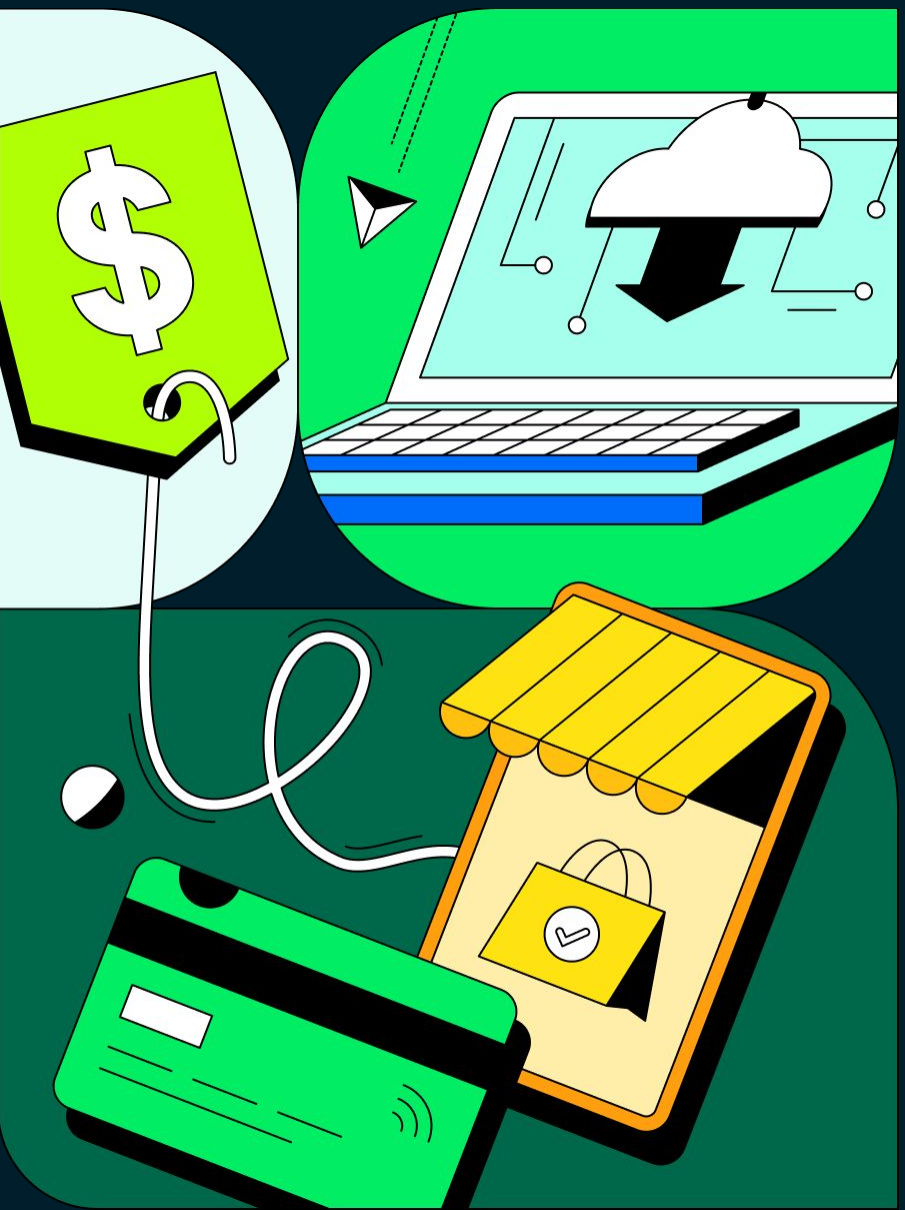
MongoDB Atlas for Industries

Enhancing Retail Operations with AI and Vector Search: The Business Case for Adoption



AI is transforming retailers' ability to maximize their competitive advantage through better understanding of their customers and improving their operating margins through intelligent decision making.

Artificial Intelligence (AI) is revolutionizing the retail industry across the globe, driving innovation and enhancing efficiency. It itself its evolving from traditional AI to generative AI.

