GPU-accelerated analytics to the test in a real production environment.

On CPUs, a single refresh previously took Nestlé 15 minutes and only ran a handful of times a day. Nestlé reports that with NVIDIA’s software and GPUs, the IBM watsonx.data Presto engine reduced query runtime down to three minutes – achieving 83% cost savings and an overall 30X price-performance improvement.

“For a company that serves billions, data underpins decision making across our global operations,”**said Chris Wright, Chief Information and Digital Officer of Nestlé**. “Working with IBM and NVIDIA, a targeted proof of concept has demonstrated the ability to refresh global operations data in a few minutes and at reduced cost. Our focus now is on turning this capability into tangible business impact — further improving decision speed in areas such as manufacturing and warehousing, and scaling these capabilities across our enterprise.”

Helping Enterprises Unlock the Full Value of Their Data

Most enterprises aren't lacking data. But often, they're unable to access and use it. SharePoint sites, CMS systems, vendor research, SME knowledge: the information exists but it is trapped in unstructured, multi-modal formats that are difficult to extract, standardize, and trust at decision speed.

IBM and NVIDIA are addressing this with Docling from IBM and NVIDIA Nemotron open models – a combination designed to make intelligent document extraction available at enterprise scale. Docling standardizes and converts documents into AI-ready formats with source-level traceability, while NVIDIA Nemotron models accelerate ingestion of multi-modal content. Early results show significantly higher throughput compared to other open-source models, while maintaining or improving accuracy wherever GPU-accelerated infrastructure is available.

GPU-Optimized Infrastructure for On-Prem and Regulated Deployments

IBM and NVIDIA are extending their data efforts to the infrastructure layer. NVIDIA has selected IBM Storage Scale System 6000 to provide 10PB of high-performance storage to serve massive data for its GPU-native advanced analytics engines, pairing IBM’s unified data access layer and massive parallel throughput with NVIDIA’s GPU pipelines. IBM Storage Scale 6000 is certified and validated on NVIDIA DGX platforms.[1]

For enterprises and governments requiring data residency and regulatory control, IBM and NVIDIA are exploring the integration of IBM Sovereign Core and NVIDIA infrastructure and NVIDIA Nemotron models that would focus on enabling GPU-intensive AI workloads that run entirely within regional boundaries – without compromising governance or compliance.

Advancing the Enterprise AI Stack with IBM, NVIDIA and Red Hat

IBM and NVIDIA are also deepening their partnership across cloud and enterprise consulting to advance clients’ enterprise AI adoption. IBM plans to offer NVIDIA Blackwell Ultra GPUs on IBM Cloud in early Q2 2026 for large-scale training, highthroughput inferencing, and AI reasoning. This technology will also be integrated across Red Hat AI Factory with NVIDIA, and VPC servers with enterprise-grade compliance and data residency controls.

Additionally, IBM Consulting plans to bring Red Hat AI Factory with NVIDIA to clients through IBM Consulting Advantage – an IBM enterprise AI platform that helps clients build and scale AI across their technology environments. Combined with Red Hat AI Factory with NVIDIA, the platform is built to simplify how companies prepare data, build models, and deploy AI, while also enhancing performance and oversight. This builds on IBM Consulting's broader efforts to help clients maximize outputs from their AI investments.

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 the competitive edge in their industries. Thousands of governments 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.

Media contacts:

Sarah Benchaita
Software Communications, IBM
sarah.benchaita@ibm.com

Bethany McCarthy
Infrastructure Communications, IBM
bethany@ibm.com

[1] IBM Storage Scale System 6000 is NVIDIA-Certified Storage