Dell Offers Complete NVIDIA-Powered AI Factory Solutions to Help Global Enterprises Accelerate AI Adoption

March 18, 2024

New Dell AI Factory with NVIDIA is the industry's first end-to-end AI solution for enterprises spanning workstations, data centers and cloud to supercharge era of generative AI

SAN JOSE, Calif., March 18, 2024 /PRNewswire/ -- **NVIDIA GTC 2024**

News summary

- Dell AI Factory with NVIDIA is industry's first comprehensive AI solution to help enterprises quickly capitalize on AI investments
- Expanded Dell Generative AI Solutions portfolio helps improve results accuracy and model training, simplifies data access and optimizes the efficiency of processing and storage
- Dell PowerEdge to support new NVIDIA GPU configurations and rack scale NVIDIA GB200 NVL72

Full story
Dell Technologies (NYSE: DELL) is strengthening its collaboration with NVIDIA to help enterprises adopt AI technologies. By expanding the Dell Generative AI Solutions portfolio, including with the new Dell AI Factory with NVIDIA, organizations can accelerate integration of their data, AI tools and on-premises infrastructure to maximize their generative AI (GenAI) investments.

"Our enterprise customers are looking for an easy way to implement AI solutions – that is exactly what Dell Technologies and NVIDIA are delivering," said Michael Dell, founder and CEO, Dell Technologies. "Through our combined efforts, organizations can seamlessly integrate data with their own use cases and streamline the development of customized GenAI models."

"AI factories are central to creating intelligence on an industrial scale," said Jensen Huang, founder and CEO, NVIDIA. "Together, NVIDIA and Dell are helping enterprises create AI factories to turn their proprietary data into powerful insights."

High quality results from high quality data
Through close collaboration between Dell and NVIDIA, additions to the end-to-end Dell Generative AI Solutions portfolio helps customers modernize with AI, accelerate business transformation and boost productivity:

- **Dell AI Factory with NVIDIA** is the industry's first end-to-end AI enterprise solution integrating Dell's compute, storage, client device, software and services capabilities with NVIDIA's advanced AI infrastructure and software suite, all underpinned by a high-speed networking fabric.[1] Delivered as a fully integrated solution, Dell AI Factory with NVIDIA takes advantage of rack-level design, with rigorous testing and validation to deliver a seamless solution for transforming data into valuable insights and outcomes. This solution also leverages existing offerings in enterprise data security with accompanying Dell services offerings in security and privacy.

  The Dell AI Factory with NVIDIA supports a wide array of AI use cases and applications to support the entire GenAI lifecycle, from model creation and tuning, to augmentation and inferencing. Customers can also take advantage of enterprise-grade professional services that help organizations accelerate their strategy, data preparation, implementation and adoption of the AI Factory, advancing AI capabilities. The Dell AI Factory with NVIDIA is available via traditional channels and Dell APEX.
- **Dell Technologies** will collaborate with NVIDIA to introduce a rack scale, high-density, liquid-cooled architecture based on the NVIDIA Grace Blackwell Superchip. These systems will support the next-generation ecosystem aiming to provide the foundation for improvements in performance density for enterprise AI workloads.
- **Dell PowerEdge XE9680 servers** will support new NVIDIA GPU models, including the NVIDIA B200 Tensor Core GPU, expected to offer up to 15 times higher AI inference performance and lower total cost of ownership.[2] Dell PowerEdge servers will also support other NVIDIA Blackwell architecture-based GPUs as well as H200 Tensor Core GPUs and NVIDIA Quantum-2 InfiniBand and Spectrum-X Ethernet networking platforms.
- **Dell Generative AI Solutions with NVIDIA – Retrieval-Augmented Generation (RAG)** leverages new microservices in NVIDIA AI Enterprise to offer a pre-validated, full-stack solution to speed enterprise AI adoption with RAG. This solution

helps organizations improve GenAI model quality and increase results accuracy with proprietary business data and knowledge bases.
- **Dell Generative AI Solutions with NVIDIA – Model Training** offers a pre-validated, full-stack solution for organizations seeking to build their own custom, domain-specific models.
- **Dell Data Lakehouse**, now globally available, is an open, modern data lakehouse that helps organizations discover, process and analyze data in one place across hybrid and multicloud environments.
- **Dell PowerScale** is the world's first Ethernet storage solution validated with NVIDIA DGX SuperPOD with DGX H100 systems, helping customers achieve faster and more efficient AI storage.[3]

Across it all, Dell Professional Services for GenAI expands with support from NVIDIA AI and infrastructure experts to help customers integrate, manage and secure these solutions to achieve business results faster. **Dell Implementation Services** now include capabilities to deliver Dell's new RAG solution, model training and the Dell Data Lakehouse, as well as new Advisory Services for GenAI Data Security that help customers assess and minimize security risks.

"Organizations are rushing to experiment with AI but there are many challenges to achieving ROI. Data sovereignty issues, legal and compliance concerns and data quality are all top of mind. New offerings, such as Dell's Generative AI Solutions with NVIDIA – RAG, provide enterprises a simpler on-ramp to GenAI, helping to increase confidence and develop their own trusted GenAI capabilities that can deliver substantial business impact," said Dave Vellante, Chief Analyst, theCUBE Research. "Our research shows that companies are turning to industry leaders like Dell and NVIDIA to help provide AI expertise and services to lower risk and get to ROI sooner."

Availability

- Dell AI Factory with NVIDIA is available globally through traditional channels and Dell APEX now.
- Dell PowerEdge XE9680 servers with NVIDIA B200 Tensor Core GPUs, NVIDIA B100 Tensor Core GPUs and NVIDIA H200 Tensor Core GPUs have expected availability later this year.
- Dell Generative AI Solutions with NVIDIA – RAG is available globally through traditional channels and Dell APEX now.
- Dell Generative AI Solutions with NVIDIA – Model Training will be available globally through traditional channels and Dell APEX in April 2024.
- As previously announced, the Dell Data Lakehouse is now available globally.
- Dell PowerScale is validated with NVIDIA DGX SuperPOD with DGX H100 and NVIDIA OVX solutions now.
- The Dell Implementation Service for RAG is available in select locations starting May 31.
- Dell infrastructure deployment services for model training is available in select locations starting March 29.
- Advisory Services for GenAI Data Security is available in select countries starting March 29.

Additional Resources

- Learn more about the Dell AI Factory with NVIDIA.
- Read more about Dell PowerEdge XE9680 servers' support for new NVIDIA GPU configurations.
- Find out more details on Dell PowerScale.
- Explore more details on the Dell Data Lakehouse.
- Discover more about AI at Dell Technologies.
- Connect with Dell on X and LinkedIn

About Dell Technologies
Dell Technologies (NYSE: DELL) helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the data era.

[1] Based on Dell analysis, March 2024. Dell offers solutions with NVIDIA infrastructure and software engineered to support AI workloads from Workstations PCs to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services.

[2] Provided by NVIDIA: Projected performance subject to change. Token-to-token latency (TTL) = 50ms real time, first token latency (FTL) = 5s, input sequence length = 32,768, output sequence length = 1,028, 8x eight-way HGX H100 air-cooled vs. 1x eight-way HGX B200 air-cooled, per GPU performance comparison.

[3] Based on Dell internal analysis, March. 2024.

View original content to download multimedia:https://www.prnewswire.com/news-releases/dell-offers-complete-nvidia-powered-ai-factory-solutions-to-help-global-enterprises-accelerate-ai-adoption-302091700.html

SOURCE Dell Technologies

Media.Relations@Dell.com