The shift from AI to generative AI in retail reflects advancements in technology that enable more sophisticated and creative applications, improving customer experiences and operational efficiencies.

Traditional AI (Machine Learning models and arithmetic algorithms) have been used extensively in retail for a variety of functions:

- **Personalization**: AI-driven recommendation engines analyze customer data to provide personalized product suggestions.
- **Demand Forecasting**: Predictive analytics help retailers manage inventory by forecasting demand and optimizing stock levels.

Generative AI represents a leap forward by not only analyzing data but also creating new content and solutions:

- **Content Creation**: Generative AI can produce personalized marketing content, such as product descriptions, advertisements, and social media posts, tailored to specific audiences.
- **Hyper Personalization**: Beyond recommendations, generative AI can create personalized shopping experiences by dynamically generating web and mobile interface elements based on user behavior.
- **Conversational Chat**: AI-generated virtual shopping assistants can provide more natural and engaging interactions with customers, improving the overall shopping experience.

Major consulting firms have extensively documented these advancements, for example, on a 2023 survey by McKinsey about one third of all respondents say their organizations are already regularly using generative AI in at least one function (McKinsey).

MongoDB sees AI as having a transformative impact on global retail by driving innovation and enhancing customer experiences. Leveraging MongoDB Atlas and its integration with different platforms, retailers can manage massive datasets

required for generative AI applications effectively. This enables advanced data ingestion, seamless AI model training, and efficient data retrieval through features like vector search. These capabilities allow retailers to automate tasks, personalize customer interactions, and innovate with new content formats, ultimately leading to faster time-to-market and cost-effectiveness. McKinsey & Company estimates generative AI could contribute roughly $310 billion in additional value for the retail industry by boosting performance in functions such as marketing and customer interactions (McKinsey).

Generative AI has significantly impacted the retail industry, driving notable growth and transformation. According to NVIDIA's "State of AI in Retail and CPG" survey, 69% of retailers reported an increase in annual revenue due to AI adoption, and 72% experienced a decrease in operating costs. Generative AI is particularly influential in enhancing customer experiences, with 86% of retailers recognizing its potential in this area (Nvidia).

This chapter delves into these insights, showcasing real-world applications and exploring the future trajectory of AI in retail.

Retail leaders who invest more in technology to improve customer experience are 17% more likely to outperform other retailers in organic sales growth and get about 37% more sales from digital channels (Bain and Company).

With the use of AI, retailers can leverage the advantages of augmented and semantic search, generate marketing materials based on the market conditions, get the most of predictive analytics to forecast demand use conversational chatbots and enhance customer experiences.

The most creative AI use cases for retailers is to 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 amount 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

Retail is a *customer centric* business. Customers have more choice than ever in where they purchase a product. To retain and grow their customer base, retailers need to keep innovating in order to offer each customer a differentiated buying experience. To do this, it is necessary to use a large amount of data from the customers such as buying patterns, interests, and interactions and to be able to quickly make complex decisions on that data.

One of the key customer interactions in an ecommerce experience is search. Through the implementation of full-text search engines, customers can 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. Traditionally, 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, and subsequently running classical AI with various Machine Learning algorithms on such data. Typically, this would run in a batch mode (every 24, 48 or even 72 hours or a few days), and the next time a customer logs in, they will have a personalized experience. It does not, however, capture the customer's true desire in real time.

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 clickstream and/or trending purchases across shoppers.

The first step in truly understanding the customer is to build a customer operational data store 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 enables bringing data of different types and formats in one document 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 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 simultaneously without interrupting the other. Retailer can build dynamic ranking by using MongoDB 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.

The benefit of an all-in-one platform is huge here, as instead of having to update your search indexing to incorporate your AI augmentation, MongoDB has Lucene Search built in. This whole flow can be completed in one data platform automatically- 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 are able to 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.

Vector search technology in retail provides notable economic benefits, as highlighted by Deloitte: Sales Uplift and Customer Engagement: Deloitte reported that retailers implementing personalized search have seen a sales uplift of about 40% (Deloitte).

