Enterprise Strategy Group by TechTarget

SHOWCASE

NOVEMBER 2024

Modernizing Retail With AI, Cloud, and Edge Computing

Jim Frey, Principal Analyst

Abstract: Innovation is essential for retailers to keep pace with the competition while simultaneously growing revenues, managing costs, and delighting customers. AI will play an important role for those ready to embrace and deploy the applications that can leverage the latest in AI foundational models via cloud edge computing infrastructure deployed at store sites. Google Cloud provides a fully managed hardware and software solution that extends cloud infrastructure to the edge via its Google Distributed Cloud (GDC) offering, powered by Intel, that puts retail operators in position for success today and in the AI-driven future.

Enabling AI for Retail

"Innovate or perish!" Peter Drucker's advice continues to ring true to this day, and running retail store operations is no exception. Customers' expectations of technological savvy and capability keep rising, while at the same time operating margins must be protected for profitability. The retailers that get this balance right are in position to thrive. Those that don't will struggle to keep pace with the competition. For retail operations at the store level, the goals are clear:

- **Store analytics:** Enable real-time insights to understand and improve customer experiences.
- **Loss prevention:** Provide real-time inventory tracking and movement insights, enabling proactive responses to potential loss events.
- **Fast transactions:** Ensure efficient transaction processing even during busy store times and when external network connectivity is congested for an ideal customer checkout experience.
- **Personalization support:** Leverage data insights processed directly in-store to deliver personalized shopping experiences for customers.
- **Employee enablement:** Empower staff to focus on enhancing shopper experiences, relationships, and value by automating operational tasks.

The most compelling innovation strategies for retail will leverage AI, cloud infrastructure, and integrated, localized computing, in ways that are optimized for retail store management. But where to start? There are myriad technology choices and all too many risks along the way in terms of fit, cost, and impact. No one can afford to get this wrong.

The Ideal AI-optimized IT Model for the Modern Retail Store

New technologies in cloud, AI, and edge computing are important enablers of innovation; the question is how to best use them in the context of store operations. Research from TechTarget's Enterprise Strategy Group has found

that on-premises, customer-facing locations such as retail stores are the second most common physical location for deployment of edge infrastructure as a way of enhancing remote operations and delivering AI optimizations.[1]

The ideal overall model will include AI-enabled retail application software that can be built and maintained using the power and flexibility of the cloud and then will efficiently deploy and host that software locally at the store level. This approach enables deployment on an ongoing basis while minimizing impact on both corporate and store staff, opening the door to continuous innovation and regulatory compliance. Delivered in this manner, store sites become first-class citizens to integrated IT architectures, just as the cloud is at the corporate level, so that the skills that corporate IT has built in using the cloud can be extended directly to the store level, reducing the cost of engineering and support.

