October 8, 2024

NEWS ANALYSIS

Teradata Offers New Options for Innovating With Generative AI

Teradata's Bring-Your-Own-LLM Choice and GPU Acceleration Give Customers Flexibility

Doug Henschen
VICE PRESIDENT AND PRINCIPAL ANALYST

constellation RESEARCH

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Once generative artificial intelligence (GenAI) grabbed the world's attention, in 2023, a technology arms race ensued, with the leading purveyors of large language models (LLMs) outdueling each other to deliver ever larger and more capable models. Innovative businesses jumped on the GenAI bandwagon, experimenting with copilots and plugging into foundation models on the leading public clouds to build custom GenAI capabilities. What many businesses soon learned is that GenAI development costs can be steep and the road to successful production is fraught with pitfalls, roadblocks, and security risks.

As 2024 unfolded, model options grew more numerous, and innovators soon learned that open source models; midsize models; and even small, purpose-built models could often outperform LLMs at a much lower cost. What's more, businesses learned that high-cost graphical processing unit (GPU) capacity isn't always necessary but that in the most demanding circumstances, it can be more economical than conventional central processing unit (CPU) capacity, because it whips through GenAI tasks much more quickly.

Building on these learnings, Teradata, in early October 2024, announced two key offerings: ClearScape Analytics Bring Your Own LLM (BYO-LLM) in Teradata VantageCloud Lake and Teradata VantageCloud Lake GPU-Accelerated Compute. This report details these two offerings and related GenAI-based Teradata solutions supporting three use cases: understanding customer complaints, helping banks ensure regulatory compliance, and helping healthcare providers analyze doctors' notes. Early adopters should use this report to gain a better understanding of Teradata's emerging offerings that support GenAI innovation.

BUSINESS THEMES

- Data to Decisions
- Next-Generation Customer Experience
- Technology Optimization

TERADATA LAUNCHES BRING-YOUR-OWN-LLM AND GPU ACCELERATION OPTIONS

TERADATA

Company: Teradata
Headquarters: San Diego, CA
Founded: 1979
Type: Public company—TDC on the New York Stock Exchange
Company size: $1.8 billion in revenue in fiscal year 2023
No. of Employees: 7,000+
Website: *Teradata.com*

Back in the early 2000s, when enterprise data was quickly becoming "big data," the best practice for generating predictive analytics shifted to bringing the model to the data rather than moving the data to the model. It only made sense, because the cost and complexity of moving terabytes of data to a separate system surpassed that of the easier, more cost-effective option of bringing the model to the data.

Flash-forward to the present day, and the same best practice of bringing the model to the data still applies for any and all forms of data science, including AI. It also usually makes sense to bring the compute to the data, rather than moving the data to a separate, high-performance compute environment. Thus, most modern cloud data platform providers have added GPU acceleration to their compute options. GPUs handle demanding workloads that might take more time and, in the long run, cost more with CPUs. Thus, customers want both options.

