Enhancing Retail Operations with AI and Vector Search

The business case for adoption

Table of Contents

Introduction

AI, or artificial intelligence, is quickly becoming a universal tool that fits in every retailer's toolbox. Soon after early machine learning and AI predictive capabilities harnessed the power of big data to give enterprises deeper business analytics at eye-popping speed, and new advances in generative machine learning applications like OpenAI and Hugging Face opened up possibilities for generating and analyzing text data. Today, generative AI-enriched applications go beyond text data, creating hyper-personalized experiences with capabilities like advanced semantic search or user-prompted conversations. While implementing AI technology can be risky, complex, and time consuming, the potential for benefits like higher profits, faster innovation, and lower costs are driving retailers toward an AI-powered future.

The world's most innovative retailers haven't missed a beat in turning robust operational data into the ingredients for powerful applications with the potential to transform the status quo. In this paper, we will explore the top use cases across the retail industry that are infused with MongoDB Atlas AI capabilities.

Let's look closer at the use of generative AI in retail. GenAI is creating unique and personalized experiences for customers at one end of the spectrum, improving value in customer touch points as well as improving back office inventory management systems, and optimizing marketing campaigns. With the use of genAI, retailers can come up with new products and offerings, define upsell strategies, create continually evolving marketing materials, and enhance customer experiences. One of the most creative uses is to understand the customer needs and preferences that change continually with season, trends and socio economic changes.

Personalized recommendations and experiences for customers is a key benefit and prime use case for Retailers. By analyzing **customer data** and **behavior**, genAI can create personalized **product recommendations**, customized marketing materials, and unique shopping experiences that are tailored to individual preferences.

Back office use case for genAI is to improve **inventory management** by **predicting demand** for products and optimizing inventory levels accordingly. This can help retailers to **reduce waste** and optimize their supply chain, leading to cost savings and increased efficiency.

Fraud detection is another important application of genAI in retail. By analyzing patterns in transaction data, retailers can identify and help alert the fraudulent activity that could be used in protecting from financial losses.

AI and the Developer Data Platform

MongoDB Atlas, the ground-breaking developer data platform, integrates operational, analytical, and generative AI data services, simplifying the development of intelligent applications. Whether you're deploying machine learning models or integrating cutting-edge generative AI into your applications, MongoDB Atlas is an indispensable component of your technology stack. From inception to deployment, MongoDB Atlas ensures that your applications are grounded in accurate operational data while meeting the demands of scalability, security, and performance expected by users.

MongoDB has already seen widespread adoption for traditional AI use cases. Continental selected MongoDB for the feature engineering platform in its Vision Zero autonomous driving initiative. Both Bosch and Telefonica use MongoDB in their AI-enhanced IoT platforms. Kronos trades billions of dollars of cryptocurrency every day using ML models configured and built with data from MongoDB. Iguazio uses MongoDB as the persistence layer for its data science and MLOps platform, while H2O.ai and Featureform support MongoDB as feature stores in their respective platforms.

Flexible Data Model

At the heart of MongoDB Atlas lies its flexible document data model and developer-friendly query API. Together, they empower developers to accelerate innovation, gain a competitive edge, and seize new market opportunities presented by generative AI. Documents, which align seamlessly with code objects, offer an intuitive and adaptable way to manage data of any structure. Unlike traditional tabular data models, and documents afford the flexibility to accommodate diverse data types and application features, facilitating data rationalization and utilization in ways previously unattainable.

Rapid Querying

Paired with the document model, the MongoDB Query API provides developers with a unified and consistent approach to data access & manipulation across various data services. From basic CRUD operations to complex analytics and stream processing, the MongoDB Query API offers developers the flexibility to query and process data according to the application's requirements. In the realm of Generative AI, this flexibility enables developers to define additional filters on vector-based queries, such as combining metadata, aggregations, and geo-spatial search, enriching the user experience and expanding application capabilities. MongoDB Atlas stands apart by offering a comprehensive suite of query functionality within a single, unified experience. This eliminates the need for developers to manually integrate query results from multiple databases, reducing complexity, errors, costs, and latency. Moreover, it maintains a compact and agile technology footprint, enabling developers to focus on building end-user functionality with greater ease and efficiency.

The Rise of Real-Time Analytics

Across the retail industry, companies are often falling short on their ambitions to build data-driven operations as they struggle to perfect real-time analytics using real-time events data.

With MongoDB Atlas App Services, these businesses are able to reinvent pricing strategies to reflect real-time market fluctuations, demand surges, or coverage changes. It's key to recognizing the importance of transforming raw data into a more usable structure and understanding the benefits of serverless functions and triggers, which can automatically respond to changes in data and execute predefined actions with a dedicated server.

