DELL Technologies

WHITE PAPER

A DELL TECHNOLOGIES PERSPECTIVE

GenAI for Retail

Chandra Venkatapathy, CTO, Retail Industry Solutions, Dell Technologies

The retail industry is rapidly changing, driven by shifting customer buying patterns. The industry has embarked on customer experience initiatives to improve sales and to lower costs with operational efficiencies including inventory and supply chain optimizations. Data is foundational to drive these initiatives and artificial intelligence (AI) is playing a key role in accelerating decision-making with accurate forecasting and predicting patterns. Generative AI (GenAI) technology has taken this further by democratizing the interaction with AI systems with conversational and interactive interfaces that can generate new and exciting innovations empowering both customers and employees alike. The goal of this paper is to provide a glimpse into GenAI technologies and how retailers can adopt and drive innovation.

What are GenAI and LLM?

Many of you have heard about GenAI, specifically ChatGPT[1] and Bard.[2] Let us quickly establish what GenAI is and how it is different from traditional AI and machine learning (ML).

Retailers have been using traditional AI technologies for pattern detection, forecasting and predictive events. Traditional AI is a technique where the AI model is trained with known patterns and rules. Once learned, the system can automatically detect, forecast and predict without any human intervention. This has helped retailers with innovations in demand forecasting, price promotions, and recommendation engines.

GenAI, unlike a traditional AI, can create new content based on how it has trained. The foundation of GenAI is a large language model (LLM) that is a special class of AI algorithms called transformers.[3] LLMs are trained with corpora containing billions or trillions of words, teaching the model to know the context, priority and placement of words. The trained model, called a foundational model, can generate new content to summarize a text, answer a question with context and accuracy—similar to how humans will respond, or generate a rich document with text, sound, images and data. This opens new opportunities for how retail processes can be improved where humans are involved.

Though LLMs have been available for the past few years in the AI community, advancements in deep learning and the increase in computing power have allowed for the proliferation and refinement of these solutions. Recent high-profile announcements with simple and highly intuitive interfaces have captured everyone's imagination, including that of retailers. The simplification and the usability improvements have democratized the technology that is immensely helpful to any industry vertical, including the retail industry. A recent study shows 69 percent[4] of CEOs see broad benefits of generative AI across the organization, revealing a broad interest from the top in the adoption of GenAI.

[1] ChatGPT is a large language model developed by OpenAI and widely used for general conversation, text generation, entertainment and learning. https://openai.com/blog/chatgpt

[2] Bard is a large language model chatbot developed by Google AI. https://bard.google.com/

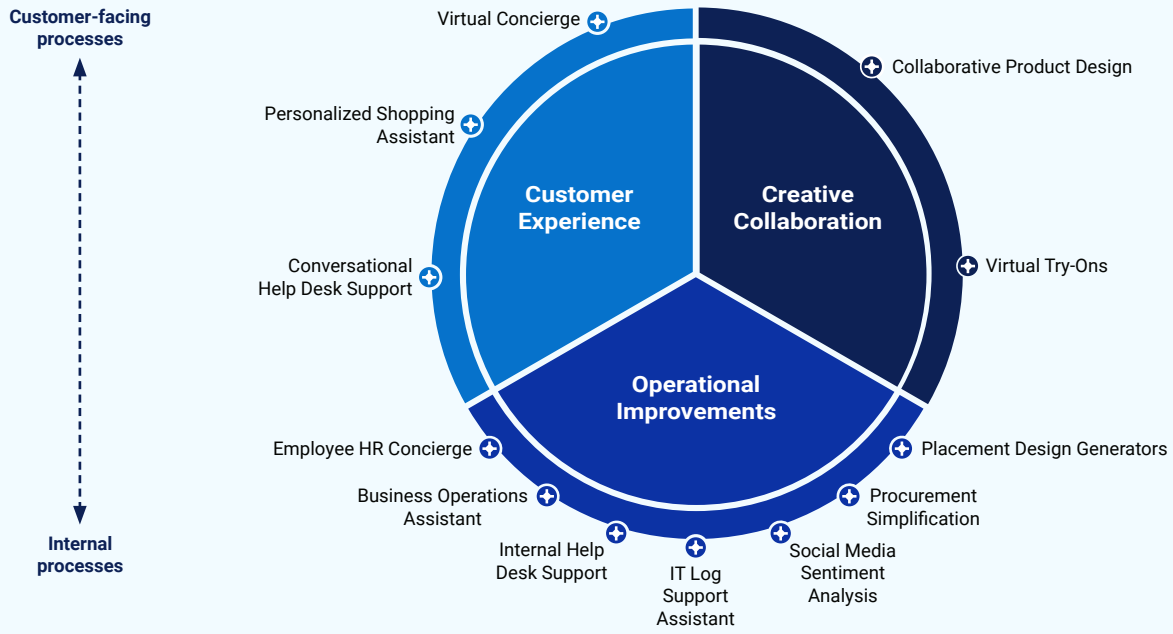
[3] A transformer is a deep-learning model that can capture relationship, dependencies and significance within large pools of data, such as sequential text.