READ MORE

AI-Enhanced Search in Ecommerce With MongoDB

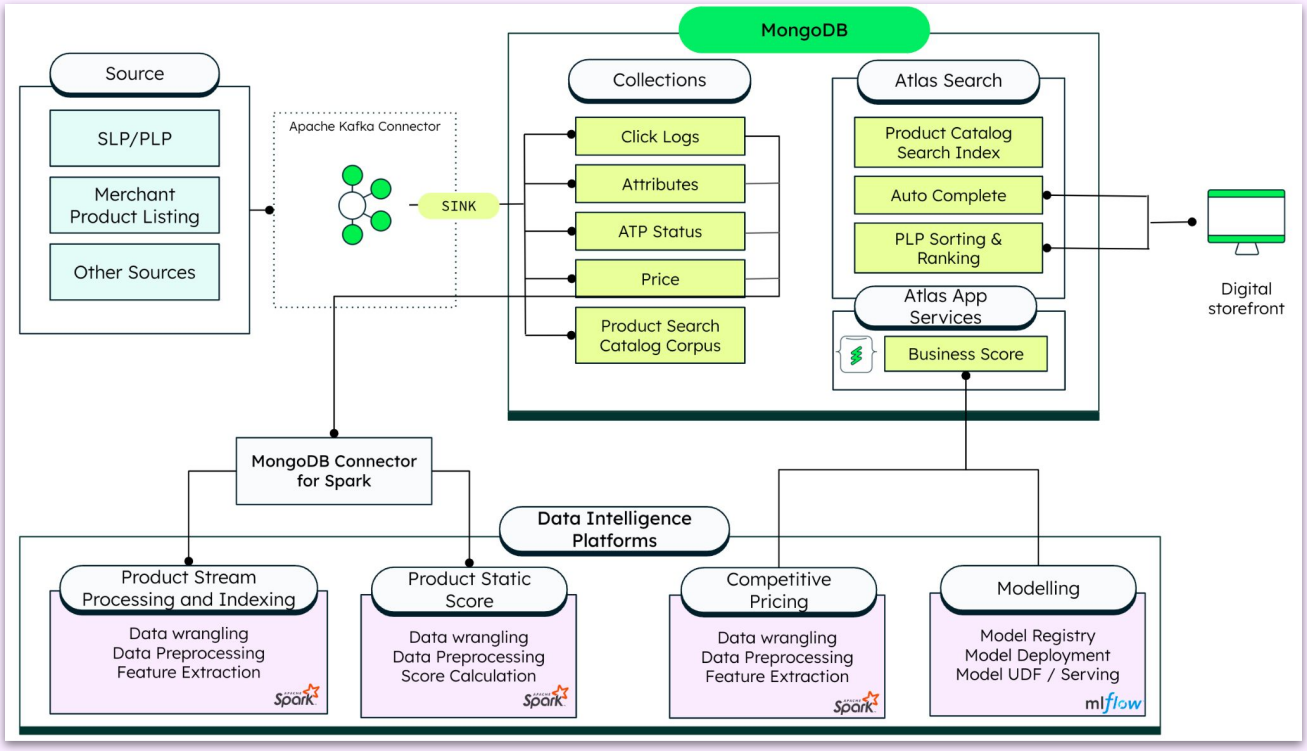


Figure 1. Architecture of an AI-enhanced search engine explaining the different MongoDB Atlas components and Data Intelligent Platforms and workflows used for data cleaning and preparation, product scoring, dynamic pricing, and vector search.

Delivery Hero

Delivery Hero Helps Customers Navigate more than 100 Million Products with MongoDB Atlas Search

Delivery Hero, a food delivery service based out of Germany, has built a new Item Replacement Tool providing hyper-personalized product recommendations in real time using state-of-the-art AI models and MongoDB Atlas Vector Search.

The challenge was that around 10% of the inventory is perishable produce that can quickly go out of stock. Without being able to recommend a suitable alternative to the customer, the company risks revenue loss and customer churn.