LEARN MORE

Real-Time Analytics and Dynamic Pricing

Vectors, Unstructured Data, and MongoDB Atlas Vector Search

To feed AI models with proprietary data, there is a need to create vector embeddings. Data in any digital format and of any structure – i.e., text, video, audio, images, code, tables – can be transformed into a vector by processing it with a suitable vector embedding model. This incredible transformation turns data that was previously unstructured and, therefore, completely opaque to a computer into data that contains meaning and structure inferred and represented via these embeddings. Now users can search and compute unstructured data in the same way they've always been able to with structured business data. Considering that more than 80% of data is unstructured, it's easy to appreciate how transformational vector search combined with GenAI really is.

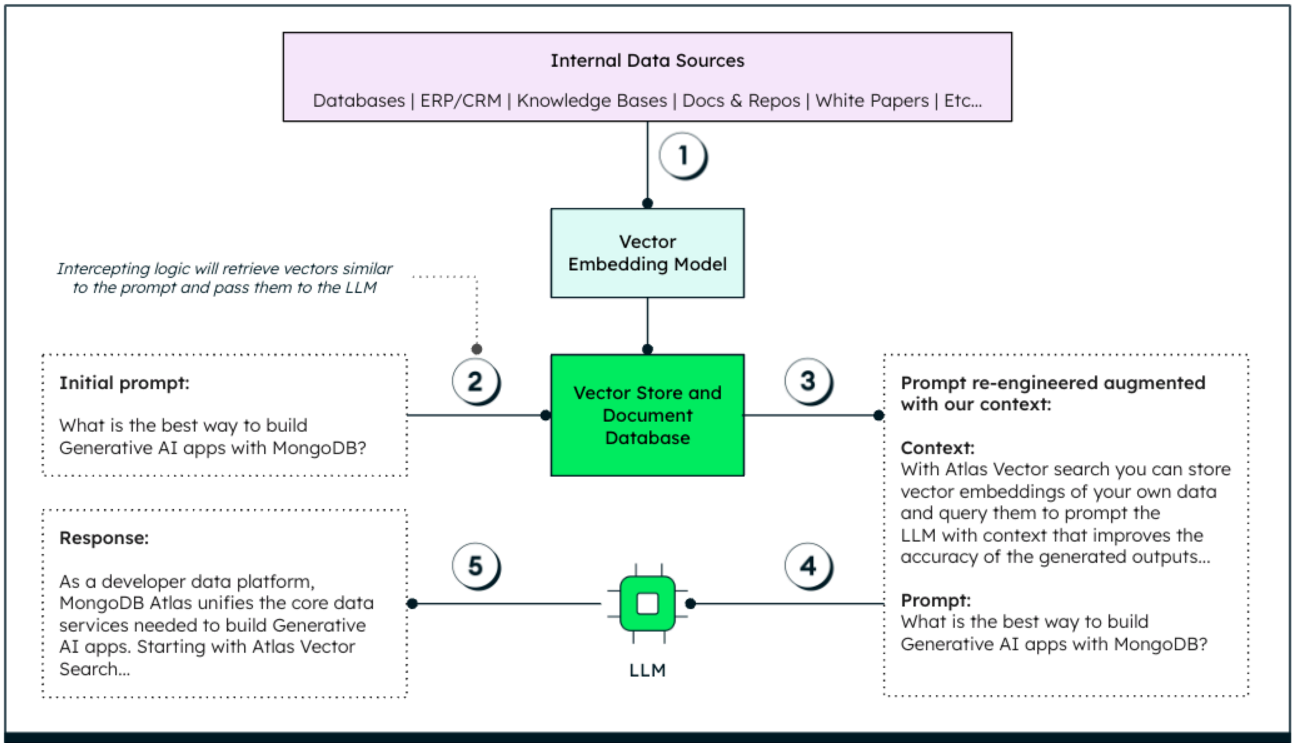
Internal Data Sources
Databases | ERP/CRM | Knowledge Bases | Docs & Repos | White Papers | Etc...
1
Vector Embedding Model
Intercepting logic will retrieve vectors similar to the prompt and pass them to the LLM
2
Initial prompt:
What is the best way to build Generative AI apps with MongoDB?
Vector Store and Document Database
3
Prompt re-engineered augmented with our context:
Context:
With Atlas Vector search you can store vector embeddings of your own data and query them to prompt the LLM with context that improves the accuracy of the generated outputs...
Prompt:
What is the best way to build Generative AI apps with MongoDB?
Response:
As a developer data platform, MongoDB Atlas unifies the core data services needed to build Generative AI apps. Starting with Atlas Vector Search...
5
4
LLM

Figure 1. Data is transformed from unstructured internal sources to actionable, impactful insights.

Once data has been transformed into vector embeddings, it is persisted and indexed in a vector store such as MongoDB Atlas Vector Search. To retrieve similar vectors, the store is queried with an Approximate Nearest Neighbor (ANN) algorithm to perform a K Nearest Neighbor (KNN) search using an algorithm such as 'Hierarchical Navigable Small Worlds' (HNSW).