Teradata Bring-Your-Own-LLM Option Gives Customers Choices

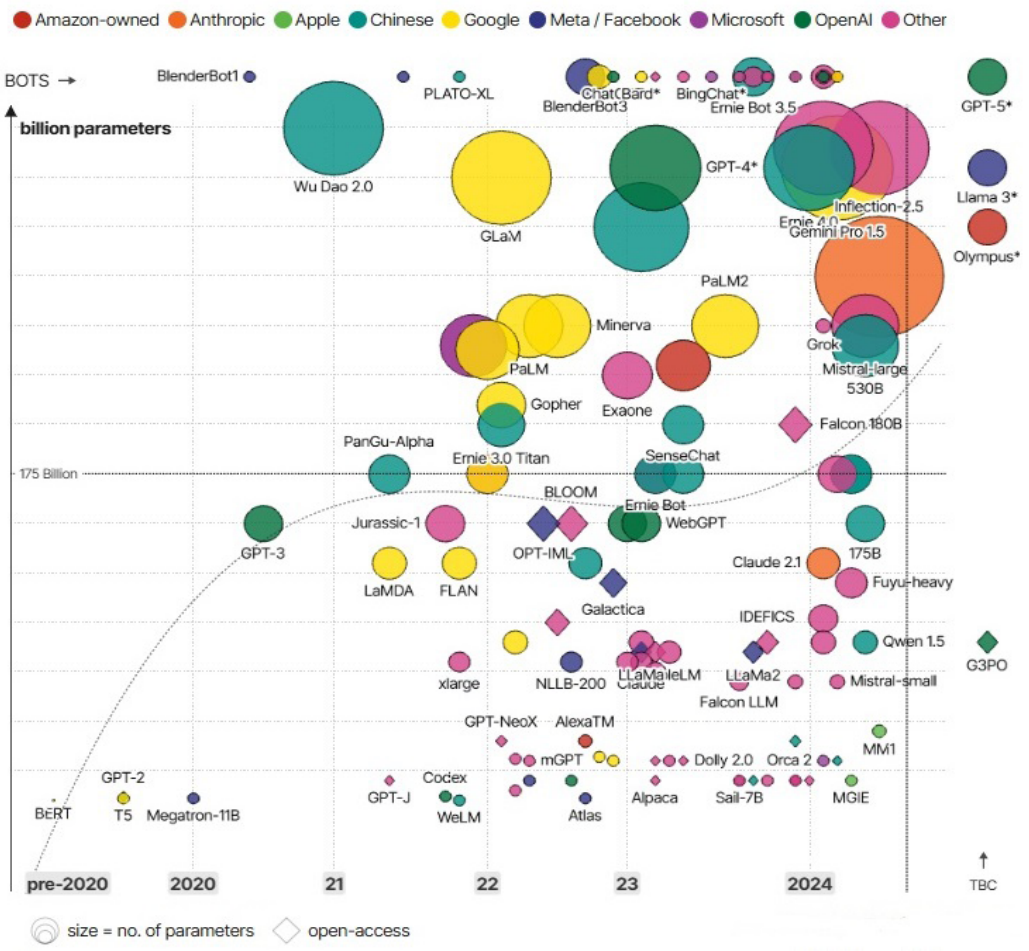
As GenAI took off in 2023, most cloud service providers and some data platform vendors started developing model libraries (aka model gardens). Some data platform companies even launched their own "native" models to jump on the GenAI bandwagon.

What has become apparent in 2024, as shown in Figure 1, is that models are proliferating almost to the point of becoming commoditized. Small models, midsize models, and open source models (aka open access models) are gaining adoption alongside the also-growing list of commercial LLMs. The investments required for model development are massive, so the competition is being led by hyperscale cloud service providers, such as Amazon Web Services (AWS), Google, and Microsoft; social networking giants, such as Meta; and dedicated AI model builders, such as OpenAI and Anthropic.

In the face of this competition, the appeal of any one-off model library is only as strong as the diversity and vibrancy of the models available. And the appeal of any one model is only as strong as its performance, which is closely tied to the investment put into it, the purpose-built focus of the model, and the community that grows up around that model.

Recognizing these realities, Teradata, in early October 2024, announced ClearScape Analytics BYO-LLM. As the name suggests, the offering gives customers the flexibility to bring models of their choice to the Teradata VantageCloud Lake platform by way of ClearScape Analytics, the vendor's proven engine for supporting end-to-end AI/machine learning (ML) pipelines and in-database analytics. ClearScape Analytics capabilities include data prep, model training, and model operationalization at scale.

Figure 1. The "Rise and Rise of AI" Visualization of Emerging Large Language Models and Associated Chatbots



Source: LifeArchitect.AI (David McCandless, Tom Evans, and Paul Barton)

VantageCloud Lake is the vendor's multicloud lakehouse platform, which harnesses object storage and is designed to run independently and elastically scalable workloads spanning structured, unstructured, and semistructured data. The lakehouse architecture fits, because innovative GenAI use cases invariably are focused on semistructured data, such as text. ClearScape Analytics BYO-LLM is set to become generally available on AWS in November and is expected to be available on Google Cloud and Microsoft Azure in the first half of 2025.