[4] "CEO Decision Making in the Age of AI." https://bit.ly/3YKiTtO

Retail use cases

Let us look at the retail use cases that can be enabled by GenAI. As established, GenAI is ideal where there is language, communication, expression or interaction involving humans and machines or humans and humans. Here are examples of the use cases touching both internal and external processes. From an adoption standpoint, internal use cases will help with learning and the best practices around data, security, governance and technology.

Figure 1. Retail use cases enabled by GenAI



USE CASES FOR IMPROVING INTERNAL PROCESSES FOR COST AND EFFICIENCIES

IT and HR Support Functions

- **Internal Help Desk Improvements:** Increasingly, internal IT support functions are required to do more or be outsourced, and retailers are looking at ways to improve the quality of service for internal support functions. LLMs can play a role in providing customized and context-sensitive help, reducing IT support costs.
- **Log Scans / Event Monitoring:** Retail IT operations teams collect millions of logs and events spanning multiple systems. Translating the voluminous text and numbers into real-time data to drive actionable insights is often hard. LLMs' capabilities will help drive added insights with summarization and sentiment analysis.
- **HR Virtual Concierge:** Employee welfare is a major priority for retailers in a tough labor market. Employers want to offer personalized assistance to help navigate scenarios and provide the best customer experience, augmenting existing customer and chatbot operations where LLMs can play a role.

Core Retail Process Improvements

- **Business Operations Support Assistants:** Intelligent Decision Agents for Price Promotions / Demand Forecasting / Supply-Chain Optimization / Core Business Operations: Today, retailer business operations teams generate business intelligence (BI) reports that are targeted at a specific audience. Often these reports must be acted on by a broader community within retail operations that require more clarifications and insights. An intelligent LLM agent can help with queriess, questions and answers, and recommendations. For example, a store manager can work with a store system asking simple questions about what the data says and hear suggestions on what actions to take. Earlier, such queries required an IT or database business analytics person to interpret the data and provide actionable recommendations.

- **Creative Marketing and Marketing Campaign Optimization:** Personalized marketing and campaigns involve creative content generation with inputs from internal sources and agencies, but these inputs are usually compiled manually, costing time and money. LLMs can automate this task by generating personalized marketing collaterals with rich content—text, images and sounds with context and accuracy to accelerate campaign execution.

- **Customer Review / Social Media Sentiment Analysis:** Retailers' marketing teams are hard at work to find, scan, filter, summarize, and classify voluminous and growing social media data. With LLMs' ability to provide sentiment analysis and summarization, marketers can derive valuable information from social media content.

- **Procurement Process Simplification:** Procurement teams deal with contracts, specifications, support and delivery documents that are rich with text and images. LLM-based applications can simplify document processing with text extraction, summarization, and contract generation and validation, saving time for both procurement teams and vendors.

