How Leading Industries are Transforming with AI and MongoDB Atlas

April 2024

Table of Contents

Editor's Introduction

I am delighted to present an insightful exploration into the dynamic intersection of artificial intelligence (AI), innovation, and industry solutions. This eBook serves as a beacon, guiding readers through the intricate landscape of AI solutions across not only the industry your organization sits in, but also provides insights into how other industries are innovating with AI. Along the way, we will explore the top use cases across the six core industries that are infused with MongoDB Atlas AI capabilities.

Why read it, you ask? Because within these pages lie invaluable insights into the critical role of AI. Understanding its significance and harnessing its power is paramount for businesses striving for success. You can also delve into our partner section highlighting organizations that have built AI solutions using MongoDB. Whether a SaaS end-to-end solution you can implement, or component-based solution you can plug in, there is something there for you.

Take advantage of our **innovation workshops** available to you where you can meet MongoDB industry experts and discuss the art of the possible.

Boris Bialek: Field CTO, Industry at MongoDB

AI and the Developer Data Platform

AI is quickly becoming a universal tool that fits in every industry'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, 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.

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 industries toward an AI-powered future. 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 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, documents afford the flexibility to accommodate diverse data types and application features, facilitating data rationalization and utilization in ways previously unattainable.

Rapid Querying

LEARN MORE

Real-Time Analytics and Dynamic Pricing

Paired with the document model, the MongoDB Query API provides developers with a unified and consistent approach to data 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 & Dynamic Pricing

Across retail, manufacturing, telecommunications, and insurance industries, 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 industries 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.

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.

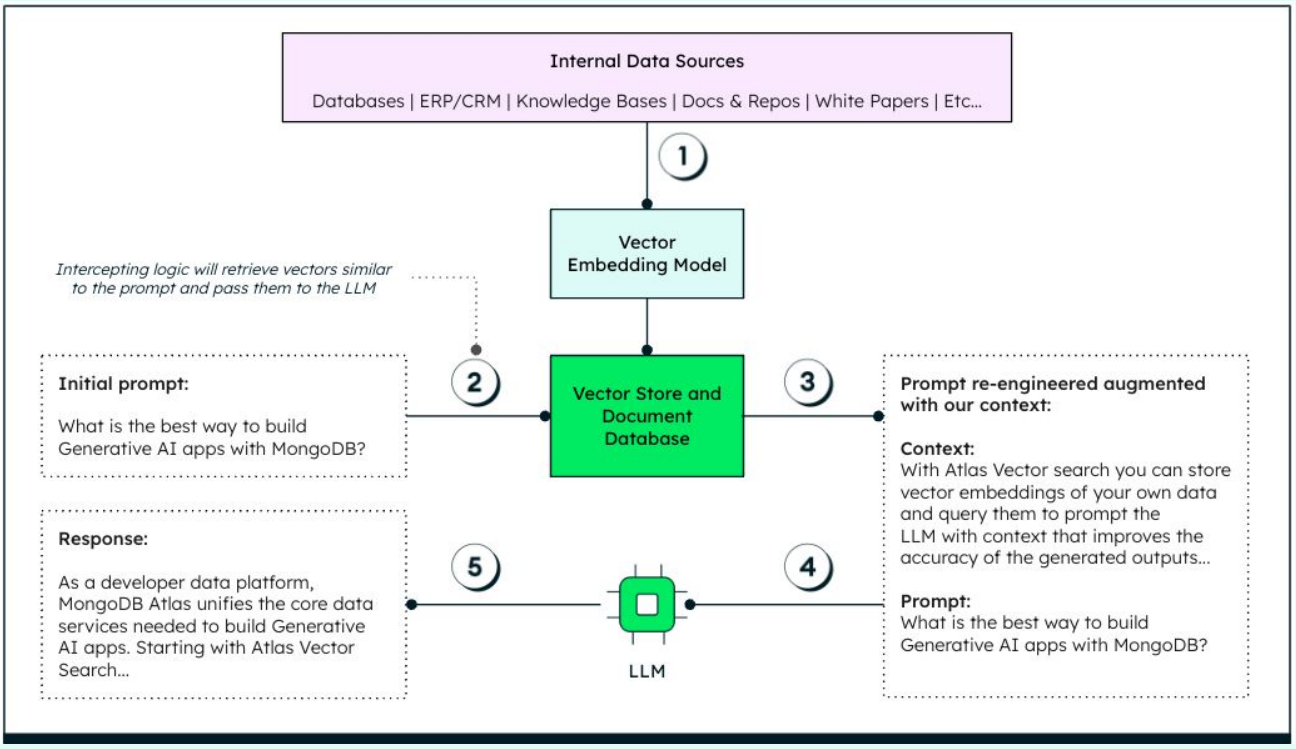


Figure 1. Data is transformed from unstructured internal sources to actionable, impactful insights.