The model source for ClearScape Analytics BYO-LLM is Hugging Face, which is used by more than 50,000 organizations and is highly regarded for its ease of use and community support. Hugging Face offers a selection of more than 350,000 models across audio, computer vision, multimodal, and reinforcement learning use cases, but it's best known for natural-language processing (NLP) models, making it a top choice for AI practitioners.

BYO-LLM isn't the only option for Teradata customers. As shown in Figure 2, Teradata supports three AI/GenAI design patterns. ClearScape Analytics has long enabled customers to bring their own models into the platform for in-database management system (DBMS) processing. The new ClearScape BYO-LLM offering is billed as supporting "performant execution" of medium-complexity commercial and open source models.

Many small and midsize LLMs available on Hugging Face are domain-specific, which, in many cases, makes them more performant for specific tasks than large, general-purpose models. Smaller models are also cheaper to run and easier to deploy, leveraging parallel CPU inferencing within VantageCloud Lake rather than higher-cost GPU infrastructure.

With the in-DBMS and in-platform BYO-LLM approaches, there's no need for special-purpose infrastructure or for data to move beyond corporate firewalls. The model is applied where the customer's data lives, in Teradata VantageCloud, so organizations avoid data movement and ensure data security and governed access control.

Figure 2. Teradata Enables Customers To Apply Models to Data Analysis Through Three Design Patterns: In-Database (DBMS), In-Platform, and via APIs

Source: Teradata

For models that are truly large, the third option is to use model endpoints that cloud service providers make available for foundation models such as Azure OpenAI and Google Gemini. This option, which is currently in preview, enables customers to use contextual data in VantageCloud Lake to improve and validate responses in the latest foundation models.

Constellation's analysis: Teradata isn't the first mover among cloud data platform providers to offer a way to bring LLMs to the data. But Constellation applauds Teradata for not building a tiny library through quid pro quo partnerships with companies willing to pay or be paid to be part of the library. Nor is it adding to that library what are best described as vanity models, designed to cause a splash in time for the big conference but that see little long-term adoption. Teradata, by contrast, is relying on one of the top sources of models chosen by developers: Hugging Face. As discussed in the next section, it's also helping customers with model selection and solutions that promise a short time to value.

Three Solutions Accelerate Generative AI Use Cases

Teradata is doing more to help customers than making it easier for them to bring models from Hugging Face onto the Teradata VantageCloud Lake platform. For starters, Teradata has come up with a curated short-list of best-fit small and midsize open source models (see Figure 3) that are known to run efficiently on VantageCloud Lake. Most importantly, Teradata is developing GenAI use case solutions that detail the steps users need to go through to apply models, perform multistep

Figure 3. Teradata Has Curated a Shortlist of Small and Midsize Open Source Models Available on Hugging Face and Suited to Specific Use Cases

Capability	Enterprise Use-case	Model Name	License
Analyze sentiment - Detect sentiment of text across your table.	Customer analytics	distilbert-base-uncased-emotion	Apache 2.0
Summarize - Summarize long documents for faster consumption	Customer analytics, product recommendation	bart-large-cnn	MIT
Language detection - Identify the language of given text and labels the language code	Chatbots	xlm-roberta-base-language-detection	MIT
Classify text with labels provided	Customer analytics	Facebook/bart-large-mnli	MIT
Entity-recognition	Customer analytics	tner/roberta-large-ontonotes5	MIT
Masking PII entities	Trusted AI	ab-ai/pii_model	Apache 2.0
Grammar Correction	Product management	pszemraj/flan-t5-large-grammar-synthesis	Apache 2.0
Sentence Similarity	Product recommendation	tner/roberta-large-ontonotes5	MIT
Extract information or phrases from your unstructured data.	Medical document analysis	ml6team/keyphrase-extraction-kbir-kpcrowd	MIT
Sentiment Analysis- (positive, negative, neutral)	Customer analytics	twitter-roberta-base-sentiment-latest	MIT
Embeddings	Product recommendation, document search	sentence-transformers/all-mpnet-base-v2	Apache 2.0