- **Creative Design Generation of Products and Layouts:** Privately labeled products deliver higher margins; however, they need a strong product design, sourcing, and manufacturing teams. LLMs offer rich content generation with image, text, sound and synthetic data to help with product designs. The same applies to store planners who can simulate several different configuration ideas.

USE CASES FOR IMPROVING EXTERNAL/CUSTOMER-FACING PROCESSES

Improving customer experiences is a top priority of many retailers. Here are some of the use cases.

- **Personalized Recommendations:** Customers want personalized service and data shows that 80% of customers expect personalization.[5] LLMs can enhance with personalized recommendations and provide product help to guide customers through the buying process.

- **Drive-Through Assistance:** Quick-service restaurants (QSRs) have already started using drive-through assistants that can listen, recommend, and help with ordering using voice- and text-based chatbots to reduce ordering time and to improve accuracy.

[5] "Personalizing the customer experience: Driving differentiation in retail." McKinsey & Company

DOMAIN-SPECIFIC DATA REQUIRES QUALITY, INTEGRITY AND PERFORMANCE

Enterprises need to have good quality controls and performance KPIs for enterprise retail AI applications. Here are a few key performance monitors:

- **Model Quality and Performance:** Poorly trained models "hallucinate,"[6] a term often used to describe content that looks good from a language perspective but may have incorrect or false information. Monitoring the model's performance is important.
- **End-User Application SLAs:** Inference infrastructure should be capable of handling end-user applications load and query serving time.

- **Customer Support Desk:** Today, customer support functions are offered via voice lines that are expensive to operate or text-based chatbots that are rudimentary in function and often incapable of answering questions, driving more call volumes. Chatbots with LLMs can now deliver more context-sensitive and meaningful, human-like conversations that are customizable with parameters like style, tone and language choice.
- **Virtual Concierge / Shopping Assistants:** LLM-powered virtual concierges can further personalize the experience. Realistic GenAI-based virtual avatars, with highly contextualized facial expressions, tones, voice gradients and accents, can have conversations that feel real. For example, Dell Technologies is introducing Clara, a GenAI chatbot, to help with internal sales assistance.
- **Virtual Try-Ons:** GenAI can generate images and artifacts that can help with personalized try-ons.

Enterprise GenAI: Considerations

Thanks to the popularity of services like ChatGPT and Bard, the usual mindset is to start with publicly available services that are largely designed for general conversations, learning and entertainment. Here are the considerations for the use of these services for retail enterprise needs:

- **Model Fit and Licensing:** It is safe to assume that many commercially well-known models have restricted commercial use. In addition, there is no guarantee that the model matches the needs of retail use cases that requires domain-specific data.
- **Content Ownership:** Retail enterprises should have visibility on the ownership and authenticity of the content. Knowing more about the source of training data, control and ownership of the model will simplify dealing with intellectual-property related challenges in the future.
- **Bias:** In AI, bias is the systemic behavior of the foundational model to like or dislike a specific attribute like gender, race, politics or nationality which could result in incorrect, unethical or harmful results. Data scientists should evaluate the foundational model for training bias to reflect core enterprise values.
- **Domain-Specific Data:** Models need domain-specific information like inventory, customer relationship management (CRM), pricing, and customer data that are confidential and cannot be exposed to commercially available LLM services without sufficient safeguards. Operating on a known security infrastructure will help with data protection and governance. *See sidebar for additional considerations.*
- **Ethics:** The generated content should reflect and exemplify ethical values and social responsibility. The model should be vetted for maintaining retailers' ethical values.
- **Cost:** With the potential for GenAI-enabled retail applications to become mainstream, cost will be a factor. Commercial services offer different business models including cost-per-transaction by size. Enterprises understand how the costs of services grow with scale.