LEARN MORE

AI-Augmented Search in Ecommerce

A Dive Into Ecommerce

With the use of generative AI, retailers can create new products and offerings, define and implement upsell strategies, generate marketing materials based on the market conditions, and enhance customer experiences. One of the most creative uses helps retailers understand customer needs and choices that change continually with season, trends and socio economic shifts. By analyzing customer data and behavior, Generative AI can also create personalized product recommendations, customized marketing materials, and unique shopping experiences that are tailored to individual preferences.

AI plays a critical role in decision making at retailer enterprises; product decisions such as design, pricing, demand forecasting, and distribution strategies require complex understanding of a vast array of information from across the organization. To ensure that the right products in the right quantities are in the right place at the right time, back office teams leveraged machine learning arithmetic algorithms for years.
As technology has advanced and the barrier for entry is lowered for adopting AI, retailers are moving towards data-driven decision making where AI is leveraged in real time. Generative AI is used to consolidate information and provide dramatic insights that could be immediately utilized across the enterprise.

AI-Augmented Search and Vector Search

Modern retail is a *customer centric* business, and customers have more choice than ever in where they purchase a product. To retain and grow their customer base, retailers are innovating at speed to offer each customer a differentiated buying experience. To do this, it is necessary to capture a large amount of data on the customers themselves, such as buying patterns, interests, interactions, and to be able to quickly make complex decisions on that data.

One of the key interactions in an ecommerce experience is search. Through the implementation of full-text search engines, customers are able to more easily find items that match their search, and retailers are given the opportunity to rank those results in a way that will give the customer the best option. In the past, decisions on how to rank search results in a personalized way were made by segmentation of customers through data acquisition from various operational systems, moving it all into a data warehouse then subsequently running classical AI with various Machine Learning algorithms on such data. Typically this would run every 24 hours or a few days, in batches, and the next time a customer logs in, they will have a personalized experience. It does not, however, capture the customer's true desire now they have returned to the website.

These days, modern retailers augment search ranking with data from real-time responses and/or analytics from AI algorithms. Also, it's now possible to incorporate factors such as the current shopping cart/basket and customer click stream and/or trending purchases across shoppers.

The first step in truly understanding the customer is to build a customer data platform that combines data from disparate systems and silos in the organization: support, ecommerce transactions, in-store interactions, wish lists, reviews, and more. MongoDB's flexible document model allows for the easy combination of data of different types and formats with the ability to embed sub-documents to get a clear view of the customer in one place. As the retailer captures more data points about the customer, they can easily add fields without the need for downtime in schema change.

Then comes the ability to run analytics in real time rather than retroactively in another separate system. MongoDB's architecture allows for workload isolation, meaning the operational workload (the customer's actions on the ecommerce site) and the analytical or AI workload (calculating what the next best offer should be) can be run at the same time without interrupting the other. Then using MognoDB's aggregation framework for advanced analytical queries or triggering an AI model in real time to give an answer that can be embedded into the search ranking in real time.

Then comes the ability to easily update the search indexing to incorporate your AI augmentation. As MongoDB has Search built in, this whole flow can be completed in one data platform- as your data is being augmented with AI results, the search indexing will sync to match.

MongoDB vector search brings the next generation of search capability. By using LLMs to create vector embeddings for each product and then turning on a vector index, retailers can offer semantic search to their customers. AI will calculate the complex similarities between items in vector space and give the customer a unique set of results matched to their true desire.