The solution was MongoDB a scalable, high-performing database platform enhanced by AI. With it the new Item Replacement Tool is being piloted first in the Middle East. By providing personalized recommendations against live inventory, Delivery Hero expects to see an increase in its monthly Gross Merchandise Value.

"With MongoDB Atlas' distributed architecture, we can cost-effectively scale-out the service across our global operations."

Learn more

Personalized Marketing & Content Generation

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.

Customer buying patterns, constantly updating product catalog and inventory availability are critical components of Retail operations. Along with this ensuring that the product literature is in the right tone of voice to reflect the brand in multiple languages. The product images should be relevant to the audience in the locale. Traditionally, this required a huge amount of labor in copywriting 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 is 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 is extremely valuable provided Retailer has the right insight.

Companies will take advantage of the sharp rise in consumer touchpoints to personalize and reach the growing population of consumers who use digital channels to discover, consider, and purchase products. 65% of consumers research products online, and 30% buy online. These numbers have doubled over the past 3 to 4 years. This creates an enormous need for brands to target online consumers with personalized content—an opportunity enabled by generative AI's lower content creation costs (Bain and Company).

GenAI has also rapidly increased retailers' ability to personalize the interactions with their customers. Retrieval Augmented Generation using Large Language Models (LLMs) is 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 amount of data the retailer already has. This reduces manual work and speeding up time to market.

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

The key to creating content that is personalized to the customer and the brand is 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