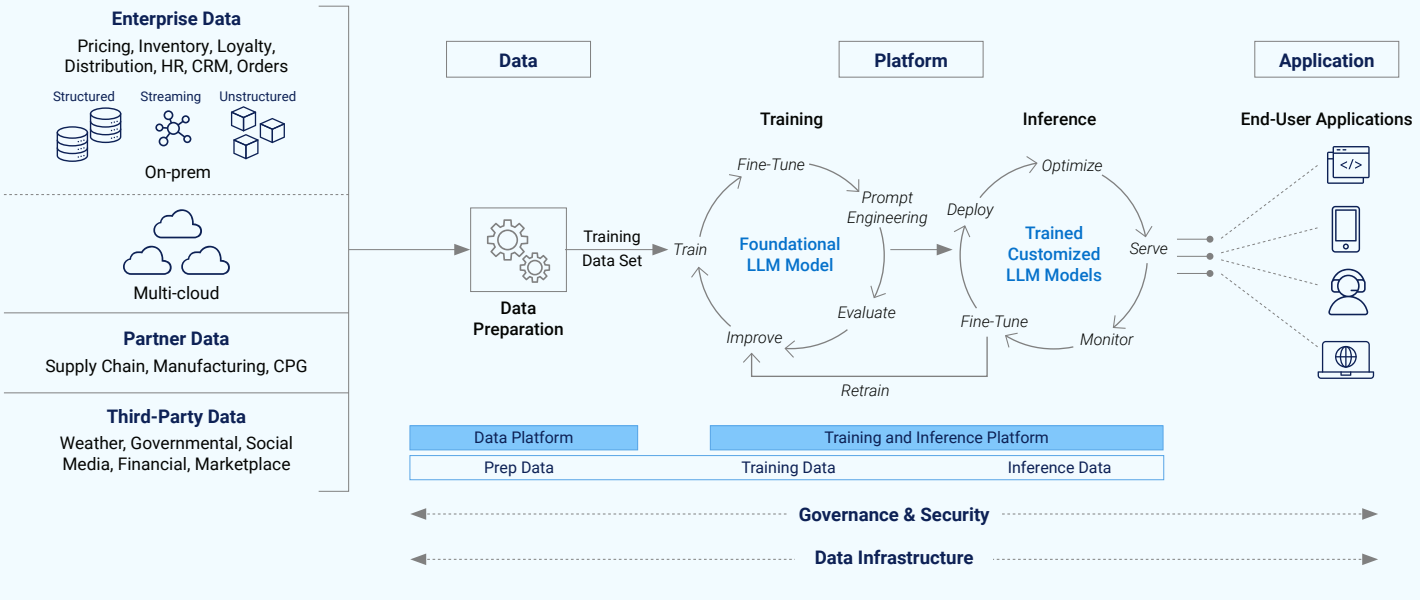
Taking all the above factors into consideration, a private GenAI platform with its own private model that can be customized with enterprise data and have the needed security and governance can help retailers in their LLM journey.

[6] https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/

Elements of Enterprise GenAI solutions

Enterprise GenAI solutions broadly consist of data needed for the use case, a platform that helps to train the model with enterprise data, and the applications that use the trained models to drive outcomes.

Figure 2. Enterprise GenAI solution



Data: Data plays a key role in customizing and fine-tuning the LLM to your needs. Retail data comes from diverse sources—core retail enterprise applications such as point of sale (PoS), inventory merchandizing, eCommerce, and enterprise resource planning (ERP); software-as-a-service (SaaS) applications like CRM; marketing; HR; and third-party data for personal finance, weather, credit agencies, and economic information. Data comes in various types: structured, semi-structured, unstructured, real-time, and time-series. A training data set is created by extracting and preparing the right data for the chosen use case. For example, an internal IT help desk may need all support documents, how-to-guides, videos, product support information, past incidents and how they were resolved, internal social media groups, FAQs and other similar data points to help train the model for IT support functions.

Training and Inference Platform: Building an LLM from scratch is time-consuming and costly, requiring deep skills. Pre-trained foundational models are a good starting point that can be trained with domain-specific data and further customized. Enterprise data scientists need a platform and tools to help with the training and inference life cycle. Large language model operations (LLMOps) tools help with monitoring the performance of the model for quality and throughput. Any deviations will be addressed by further fine-tuning and/or retraining. The infrastructure should be scalable as requirements can change with the model's training and inference loads.