READ MORE

AI-Enhanced Search in Ecommerce With MongoDB

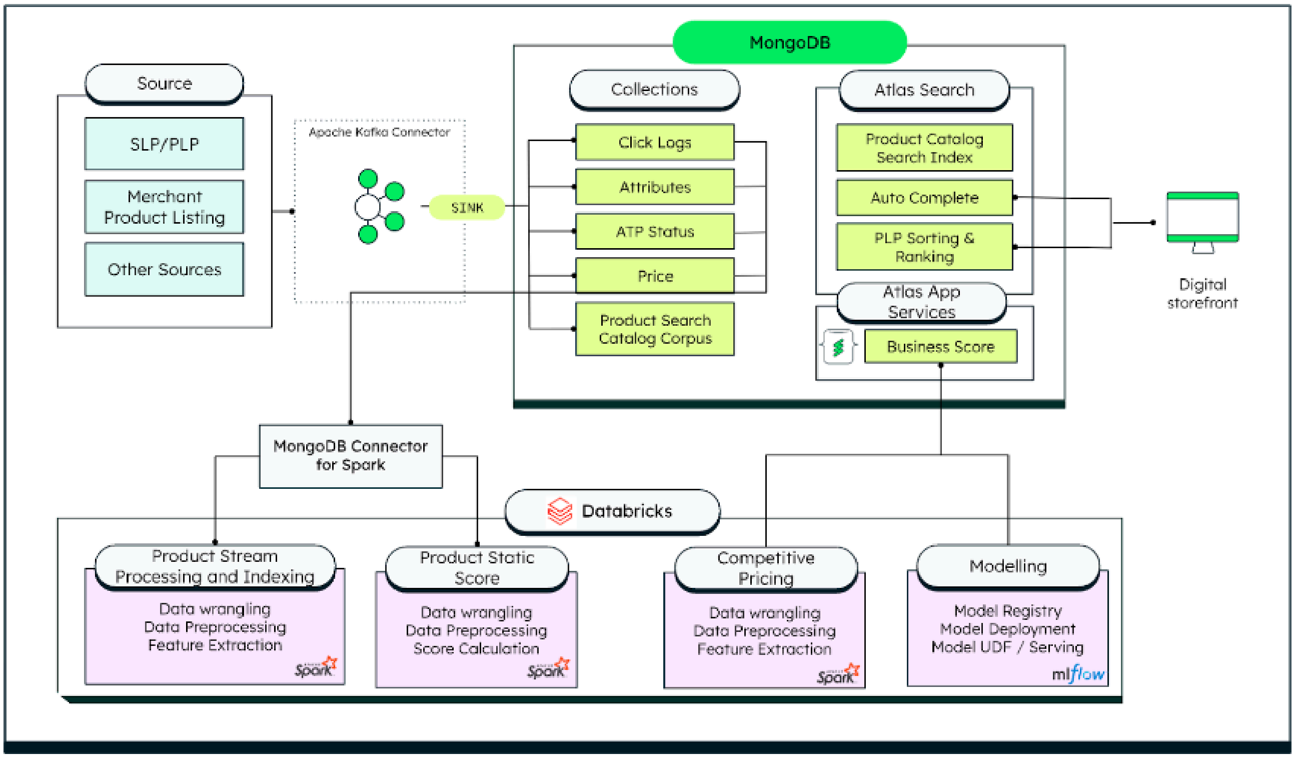

Figure 5. Architecture of an AI-enhanced search engine explaining the different MongoDB Atlas components and Databricks notebooks and workflows used for data cleaning and preparation, product scoring, dynamic pricing, and vector search

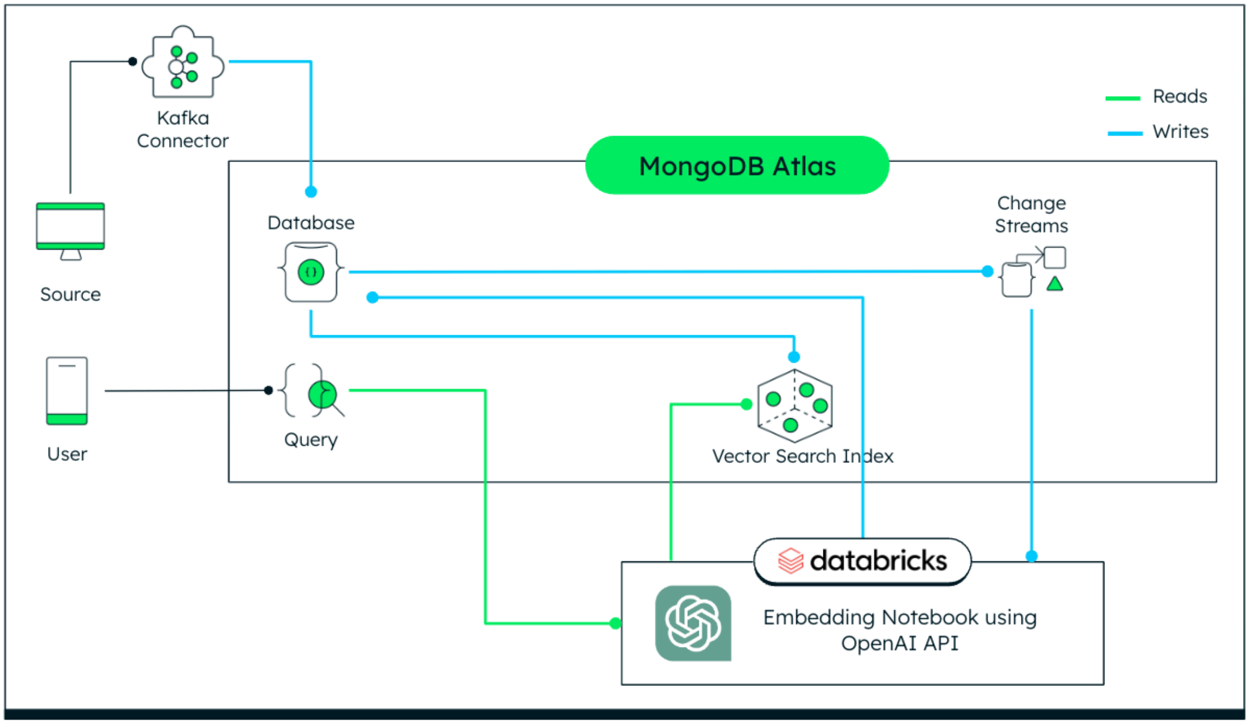

Figure 6. Architecture of a vector search solution showcasing how the data flows through the different integrated components of MongoDB Atlas and Databricks