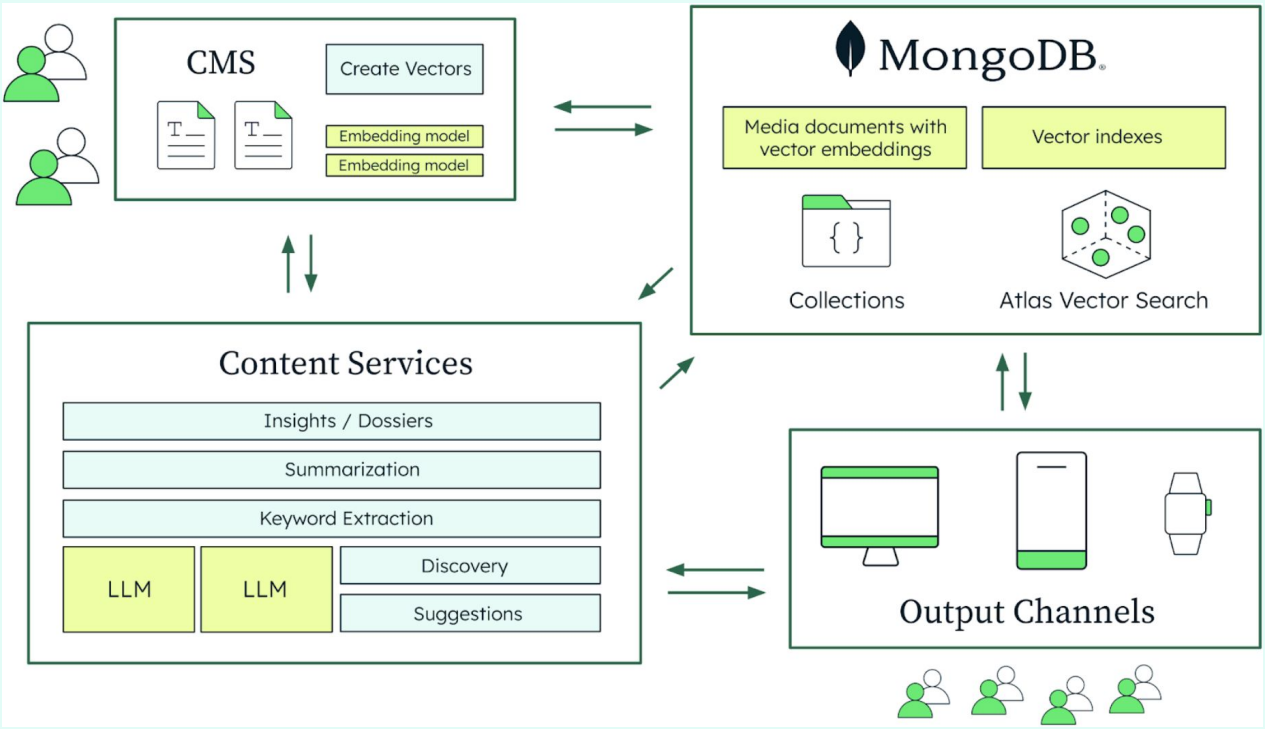


Figure 2. The above image shows a reference architecture highlighting where MongoDB can be leveraged to achieve AI-powered personalization. By utilizing user data and the multi-dimensional vectorization of media content, MongoDB Atlas can be applied to multiple AI use cases. This allows utilization of media channels effectively and improve end-user experiences. By doing so, media organizations can suggest content that aligns more closely with individual preferences and past interactions. This not only enhances user engagement but also increases the likelihood of converting free users into paying subscribers.

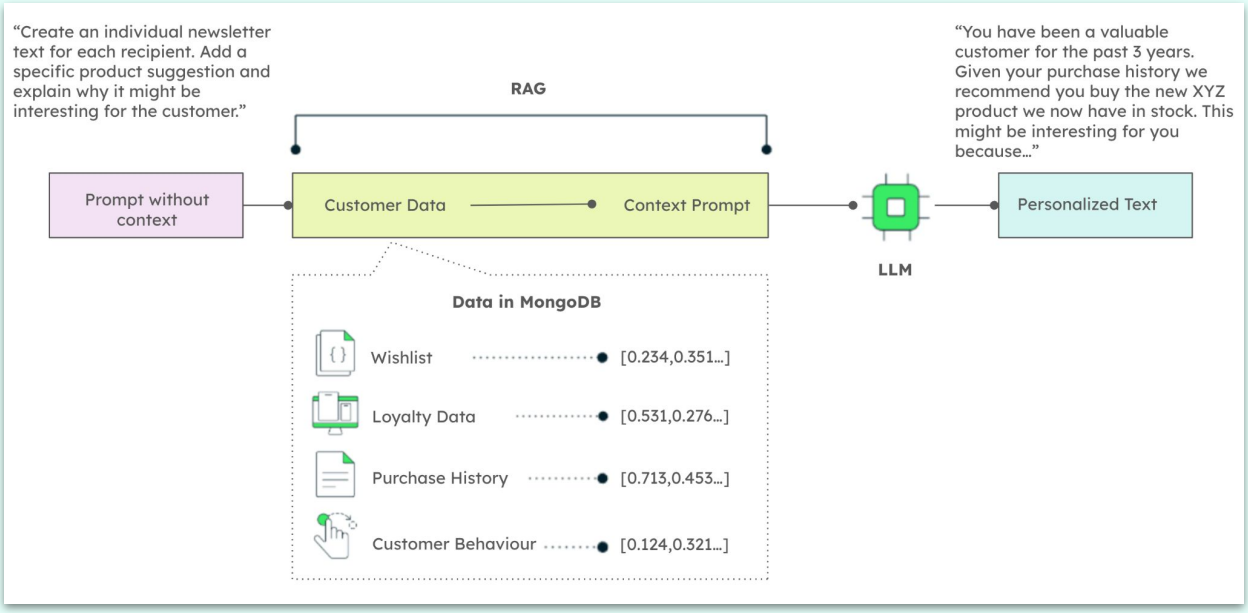


Figure 3. Example of the data flow for an AI-generated personalized newsletter. The prompt is entered by a user on the left hand side and context is added via the vectorised data in MongoDB- wishlist, loyalty data, purchase history, and customer behavior. Using RAG, the LLM can produce a personalized newsletter per customer in seconds, allowing the retailer to create vast amounts of personalized content.

Demand Forecasting & Predictive Analytics

Accurate demand planning using AI in retail optimizes inventory levels, reducing costs and stockouts, while enhancing customer satisfaction through better availability of products. It also enables data-driven decisions, leading to improved sales forecasts and efficient supply chain management. 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 optimal performance and provide value to the businesses.