Source: Teradata

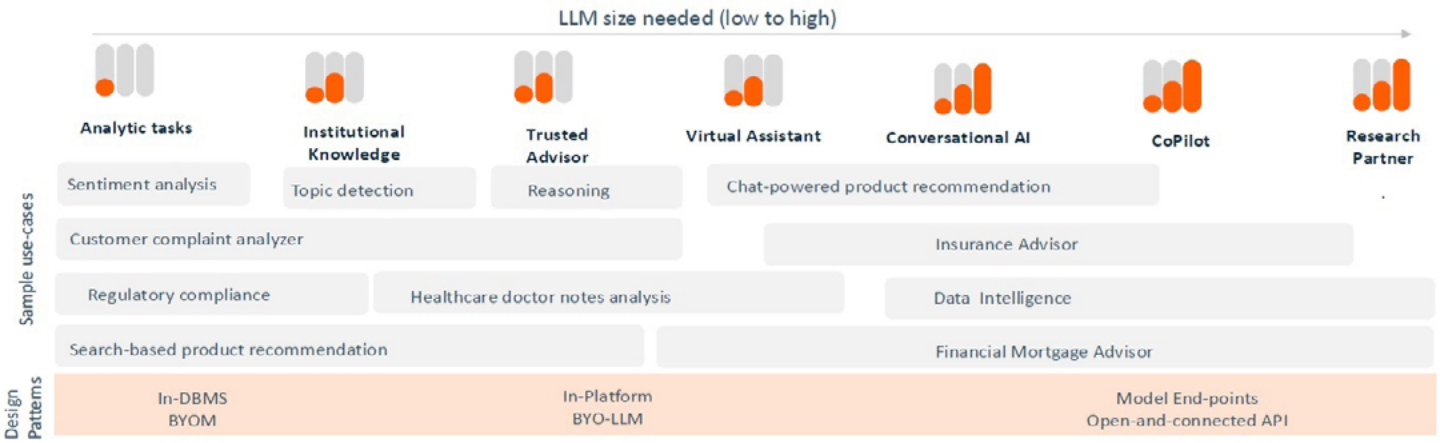
analyses, and deliver insights. The solutions also include templated user interfaces for analyzing and sharing the results.

As a starting point, Teradata has introduced three use-case-specific solutions:

- **Customer Complaint Analyzer.** Customers can use small open LLMs from Hugging Face to identify complaint topics and then measure sentiment via clustering and summarization techniques supported in ClearScape Analytics.
- **Regulatory Compliance.** Banks can use small open LLMs to locate email messages that might have regulatory implications.
- **Healthcare Doctor Notes Analysis.** Healthcare providers can automate the extraction and interpretation of information in doctors' notes to ensure consistency with patient diagnoses and records. The application saves administrative time while handling confidential information within the confines of Teradata's secure platform, with appropriate access controls.

Other low-hanging-fruit use cases for GenAI include chat-based product recommendation capabilities for retailers and AI-based advisors for insurance agents and financial advisors (see Figure 4). As also

Figure 4. Teradata Believes That Customers Will Need To Address a Variety of Use Cases Requiring a Variety of Model Types and Sizes



Source: Teradata

shown in Figure 4, certain use cases might call on combinations of the in-DBMS, in-platform. and model endpoint design patterns detailed in Figure 2.

Constellation's analysis. It's early days for Teradata's ClearScape Analytics BYO-LLM offering, and work is under way on more use case solutions applying GenAI to semistructured and unstructured data within the secure, governed confines of Teradata VantageCloud Lake. As required, insights from unstructured and semistructured sources can be correlated with structured data insights developed in-database, through ClearScape Analytics.