Personalized Marketing & Content Generation

In modern retail, advertising and marketing material are vital to capturing a customer's interest and driving towards a purchase. With the advent of social media there are now many more ways to reach the customer than before: Instagram, Facebook, email outreach, newsletters, and promotional banners on sites. This creates a lucrative opportunity for retailers but also a challenge when it comes to a huge amount of content generation.

Capturing current customer buying patterns, a constantly updating product catalog and ensuring that the items being advertised are in inventory locally is part of it. The other important piece is ensuring that the content is in the right tone of voice to reflect the brand, available in multiple languages and that the pictures used reflect the audience. Traditionally, this has required a huge amount of labor in copy writing and editing, photography of different models and generation of visuals and graphics.

The retailer must also understand in real time what the impact of campaigns are so they can quickly redirect their marketing spend and strategy to reflect what is working. In an industry where marketing and branding budgets are high and the opportunity to reach customers extremely valuable if done correctly, insight is key.

GenAI has also rapidly increased retailers' ability to personalize the interactions with their customers. Retrieval Augmented Generation using Large Language Models (LLMs) are capable of creating individualized marketing material, newsletters, social posts and email outreach that is unique to each customer in seconds. Visuals, graphics and even photo-realistic images can be generated using AI to leverage the vast array of content that the retailer has- reducing manual work and speeding up time to market.

Conversational chatbots either in product recommendations or customer support also leverage GenAI to allow retailers to scale-up their ability to provide customers with personalized responses generated from internal data sets.

AI can also be used to understand quickly and easily the complex impact of campaigns, giving insights to drive intelligent strategic decisions.

The key in creating content that is personalized to the customer and the brand is in leveraging the vast amount of data that retailers have in house to provide an LLM with context. In MongoDB, the Apache Spark Connector allows for model training of LLMs, so that prompts such as "create a personalized newsletter for each customer suggesting an item based on what is on offer and their previous purchases" can use data, images, tonal or language references to create outreach.

With the MongoDB platform approach, as new items are added to the product catalog, or new images and visuals, change streams can be used to trigger the vectorization of new data so that the process becomes seamless. Keeping the model training with your internal data provides an invaluable resource to retailers in reaching their audience easily.

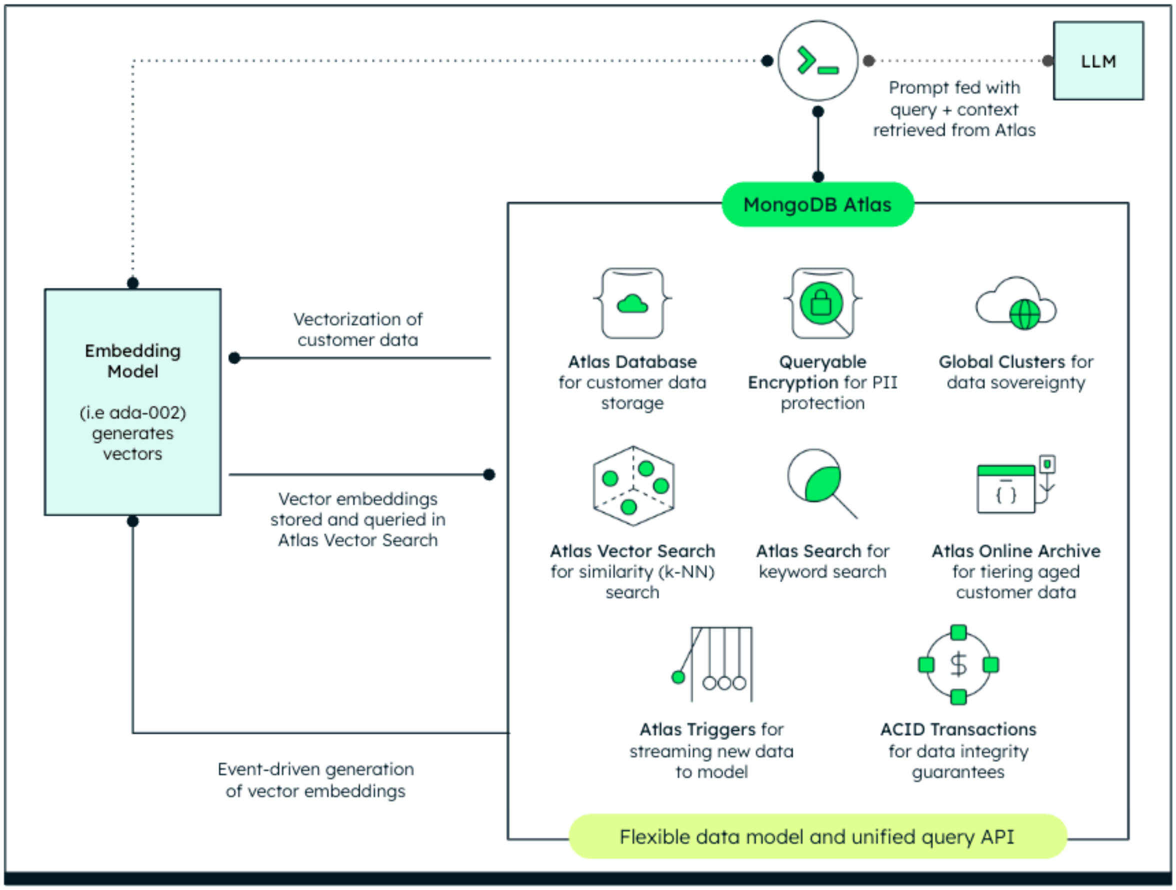
LLM
Prompt fed with query + context retrieved from Atlas
MongoDB Atlas
Embedding Model
(i.e ada-002) generates vectors
Vectorization of customer data
Vector embeddings stored and queried in Atlas Vector Search
Event-driven generation of vector embeddings
Atlas Database for customer data storage
Queryable Encryption for PII protection
Global Clusters for data sovereignty
Atlas Vector Search for similarity (k-NN) search
Atlas Search for keyword search
Atlas Online Archive for tiering aged customer data
Atlas Triggers for streaming new data to model
ACID Transactions for data integrity guarantees
Flexible data model and unified query API