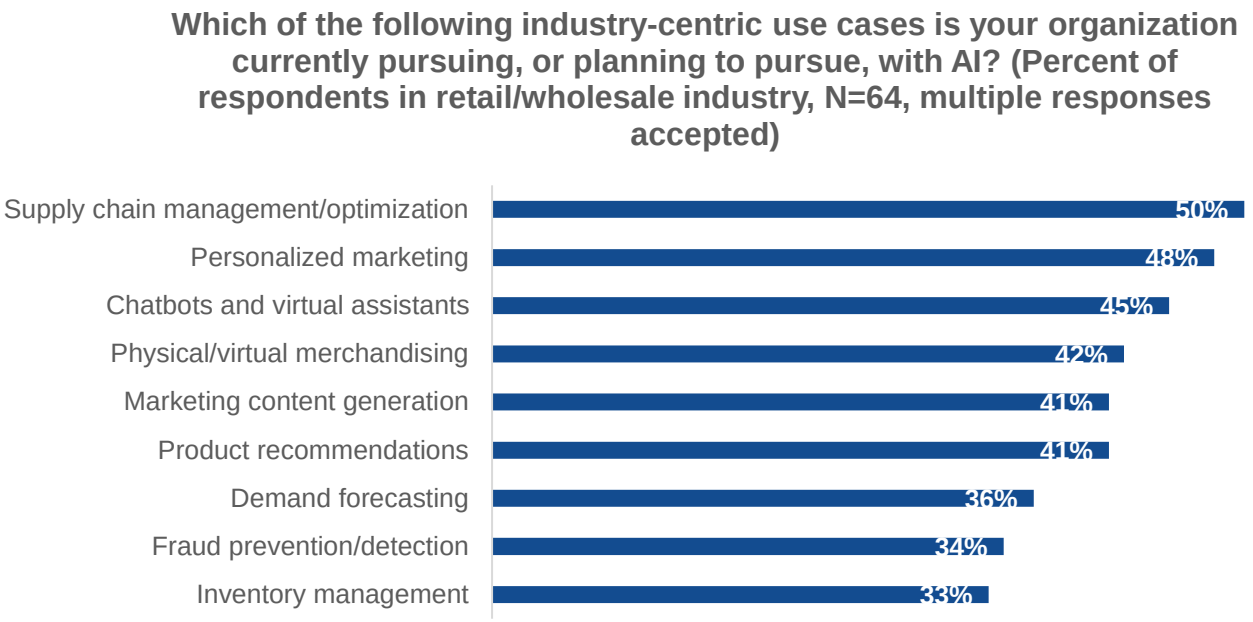
AI-enabled retail application software might either be designed and built by an independent developer specializing in retail solutions or within a company's own development teams as part of its "secret sauce." Whichever approach is taken, it must maximize flexibility to leverage whichever combination makes sense today while not creating barriers to changing the mix based on future changes and opportunities.

The next step is to find a means to deploy those applications locally, in-store, to minimize or eliminate operational risks around internet availability, network bandwidth, and application performance. This means having local compute resources to host the applications and data necessary to operate and secure the store. One way to get there is to leverage "edge infrastructure" that acts as a local IT hub while also being connected to cloud-based applications and resources. Edge infrastructure retains the ability to support all critical store operations, even when internet connectivity is lost or degraded.

The scope of uses for local edge computing starts with hosting common pre-existing functions like point-of-sale transaction processing, customer personalization and loyalty programs, and inventory management. But there is much more to consider, including sales reporting and analytics, worker schedule management, kiosks, environmental controls, and fraud loss prevention, just to name a few. These could be integrated applications or unique applications, and some require computing and connectivity that must be networked and monitored. The platform should also be ready to support emerging AI use cases for retail. Figure 1 shows findings from Enterprise Strategy Group's research into the most common AI use cases for retail organizations.[2] While some of them will be centered in corporate headquarters, almost all AI use cases will have a direct impact on IT requirements at the store level, gathering sales and customer data (learning) as well as executing local decision support (inferencing).

[1] Source: Enterprise Strategy Group Research Report, *Unleashing the Edge: Use Cases, Challenges, and Requirements for Edge Infrastructure and Environments*, March 2024.
[2] Source: Enterprise Strategy Group Complete Survey Results, *The State of Analytics and Business Intelligence Platforms*, April 2024.

Figure 1. Top AI Use Cases for Retail/Wholesale



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Beyond these examples, there are many more potential applications of AI and generative AI coming, such as pricing and promotion optimization, wait time analysis, in-store customer sentiment analysis, demand forecasting, workforce assistance, planogram compliance, and planogram optimization. Each of these will require a flexible compute platform at every store location.