Constellation expects to see more vendor- and community-developed content made available for these solutions, likely to include best-practice recommendations, sample code, and more. And as new small, open, and domain-specific models become available on Hugging Face, customers will have new opportunities to deliver more performance and value without spending more money on the largest models and GPU processing.

GPU Acceleration Comes to VantageCloud Lake

Depending on the size of the model required, solutions developed by customers and/or by Teradata might require GPU acceleration. When organizations need to develop broadly capable conversational agents, copilots, or assistants, for example, there will likely be a need to use larger models and GPU

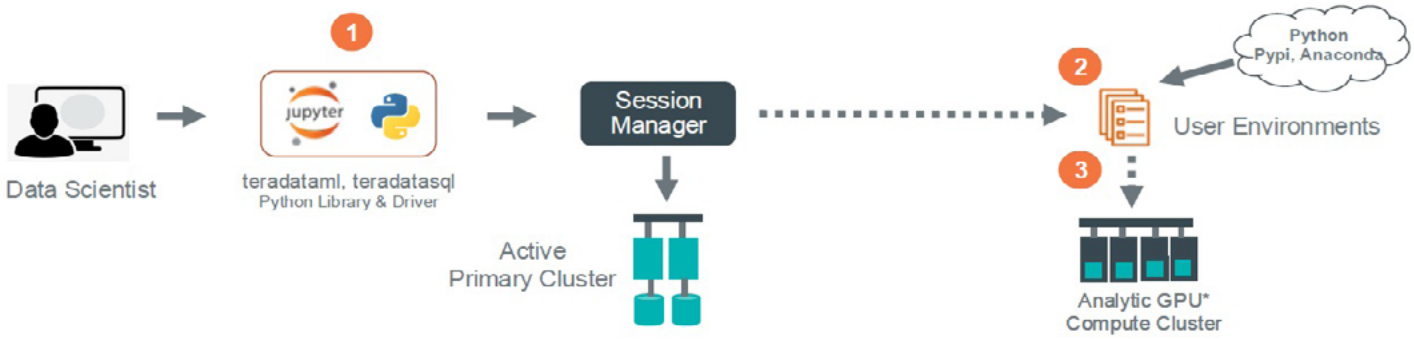
acceleration. To meet this need, the vendor has added Teradata VantageCloud Lake GPU-Accelerated Compute, with general availability on AWS scheduled for November 2024.

With GPU acceleration, customers get the performance and speed they need in order to analyze LLM results quickly and at scale. The GPU option ensures that as AI needs grow, the infrastructure can keep pace. In these use cases, conventional CPUs would likely slow to a crawl and end up costing more than applying GPU capacity for a shorter period.

GPU capacity can be accessed almost instantaneously via the following steps (see Figure 5):

1. The data scientist uses the Python language and packages to connect to VantageCloud Lake via the teradataml client library.
2. Using teradataml application-programming interfaces (APIs), the data scientist creates code artifacts and uploads the model to the user environment.
3. The data scientist executes the APPLY query to perform model inferencing on analytic GPU compute clusters preconfigured via the VantageCloud Lake console. The compute clusters read the Python script and AI/ML/LLM model stored in the user environment. The inference script is executed in parallel in each container in the compute cluster nodes.

Figure 5. To Tap GPU Acceleration, Customers Connect to VantageCloud Lake, Upload the Desired Model, and Inference on Preconfigured GPU Clusters



Source: Teradata

Constellation's analysis: GPU capabilities are essential for certain GenAI workloads, so the GPU Acceleration option was a must-have capability that Teradata had to deliver. Making the option generally available on AWS was a good first step, but Constellation is eager to learn more about planned availability on Azure and Google Cloud.

RECOMMENDATIONS

Organizations are in the early days of applying GenAI models to semistructured and unstructured data managed in cloud lakehouse platforms. Anticipating customer needs, Teradata has given customers flexible options via ClearScape Analytics BYO-LLM and VantageCloud Lake GPU-Accelerated Compute. Constellation sees these options as delivering the following benefits:

- **Choice.** By choosing Hugging Face, a platform that is already hugely popular with AI developers, Teradata is giving customers the choice of more than 350,000 models, including the vendor's trove of NLP models. The next step for Teradata will be adding integrations with Hugging Face so that customers can invoke their choice models programmatically via APIs rather than having to load models manually.