Figure 7. Architecture showing how MongoDB can be used with a vector embedding model. Data in MongoDB e.g. customer data will be fed into the model (via Spark or other connector), and the generated vector embeddings will be added to each document in the collection. Then an Atlas Vector Search index can be added to the collection in MongoDB for Vector Search to be activated. An event-based architecture in MongoDB using Change Streams and Triggers can be set up so that vectors stay up to date and new additions to the database are automatically vectorized.

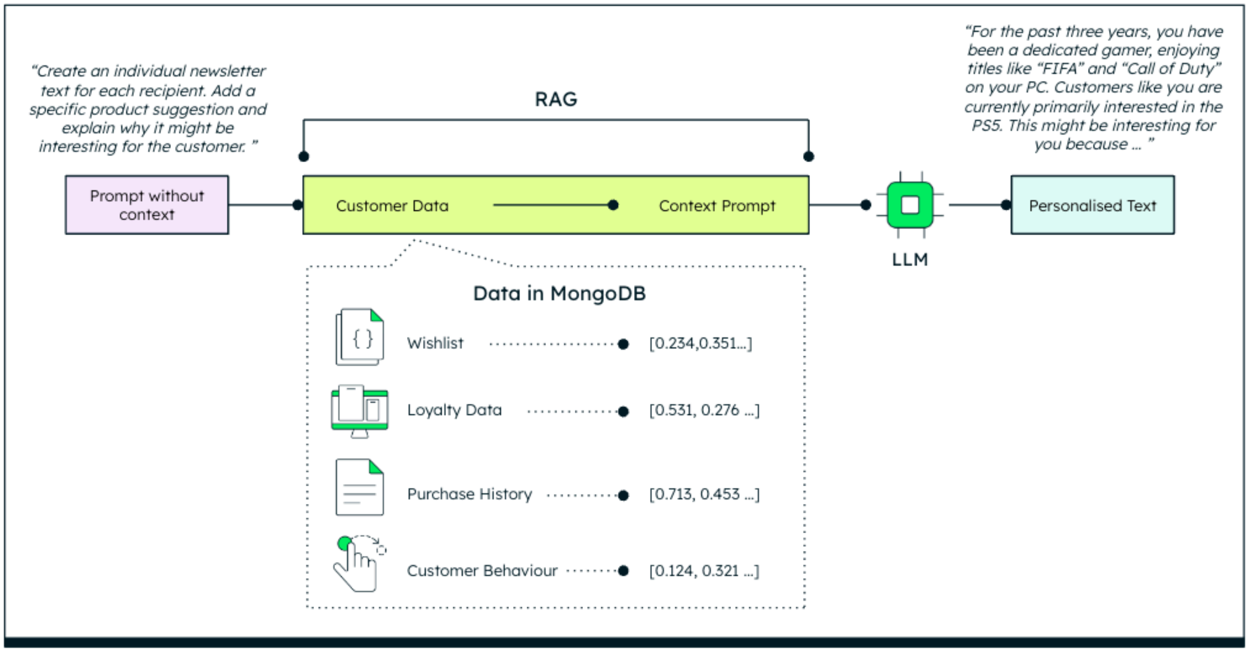
"Create an individual newsletter text for each recipient. Add a specific product suggestion and explain why it might be interesting for the customer. "
RAG
"For the past three years, you have been a dedicated gamer, enjoying titles like "FIFA" and "Call of Duty" on your PC. Customers like you are currently primarily interested in the PS5. This might be interesting for you because ... "
Prompt without context
Customer Data
Context Prompt
LLM
Personalised Text
Data in MongoDB
Wishlist
[0.234,0.351...]
Loyalty Data
[0.531, 0.276 ...]
Purchase History
[0.713, 0.453 ...]
Customer Behaviour
[0.124, 0.321 ...]