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 reduce possibility of 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
- Accuracy in inventory management
- Enhanced customer satisfaction

Traditional AI is used in demand forecasting and predictive analytics in retail by integrating data from diverse sources like sales transactions, social media, and weather patterns, leading to highly accurate and timely forecasts. Machine learning algorithms continuously learn and adapt from new data, improving predictive accuracy, while automation reduces the time and resources needed for these tasks, allowing for efficient scaling of forecasting efforts.

Generative AI is transforming demand forecasting and predictive analytics by data to find patterns from existing datasets, enhancing the accuracy and depth of predictions. By creating synthetic data, generative AI models can fill in gaps in historical data, simulate various market scenarios, and predict future trends more effectively. This leads to more precise demand forecasts, allowing retailers to optimize inventory levels, reduce stockouts, and avoid overstock situations, thus improving operational efficiency and customer satisfaction.

Well implemented demand forecasting can lead to a 3-7% increase in group operating profit, 30% reduced time to market 15-20% increase in store ordering daily, increasing product freshness (Bain and Company).

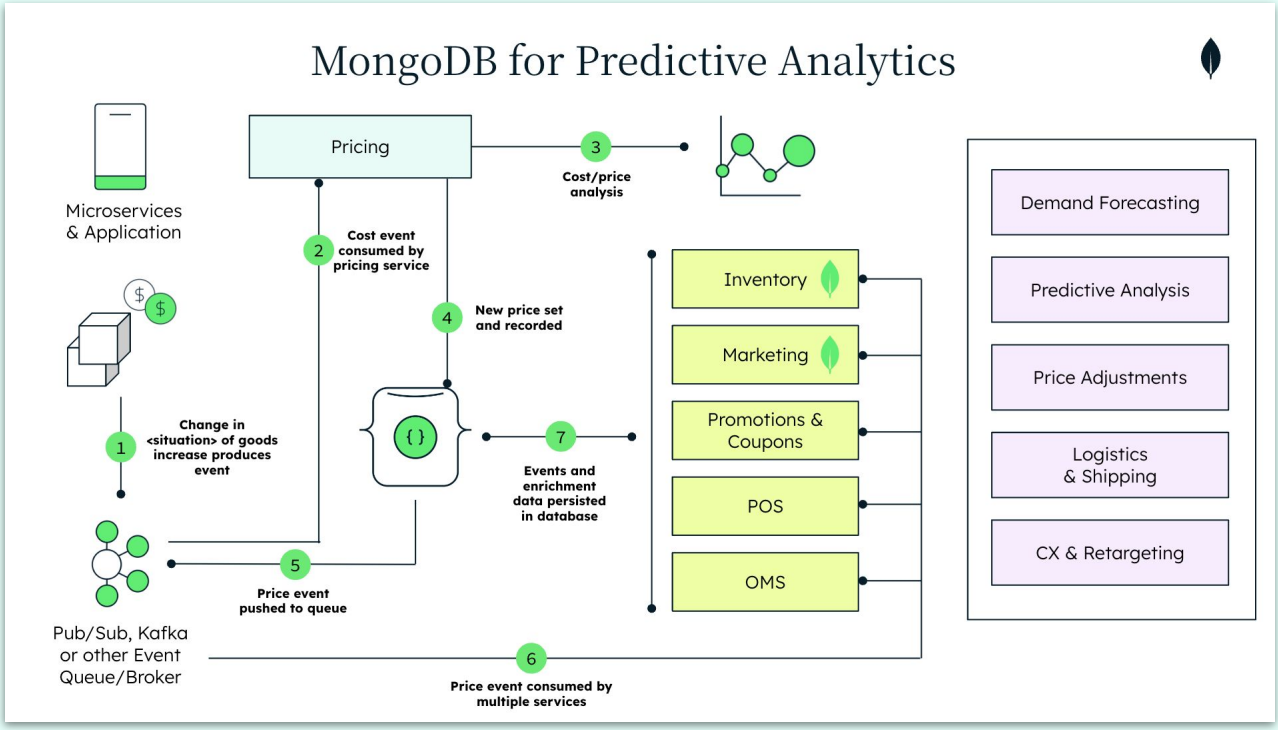


Figure 4. This figure illustrates a price change scenario. Fuel costs to ship some items have gone up, which impact pricing: [1] This produces events about the cost increase and places them in the message stream where the event queue makes them available. All microservices are listening for messages of interest. [2-3-4] Pricing microservice consumes the event, analyzes it against existing data, and produces events conveying the new pricing into the message stream. [5 - 6] The database pushes those messages to the event queue, which makes them available to all consumers who are listening for messages of interest. Microservices directly impacted by pricing changes — such as those that manage inventory, marketing, promotions & coupons, point of sale (POS), and the e-commerce provider's order management system (OMS) — consume the price change events and update their individual databases accordingly. [7] The centralized database aggregates and persists events, enriches event streams with data from other sources, including historical data, and provides a central repository for multiple event streams.

Conversational Chatbots

