RESEARCH NOTE
PROGRAM: INFRASTRUCTURE
DOCUMENT NUMBER: Y155 OCTOBER 2024

NUCLEUS
RESEARCH

TERADATA MAKES TRUSTED AI POSSIBLE

ANALYST
Alexander H. Wurm

THE BOTTOM LINE

At Teradata Possible 2024 in Los Angeles, Teradata unveiled key advancements for its VantageCloud Lake platform, introducing Bring Your Own Large Language Model (BYO-LLM) and GPU-accelerated compute features through its partnership with NVIDIA. Collectively, these capabilities better enable enterprises to take advantage of emerging AI and machine learning use cases while balancing cost-effectiveness, resource consumption, and data security. Customers can expect accelerated performance across a range of complex models as well as improved support for inference, RAG, and fine-tuning tasks.

OVERVIEW

Enterprise customers are increasingly recognizing the critical importance of AI capabilities integrated directly into their database and analytics platforms. These AI-enhanced solutions offer the ability to process vast amounts of data more efficiently, generate deeper insights, and enable real-time decision-making. As the AI landscape continues to evolve, enterprises that leverage advanced AI capabilities within their data management systems are better positioned to innovate, adapt to market changes, and maintain a competitive edge.

At Teradata Possible 2024, the company unveiled two major enhancements to its VantageCloud Lake platform: ClearScape Analytics BYO-LLM and NVIDIA AI accelerated compute integration. These new features are designed to allow organizations to harness the generative AI and large language models within their existing data ecosystems. By combining Teradata's data management capabilities with AI technologies, these enhancements enable businesses to unlock new levels of insights and operational efficiency across various industries and use cases.

AI CAPABILITIES HIGHLIGHTED

Teradata introduced two AI-focused capabilities for its VantageCloud platform, leveraging proprietary and third-party technologies to deliver AI solutions better equipped to drive enterprise ROI.

ACCELERATED COMPUTE WITH NVIDIA

Teradata’s partnership with NVIDIA harnesses NVIDIA GPUs to improve AI and ML workload performance. This integration leverages NVIDIA's CUDA architecture and optimized libraries such as cuDNN and TensorRT to accelerate machine learning and deep learning operations. The platform utilizes NVIDIA's multi-instance GPU (MIG) technology to enable efficient resource allocation and isolation for concurrent AI workloads. With the addition of inference capabilities in Q4 2024, the system will support real-time AI model deployment, leveraging NVIDIA Triton Inference Server for high-throughput, low-latency inference across various AI frameworks. The integration of NVIDIA NIM microservices enhances support for retrieval-augmented generation (RAG) applications, allowing organizations to connect custom models with business data for more accurate and contextually relevant AI-generated responses. Furthermore, the planned fine-tuning capabilities in 1H 2025 will further extend the platform's versatility, allowing for on-the-fly model adaptation using Teradata's data resources and NVIDIA's optimized training libraries. Benefits delivered include:

- **High-performance AI processing.** By integrating NVIDIA's accelerated computing technologies, Teradata enables faster and more cost-effective inferencing for highly complex models such as machine learning, deep learning and large language models.
- **Improved RAG support.** Teradata enhanced its retrieval-augmented generation (RAG) capabilities through its partnership with NVIDIA, specifically by integrating NVIDIA NeMo Retriever, a collection of NVIDIA NIM microservices, into the Teradata VantageCloud platform. This integration allows organizations to connect custom language models to business data sources, enabling highly accurate and context-aware responses in AI applications. The NVIDIA NIM microservices, part of the NVIDIA AI Enterprise software suite, will be initially available to Teradata VantageCloud customers, with plans to expand to hybrid environments globally.
- **Fine-tuning support.** This enables customers to leverage GPU-accelerated computing clusters for advanced model customization tasks, particularly for large language models (LLMs). Planned for release in the first half of 2025, this capability will allow organizations to customize and optimize AI models for their specific use cases and data environments.

BYO-LLM

BYO-LLM for Teradata VantageCloud enables organizations to integrate custom large language models (LLMs) directly into their data analytics environment. This feature leverages Teradata's in-database processing capabilities to execute LLM inference within the secure confines of the VantageCloud Lake platform. By allowing customers to bring their own models, Teradata facilitates the use of proprietary or domain-specific LLMs while maintaining data locality and security. The BYO-LLM functionality supports various model formats and can be seamlessly integrated with Teradata's existing Open and Connected ecosystem for AI, enabling complex natural language processing tasks to be performed directly on massive datasets without data movement. Example use cases include sentiment analysis, and regulatory compliance monitoring where LLMs can be used to integrate insights with operational data for actionable results. Benefits delivered include:

- **Flexible model integration.** This feature allows organizations to bring their own large language models into the Teradata environment, providing flexibility in choosing and deploying AI models that best fit their specific needs and use cases.
- **Enhanced data security**. By enabling LLM integration within the Teradata ecosystem, organizations can maintain stricter control over their sensitive data, ensuring compliance with data privacy regulations and internal security policies.

- **Streamlined AI workflows.** The BYO-LLM capability simplifies the process of incorporating AI models into existing data analytics workflows, reducing the complexity and time required to operationalize AI. The platform’s integrated data management and analytics capabilities allow enterprises to solve complex AI use cases with a unified, in-platform experience, eliminating the need for data movement and further enhancing the efficiency of AI deployment.
- **Cost-effective approach.** Leveraging both GPU and CPU parallel inferencing offers a sustainable way to execute different use cases based on the complexity of the LLMs.

WHY IT MATTERS

Teradata's introduction of BYO-LLM capabilities and GPU-accelerated compute in VantageCloud Lake and ClearScape Analytics marks a key advancement for emerging enterprise AI applications enabling organizations to better integrate inference, RAG, and fine-tuning into their existing data workflows. The integration of NVIDIA's full-stack AI platform accelerates trusted AI workloads, offering improved performance for various AI/ML tasks. Furthermore, by allowing customers to leverage small or mid-sized open LLMs, including domain-specific models, Teradata facilitates cost-effective deployment of everyday GenAI use cases that fit with customer goals. This approach allows businesses to tailor AI solutions to their needs without the substantial resource requirements of larger, more general-purpose models. By focusing on smaller, specialized models, organizations can achieve faster implementation times and more targeted results, aligning AI capabilities closely with their unique business objectives and industry-specific challenges.

With these announcements, Teradata better serves organizations seeking to harness generative AI while addressing concerns about data privacy, model customization, and performance. Additionally, the flexibility to leverage either GPUs or CPUs, depending on the complexity and size of the generative AI workload, allows organizations to optimize their use cases for both performance and cost-effectiveness. Collectively, these announcements position Teradata as a key player delivering AI solutions with a trusted a flexible foundation, enabling enterprises to recognize its potential business value while maintaining control over their data.