Figure 8. Example of the data flow for an AI-generated personalized newsletter. The prompt is entered by a user on the left hand side and context is added via the vectorised data in MongoDB- wishlist, loyalty data, purchase history and customer behavior. Using RAG, the LLM can produce a personalized newsletter per customer in seconds, allowing the retailer to create vast amounts of personalized content.

Demand Forecasting and Predictive Analytics

Retailers either develop homegrown applications for demand prediction using traditional machine learning models or buy specialized products designed to provide these insights across the segments for demand prediction and forecasting. The homegrown systems require significant infrastructure for data and machine learning implementation and dedicated technical expertise to develop, manage, and maintain them. More often than not, these systems require constant care to ensure the optimal performance and providing value to the businesses.

Generative AI already delivers several solutions for demand prediction for retailers by enhancing the accuracy and granularity of forecasts. The application of retrieval augmented generation utilizing large language models (LLMs) enables retailers to generate specific product demand and dig deeper to go to product category and individual store level. This not only streamlines distribution but also contributes to a more tailored fulfillment at a store level. The integration of generative AI in demand forecasting not only optimizes inventory management but also fosters a more dynamic and customer-centric approach in the retail industry.

Generative AI can be used to enhance supply chain efficiency by accurately predicting demand for products, optimizing/coordinating with production schedules, and ensuring adequate inventory levels in warehouses or distribution centers. Data requirements for such endeavors include historical sales data, customer orders, and current multichannel sales data and trends. This information can be integrated with external datasets, such as weather patterns and events that could impact demand. This data must be consolidated in an operational data layer that is cleansed for obvious reasons of avoiding wrong predictions. Subsequently, feature engineering to extract seasonality, promotions impact and general economic indicators. A retrieval augmented generation model can be incorporated to improve demand forecasting predictions and avoid hallucinations. The same datasets could be utilized from historical data to train and fine-tune the model for improved accuracy. Such efforts lead to the following business benefits:

- Precision in demand forecasting
- Optimized product / Supply planning
- Efficiency improvement
- Enhanced customer satisfaction

Challenges in Adopting AI and Machine Learning

Developing machine learning (ML) models in the traditional way can present various challenges:

- **Data acquisition and preparation:** Obtaining high-quality and diverse datasets can be challenging and time-consuming. Preparing data, including cleaning, preprocessing, and feature engineering, demands considerable effort.

- **Expertise and resources:** Building ML models requires expertise in machine learning algorithms, data science, and programming languages. Organizations may need to invest in specialized talent and computational resources.

- **Model selection and hyperparameter tuning:** Identifying the most appropriate ML model and optimizing hyperparameters require extensive experimentation, which can be a laborious trial-and-error process.

- **Overfitting and underfitting:** Balancing model complexity to avoid overfitting (when the model performs well on training data but poorly on new data) or underfitting (when the model lacks the ability to generalize to new data) is a challenging task.

- **Interpretability and explainability:** Traditional ML models, such as deep neural networks, may lack interpretability, making it challenging to understand and explain their decisions.

- **Long development cycles:** Building ML models from scratch can lead to long development cycles, delaying time-to-market for applications.

- **Resource intensiveness:** Training complex models can be computationally intensive and may require substantial hardware and time resources.

- **Limited domain expertise:** Developers may lack expertise in specific domains, making it difficult to understand the underlying data patterns and model behavior.

- **Maintenance and updates:** Regular model maintenance and updates are essential to keep the model relevant and accurate, requiring continuous effort.

- **Generalization to new data:** Traditional ML models may not generalize well to new data or changing patterns, necessitating frequent retraining and adaptation.

Addressing these challenges often demands a significant investment of time, resources, and expertise, making the development process complex and potentially expensive. However, as the field of ML continues to evolve, various tools and platforms aim to simplify and streamline the development process, offering solutions to some of these challenges.

Enterprise Principles for Responsible AI

The responsible AI principles that should be followed by the retail industry are similar to those applied across various sectors. Here are some key principles that retailers should consider:

- **Fairness and bias mitigation:** Ensure that applications employing AI do not discriminate intentionally or unintentionally against any particular group of customers based on factors like race, gender, or socioeconomic status. Implement measures to detect and address biases in data and algorithms to provide fair and equitable services. Essentially these implementations must be based on a strong data foundation.