- **Flexibility.** With support for three design patterns, customers have a choice of using AI and ML models built into ClearScape Analytics for in-database processing, bringing small and midsize GenAI models into VantageCloud Lake via the BYO-LLM option, or accessing truly large foundation models via model endpoints. The option of adding GPU acceleration enables customers to match the compute power with the need.

- **Security.** The "in-platform" approach, much like the in-DBMS approach that Teradata helped pioneer more than two decades ago, gives customers the assurance of applying the model to the data, thereby avoiding the cost, time delays, and many risks of moving data outside of a trusted and secure platform with rigorous access controls and governance capabilities.

- **Value.** Teradata has focused on providing cost-effective solutions, having curated a short-list of small, purpose-built open source models that are well suited to its initial use case solutions and

platform capabilities. Teradata is thereby helping customers take advantage of GenAI innovation without wasting money on large models and GPU acceleration when they are not required.

Market-leading and fast-following organizations looking to innovate with GenAI will need a capable data platform, diverse model choices, and flexible design options to deliver breakthrough insights that differentiate their organization. Constellation sees the BYO-LLM and GPU acceleration options as good reasons to consider Teradata's VantageCloud Lake platform.

ANALYST BIO

Doug Henschen

Vice President and Principal Analyst

Doug Henschen is vice president and principal analyst at Constellation Research focusing on data-driven decision-making. His Data to Decisions research examines how organizations employ data analysis to reimagine their business models and gain a deeper understanding of their customers. Data insights also figure into tech optimization and innovation in human-to-machine and machine-to-machine business processes in manufacturing, retailing, and services industries.

Henschen's research acknowledges the fact that innovative applications of data analysis require a multidisciplinary approach, starting with information and orchestration technologies; continuing through business intelligence, data visualization, and analytics; and moving into NoSQL and big data analysis, third-party data enrichment, and decision management technologies. Insight-driven business models and innovations are of interest to the entire C-suite.

Previously Henschen led analytics, big data, business intelligence, optimization, and smart applications research and news coverage at InformationWeek. His experiences include leadership in analytics, business intelligence, database, data warehousing, and decision support research and analysis for Intelligent Enterprise. Further, Henschen led business process management and enterprise content management research and analysis at Transform magazine. At DM News, he led the coverage of database marketing and digital marketing trends and news.

@DHenschen constellationr.com/users/doug-henschen linkedin.com/in/doughenschen

ABOUT CONSTELLATION RESEARCH

Constellation Research is an award-winning, Silicon Valley–based research and advisory firm that helps organizations navigate the challenges of digital disruption through business model transformation and the judicious application of disruptive technologies. Unlike the legacy analyst firms, Constellation Research is disrupting how research is accessed, what topics are covered, and how clients can partner with a research firm to achieve success. Over 350 clients have joined from an ecosystem of buyers, partners, solution providers, C-suites, boards of directors, and vendor clients. Our mission is to identify, validate, and share insights with our clients.

Organizational Highlights

- Institute of Industry Analyst Relations (IIAR) New Analyst Firm of the Year in 2011 and #1 Independent Analyst Firm for 2014 and 2015
- Experienced research team with an average of 25 years of practitioner, management, and industry experience
- Organizers of the Constellation Connected Enterprise—an innovation summit and best practices knowledge-sharing retreat for business leaders
- Founders of Constellation Executive Network, a membership organization for digital leaders seeking to learn from market leaders and fast followers

www.ConstellationR.com

@ConstellationR

info@ConstellationR.com

sales@ConstellationR.com

San Francisco Bay Area | Boston | Colorado Springs | Denver | Ft. Lauderdale | New York Metro
Northern Virginia | Portland | Pune | San Diego | Sydney | Washington, D.C.