Applications: Enterprise applications invoke GenAI services by making an API call to the inference model that is served. The inference infrastructure should be capable of handling a volume of requests with well-defined query-processing times and sufficient scalability.

Governance and Security: Governance plays an important role in the generative content. Enterprise governance teams should have a process and procedures for verifying data sources, intellectual properties (IPs) associated with generated content, and ethics.

Dell Validated Design for Generative AI with NVIDIA: An end-to-end stack for enterprise-ready LLM applications

Dell Validated Design for Generative AI with NVIDIA[7] is a joint effort with NVIDIA to provide an integrated stack for enterprise GenAI, that has been optimized and validated for rapid adoption. It offers a design blueprint with configuration guidance and best practices using Dell Technologies PowerEdge Servers with a choice of NVIDIA GPUs, **NVIDIA AI Enterprise software**, and Dell software at its core. Data is an important piece of any GenAI solution and Dell Technologies PowerScale and ObjectScale help with scalable unstructured storage.

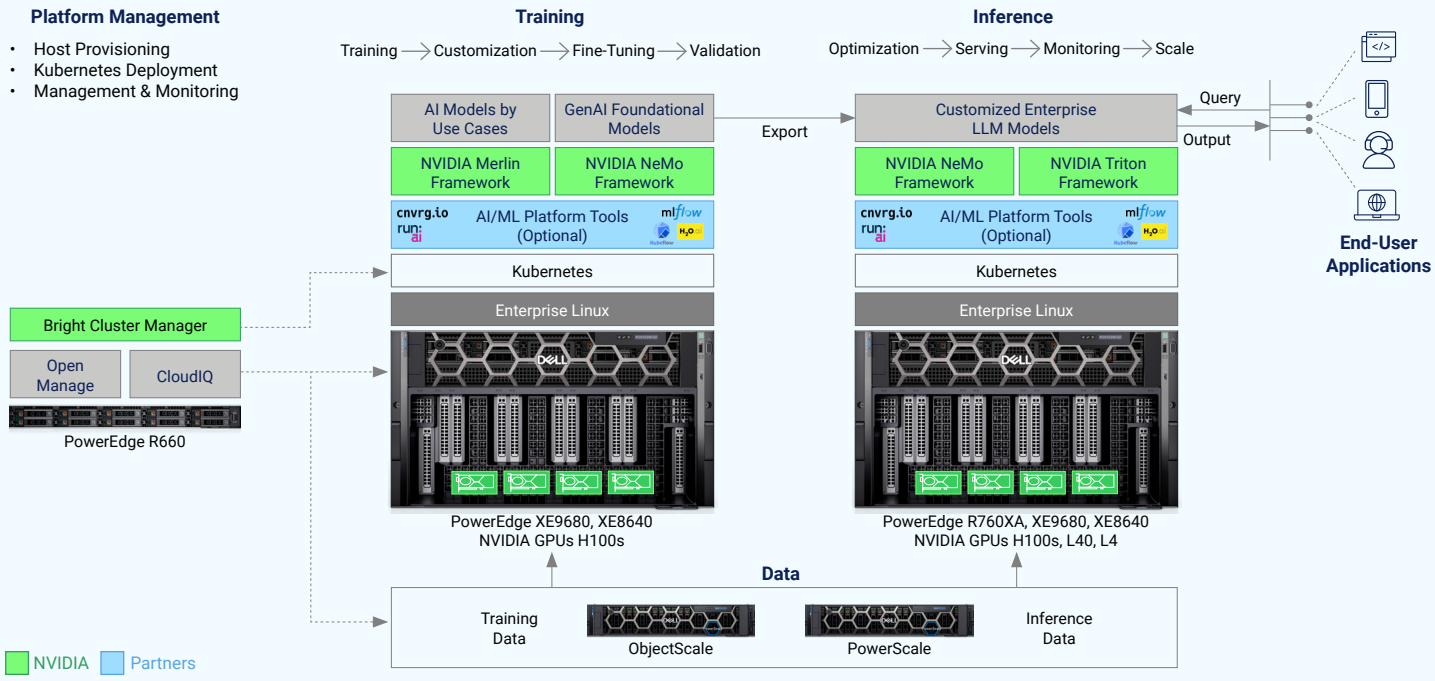
Here are the elements of the validated design:[8]