Conversational chatbots powered by Generative AI are revolutionizing the retail industry by enhancing customer service. These chatbots can handle a wide range of customer inquiries, from product recommendations to order tracking, providing instant and accurate responses. This reduces wait times and improves the overall customer experience, leading to higher satisfaction and increased loyalty. Additionally, chatbots can operate on real time data 24/7, ensuring customers receive support at any time, which is especially beneficial for global retailers.

Beyond customer service, AI chatbots are also transforming marketing and sales strategies in retail. They can analyze customer data to personalize shopping experiences, offering tailored recommendations and promotions based on individual preferences and behavior. This personalization helps retailers boost conversion rates and increase sales. Moreover, chatbots can engage customers through various digital channels, including social media, websites, and messaging apps, broadening the reach and effectiveness of marketing campaigns.

Operational efficiency is another area where AI chatbots are making a significant impact. By automating routine tasks such as answering FAQs, managing inventory inquiries, and processing returns, chatbots free up human employees to focus on more complex and value-added activities. This not only reduces operational costs but also improves accuracy and consistency in service delivery. Furthermore, the data collected by chatbots can provide valuable insights into customer preferences and behavior, helping retailers refine their strategies and improve their offerings.

Across both savvy and non-savvy digital users, 50%-60% have shown high preference to move to conversational journeys for day-to-day use cases across verticals (Bain and Company).

Following is a chatbot RAG architecture example. This chatbot is built using the retrieval augmented generation (RAG) architecture. RAG augments the knowledge of large language models (LLMs) by retrieving relevant information for users' queries and using that information in the LLM-generated response. MongoDB's public documentation is used as the information source for chatbot generated answers.

To retrieve relevant information based on user queries, MongoDB Atlas Vector Search is utilized. In this example Azure OpenAI ChatGPT API to generate answers in response to user questions based on the information returned from Atlas Vector Search. Azure OpenAI embeddings API are used to convert MongoDB documentation and user queries into vector embeddings, to help find the most relevant content for queries using Atlas Vector Search.

Here's a high-level diagram of the chatbot's RAG architecture.