- **Transparency and explainability:** Enterprises should make a conscious effort to make their applications employing AI (generative or otherwise) transparent and explainable to customers, regulators, and shareholders. Users should understand how AI is adopted and used to make decisions and have access to explanations for recommendations or actions taken by systems employing AI models.

- **Privacy and data protection:** Protecting customer data and privacy is one of the MOST key activities that any enterprise must not ignore. By implementing robust security measures and complying with relevant data protection laws and regulations, enterprises not only save themselves from tremendous regulatory oversight but also save themselves from reputational risk. Only collect and use customer data with explicit consent and use it (document it) responsibly.

- **Accountability and governance:** Establish clear lines of responsibility for AI adoption into application development and deployment within the organization. All enterprises including Retailers should have appropriate governance structures to oversee AI adoption and ensure adherence to responsible practices and established principles.

- **Customer empowerment and consent:** Retailers should empower customers to control their interactions with systems that employ AI and provide choices for opting in or out of personalized recommendations and other AI-driven features.

- **Safety and reliability:** Ensure that all systems and applications including those using AI are safe and reliable. Implement safeguards to prevent systems/applications from causing harm to customers or negatively impacting their experience.

- **Collaboration with experts:** Collaborate with field ethics experts, researchers, and stakeholders to incorporate diverse perspectives and address ethical challenges in AI implementation.

- **Continuous monitoring and improvement:** Continually monitor AI systems to identify and rectify any issues that may arise over time. Continuously improve AI models to enhance accuracy and fairness.

- **Social and environmental impact:** Consider the broader social and environmental impacts of AI applications in the retail industry. Strive to use AI to create positive impacts for customers and the community.

A Checklist for Responsible AI Adoption

Adopting the following responsible AI principles helps to ensure that AI used in the retail industry is developed and used in an ethical and socially beneficial manner, fostering trust and long-term relationships with customers. Any deviation, especially unintended, will cause a major impact and run a reputational risk to the enterprise.

Now let's look at a way to adopt the above-mentioned principles into enterprise software development practices. Here's a checklist of items a retailer should put together prior to engaging in adopting Generative AI into their application usage:

- ☑ Business objectives: Clearly define the business objectives and use cases for incorporating Generative AI into retail applications. Identify specific areas where Generative AI can add value, such as personalized recommendations, product design, or customer support.

- ☑ Data inventory: Conduct a thorough inventory of available data. Identify the relevant datasets, including customer data, product catalogs, transaction records, and any other data needed for training and fine-tuning Generative AI models.

- ☑ Data privacy and compliance: Ensure compliance with data privacy regulations and obtain necessary permissions from customers for data usage. Develop robust data privacy and security measures to protect customer information.

- ☑ Data quality: Assess the quality, accuracy, and completeness of the data. Clean and preprocess the data to ensure it is suitable for training and generating reliable insights.

- ☑ AI expertise: Evaluate the existing AI expertise within the organization. Determine if in-house expertise is sufficient for developing and fine-tuning Generative AI models or if external help is required.

- ☑ Model selection: Research and choose the appropriate Generative AI model for the specific use case. Consider factors like model size, performance, and compatibility with your infrastructure.

Creating, communicating and following these best practices, retail enterprises can optimize their implementation of genAI that aligns with business objectives, drives meaningful outcomes (revenues), and enhances customer experiences. Implementing genAI in a thoughtful and iterative manner empowers retailers to stay at the forefront of AI-driven innovation in the dynamic retail landscape.

Conclusion

Across the retail industry, AI has captured the imaginations of executives and consumers alike. Whether you're a customer of a grocer, ecommerce site, or retail conglomerate, AI has and will transform and enhance the way you do business with corporations. For the retailers that matter most globally, AI has created opportunities to minimize risk and fraud, perfect user experiences, and save companies from wasting labor and resources.

MongoDB Atlas will revolutionize retailers' abilities to incorporate operational, analytical, and generative AI data services. Leading companies like Bosch and Telefonica use MongoDB in their AI-enhanced IoT platforms, while Iguazio uses MongoDB as the persistence layer for its data science and MLOps platform.

From creation to launch, MongoDB Atlas guarantees that AI applications are cemented in accurate operational data and fulfill the demands of scalability, security, and performance by developers and consumers alike.

To learn more about industry-specific solutions for AI developers, visit the MongoDB Solutions Library to access reference architectures, product guides, and key tools for building your next generative AI application. If you are ready to dive in even further with our experts, schedule an Innovation Workshop with our team today.

Contact Information

Genevieve Broadhead
MongoDB Global Lead, Retail Solutions
genevieve.broadhead@mongodb.com

Prashant Juttukonda
Retail Industry Solutions Principal
prashant.juttukonda@mongodb.com

FOR MORE INFORMATION AND RESOURCES

Visit MongoDB for Retail →