	Capability	Description
Training & Inference Infrastructure	Training	Dell Technologies AI-optimized servers for training, powered by Dell PowerEdge XE9680 and XE8640 servers with NVIDIA H100 GPUs
	Inference	Dell Technologies AI-optimized servers for inference, powered by Dell PowerEdge XE9680 and XE8640 servers with NVIDIA H100 GPUs
	Management	Module for system and cluster management, including a head node for NVIDIA Bright Cluster Manager (BCM) powered by Dell PowerEdge R660 servers. Infrastructure management by Dell OpenManage and CloudIQ
	Network	Module for high-throughput and high-bandwidth communication between other modules in the solution powered by Dell PowerSwitch Z9432F-ON high-bandwidth GPU-to-GPU communication, powered by NVIDIA QM9700 InfiniBand switches
Training & Inference Life Cycle	Training	NVIDIA NeMo™ framework provides the capabilities for training LLMs. • Use of pre-trained foundational models • Customize model by adding tasks like RLHF, P-tuning, fine-tuning • Feed knowledge with proprietary info for the use case • Deploy the model for testing and validation
	Inference	NVIDIA Triton™ Inference Server for inference serving: loading of models, performance optimization, scaling by utilization, and support for multiple models
	LLMOps	In addition to NVIDIA AI Enterprise Suite, Dell Technologies has an ecosystem of partners for machine learning operations (MLOps) to further help with the training and deployment life cycles that have been validated on the Dell Technologies stack.
Data	Data Prep, Training & Inference	High-throughput, scale-out network-attached storage (NAS) powered by Dell PowerScale, plus high-throughput scale-out object storage powered by Dell ECS and Dell ObjectScale

[7] https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2023~07~20230731-dell-technologies-expands-ai-offerings-to-accelerate-secure-generative-ai-initiatives.htm#/filter-on/Country:en-us

[8] The currently available design is focused on inference.

Figure 3. Dell Validated Design for Generative AI with NVIDIA



Here are the steps for the LLM application workflow:

- **Provision the Infrastructure** using NVIDIA Bright Cluster Manager (BCM), which configures the cluster/node for training and inference with the needed compute, memory, storage and GPUs along with the Kubernetes layers. Dell OpenManage and CloudIQ help with monitoring and management of the server and storage infrastructure layers.
- **Model-Training Life Cycle:** NVIDIA NeMo™ framework helps with building, training and customizing of the foundational model with enterprise data. It enables ML engineers to set up the model and iterate through the model life cycle.
- **Inference / Model Serving:** The trained and customized enterprise model is ready to deploy for inference. NVIDIA Triton™ takes the trained model, evaluates it for deployment and provides an optimal configuration for model serving. The deployed model then can be scaled up and down based on utilization. End-user applications can now invoke the LLM to deliver new and innovative services.

Dell Validated Design for Generative AI with NVIDIA: Value to retailers in their LLM journeys

Dell Validated Design for Generative AI with NVIDIA is a collaboration between Dell Technologies and NVIDIA to enable high-performance, scalable, and modular full-stack generative AI solutions for large language models (LLMs) in the enterprise. This collaboration was created to simplify adoption and accelerate innovation for GenAI. Here are some of the values and benefits.