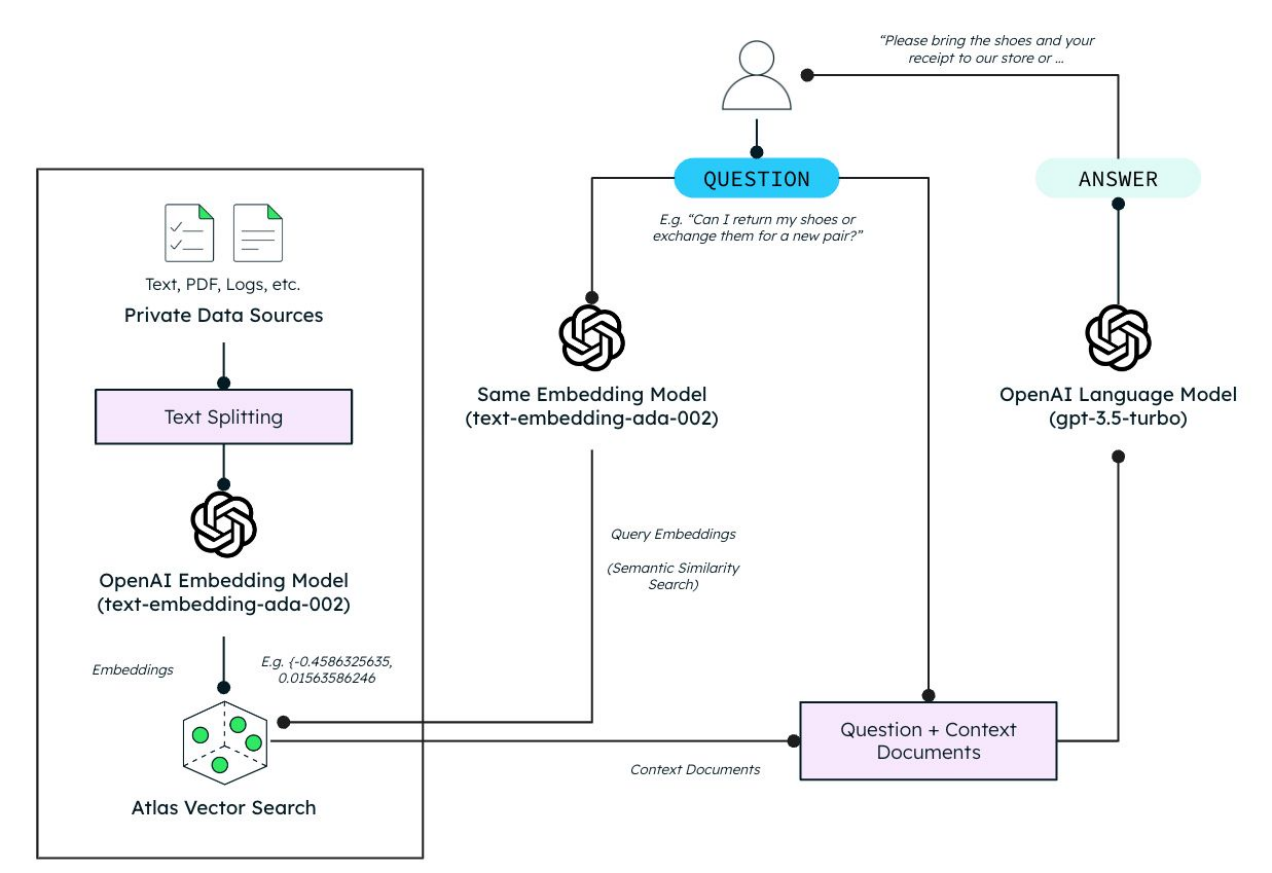


Figure 5. Example of the data flow for a chatbot RAG architecture.

L'ORÉAL GROUPE

L'Oréal Improves App Performance and Velocity with MongoDB Atlas

L'Oréal, the world leader in beauty, championing the 'Beauty Tech'.

The challenge was to complete complex calculations on vast volumes of data—without causing latency. The solution was simplifying management and maintenance while boosting performance with MongoDB Atlas.

The result was reducing latency from seconds to just 10 milliseconds.

"MongoDB Atlas doesn't just solve our performance issues It makes life easier. We have a hyper agile DevOps model."

Learn more

In conclusion, artificial intelligence is revolutionizing the way retailers enhance their competitive edge by providing deeper insights into customer behavior and optimizing profit margins through smart decision-making processes. By incorporating both Traditional and Generative AI, retailers can harness the benefits of enhanced and semantic search capabilities, create targeted marketing content based on current market trends, effectively utilize predictive analytics for demand forecasting, employ conversational chatbots, and significantly elevate the overall customer experience.

Contact Information

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

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

Rodrigo Leal
Retail Industry
Solutions Principal
rodrigo.leal@mongodb.com

FOR MORE INFORMATION AND RESOURCES

Visit MongoDB for Retail