In order to be a cost-effective, low-risk win for retailers, edge computing solutions must have the following attributes:

- **Easy to deploy.** The platform must be simple and quick to set up and configure, requiring little or no direct responsibility of in-store employees and including few, if any, requirements for upgrading store infrastructure such as electrical power and networking. Further, it should use the same tools and frameworks used to create, deploy, maintain/support, and evolve functionality built in modern cloud environments, maximizing ROI on organizational cloud development investments.
- **Flexible.** Edge platforms should not restrict retailers from adding or changing any aspect of applications and systems in use. This makes the platform a future-proofing enabler, rather than a barrier, for business growth and evolution, particularly for innovations around applying AI models. Further, they should deliver a uniform experience from the latest in open source and high-performance cloud infrastructure to on premises, effectively connecting and extending the cloud to the store.
- **High performing.** The solution should provide access to the latest in AI models along with sufficient local computing capacity to support today's applications as well as coming innovations around AI for retail.
- **Highly available.** The heart of edge computing is making sure that stores can operate smoothly and continuously, even when external communications are interrupted or degraded. Edge compute infrastructure must be engineered for such standalone operations.
- **Highly reliable.** It's one thing to acquire, deploy, and configure edge computing infrastructure, but the chosen platform also needs to have a long and productive lifetime. Ideally, the edge solution will come inclusive of 24/7 operational monitoring and support, much like today's cloud computing environments, to eliminate operational risks while also avoiding requirements for direct support by either store personnel or central corporate IT.

The Google Distributed Cloud Solution, Powered by Intel

As one of the earliest and most influential providers of cloud services and AI, Google Cloud has focused on the best ways to deliver and leverage cloud-centric and hybrid solutions. Recently, this has expanded to include fully managed hardware and software infrastructure extending AI models and AI-enabled cloud infrastructure to the edge via Google Distributed Cloud (GDC). The platform provides a fully functional local solution offering inclusive of compute, storage, and networking, powered by Intel and designed for deployment at store sites to empower modern retail operations.

Google Distributed Cloud meets the needs of the ideal local store IT model as follows:

- **Easy to deploy.** Google Cloud offers GDC, a fully managed hardware and software configuration, that typically requires no special local accommodations for power, cooling, or networking. Further, this approach eliminates the need for local store employees to support deployment and there is no need to assign corporate IT personnel to each store.
- **Flexible.** GDC delivers a cloud-native Kubernetes platform on site, enabling retail teams to either select independent software solutions from a thriving, open ecosystem of thousands of partners or develop proprietary applications themselves. The platform is AI-ready, so developers can use the latest AI models to build and train AI-native applications centrally, including the use of data gathered from store sites for training, and then easily deploy those applications from the cloud to hundreds or thousands of locations. The result is a contiguous application development and deployment environment from cloud to store.
- **High performing.** GDC leverages 5th Gen Intel Xeon Silver processors to enable optimal, scalable compute performance with minimal power and cooling demands. Intel's Xeon processor is designed for AI and has a long-proven history of successful, reliable deployments in retail settings. As the compute functions are local, this also minimizes potential application latency by eliminating reliance on external networking and remote application processing. Essential data processing is also kept locally, improving security and sovereignty.
- **Highly available.** With applications running locally, there is never a problem when internet connectivity is lost or degraded due to load. GDC can continue to operate for up to seven days without cloud connectivity at a minimum, so store operations can continue uninterrupted.
- **Highly reliable.** GDC provides 24/7 post-deployment monitoring and remote support, essentially eliminating the load on corporate IT teams for keeping GDC compute infrastructure up and running smoothly.

As an edge computing solution designed to meet the innovation and operating objectives of retail stores, GDC matches with the ideal AI-optimized, cloud-connected model as a cost-effective, flexible, low-risk approach to edge computing, so retailers can focus on delighting customers and maximizing sales.

Conclusion

Retail business is relentlessly competitive, and store owners need to prepare for a future where AI opens new opportunities, delights customers, and delivers the necessary competitive edge. Steps taken as part of a regular periodic tech refresh to engage and deploy intelligent edge infrastructure, ready to connect the cloud to the store for leveraging AI-enabled business applications and analytics, can be the platform for ensuring success. The GDC solution, powered by Intel, represents an AI-optimized, cloud-connected local approach that is ideally designed for innovating and AI-enabling retail use cases, unlocking the store of the future, working for you today.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com