- **Integrated Stack:** This fully integrated and validated stack provides for training and inference, accelerating time to value. Data sciences teams can start with either do-it-yourself models or pre-trained models. The design is modular and comes with a choice of configurations to help with growing needs for both training and inference.
- **Validated Design:** Dell Technologies and NVIDIA have published a validated design document that provides guidance for configuration and best practices to help jump-start a GenAI infrastructure program.
- **Partner Ecosystem:** Dell Technologies works with a robust partner ecosystem for data, AI/ML and value-added services to help drive outcomes.
- **Services:** Dell Technologies offers a choice of services including consultation, deployment, implementation, and managed services to help with rapid adoption of this technology.

How to get started with your GenAI project?

GenAI helps with digital transformation, and for many retailers the journey has already begun with their data, AI, and edge strategies. Though GenAI is an extension of the AI strategy, there are differences as it deals with generative content. Given the transformative nature of GenAI projects and the need for teaming, it is good to have executive support before the start of the project. Here are steps to consider as you start with your first project.

- **Plan:** GenAI offers exciting and innovative use cases. Develop a framework to evaluate a use case in various dimensions: business case, data and technology requirements, ability to execute, governance challenges, cost, skills and risk.
  - Data is critical to a GenAI strategy. A good data strategy can support current and future use cases.
  - Scope: Limit the scope of the use case to prove success that aligns with your overall business. Start simple with a limited scope for functions and the customer base. Choose use cases that are simple in technology needs.
  - Skills: Given the emerging nature of the technology, enterprises have skill and process gaps for GenAI. Have a good strategy for closing these gaps.
  - Risk is probably the most important factor for an initial project given the nature of generative content that may have correctness issues, bias, and IP challenges. It is better to start with an internal project limited to a small user base.

ABOUT THE AUTHOR

Chandra Venkatapathy is the Global CTO of Retail Industry Solutions for Dell Technologies. He has extensive experience in identifying industry trends and creating solutions for retailers pursuing digital transformations. He has engaged with many top 100 retailers in their retail transformation journeys leveraging data, AI/ML, edge, IoT and cloud technologies.

- **Team:** GenAI applications touch multiple teams, including data, IT, solution engineering, and app deployment. Given the sensitivity of generated content, it is recommended to engage the corporate ethics, legal, and compliance team upfront. Have a good process for engagement that supports an iterative process for continuous improvement.
- **Choose:**
  - **Technology Stack:** Retailers have options today to choose an integrated stack to address the end-to-end life cycle. Publicly hosted services are easy to access; however, data privacy and the long-term cost of serving queries must be considered. A privately owned stack from Dell Technologies can help a retailer explore its own data in a protected environment with tools for innovation.
  - **Model:** The market is booming with choices, and with innovation there will be more models targeting specific use cases. For your first projects, choose a good foundational model meeting the needs of the use case. The model should have the appropriate commercial terms for enterprise use.
- **Execute:** Set modest goals and continue to advance and grow with a continuous improvement process. This needs a good set of tools for measuring KPIs and a plan for improvement that may impact data, training, fine-tuning and end-use query.
- **Govern:** Governance and oversight for response for relevance, explainability, correctness, bias, and IP ownership should be in place to help with the content.

It is good to start with a "light house" project—an internal project that can capture the imagination and that can develop a blueprint for the future. Dell Technologies and our partners are ready to help to kick-start your project.

Summary

GenAI has captured the imagination of retailers with the potential to transform customer experiences and improve operations through a wide variety of use cases. Adopting GenAI for retail use cases requires using enterprise data and customizing publicly available models through a careful strategy for data protection, security, quality, performance, security, and compliance. Dell Technologies has helped in customers' data and AI journeys and is well positioned to help retailers with an infrastructure stack aligned with LLM life cycles. Dell Technologies and NVIDIA introduced generative AI solutions with a collection of capabilities and a validated design with configurations to help the data and AI stakeholders to accelerate adoption and time to value. The journey has just started, and you will see more innovations coming from Dell Technologies.

For more details, please visit **Dell Technologies Generative AI Solutions**.

Learn more at: **DellTechnologies.com/Retail** | Contact us at: **Dell.com/contact-us**