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).

Now, let's take a closer look at how AI is transforming businesses across industries.

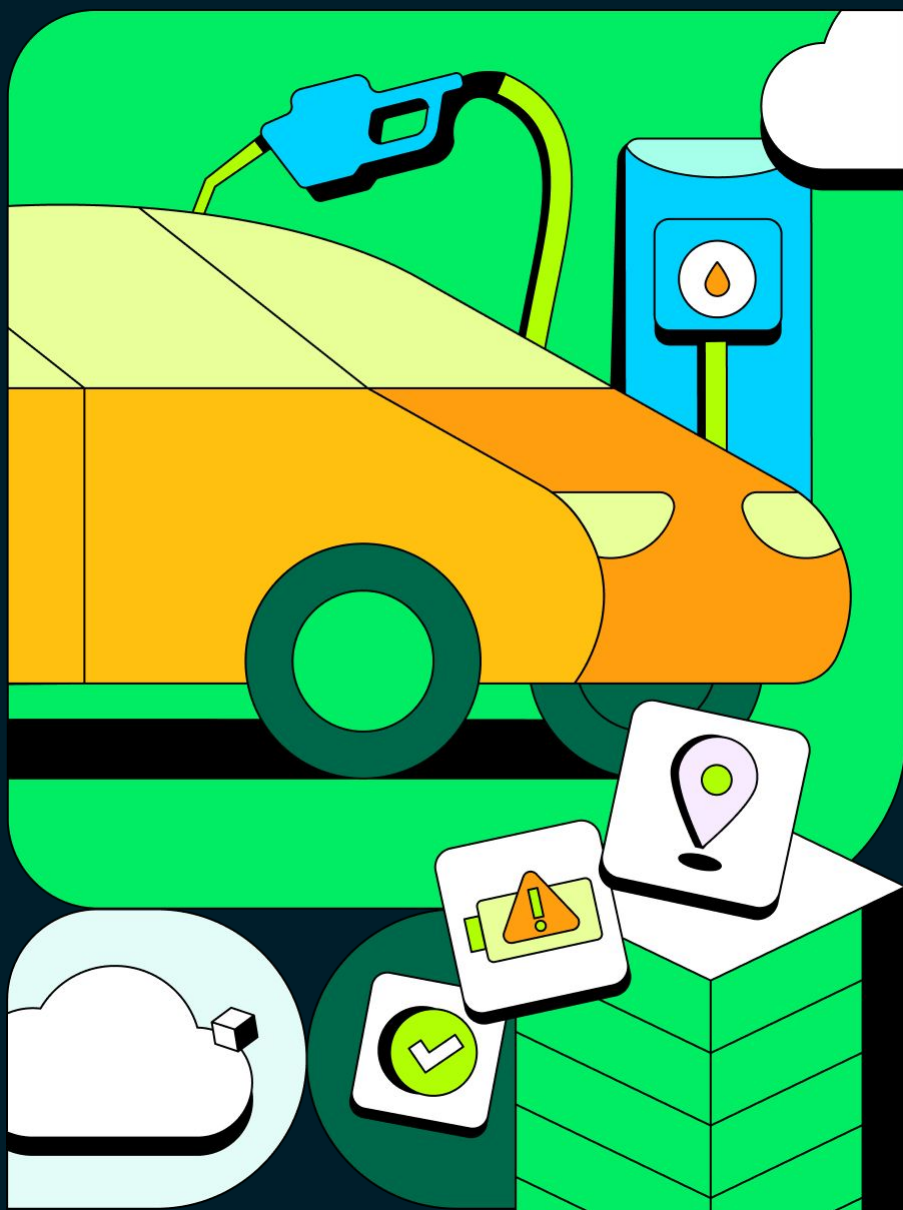
LEARN MORE

AI-Augmented Search in Ecommerce

Atlas for Industries

Manufacturing and Motion

The integration of AI within the manufacturing and automotive industry has transformed the conventional value chain, presenting a spectrum of opportunities.



Leveraging Industrial IoT, companies now collect extensive data from assets, paving the way for analytical insights and unlocking novel AI use cases, including enhanced inventory management and predictive maintenance.

Inventory Management

Efficient supply chains are able to control operational costs and ensure on-time delivery to their customers. Inventory optimization and management is a key component in achieving these goals. Managing and optimizing inventory levels, planning for fluctuations in demand, and of course, cutting costs are all imperative goals.

However, efficient inventory management for manufacturers presents complex data challenges too, primarily in forecasting demand accurately and optimizing stock levels. This is where AI can alleviate some of the pain.

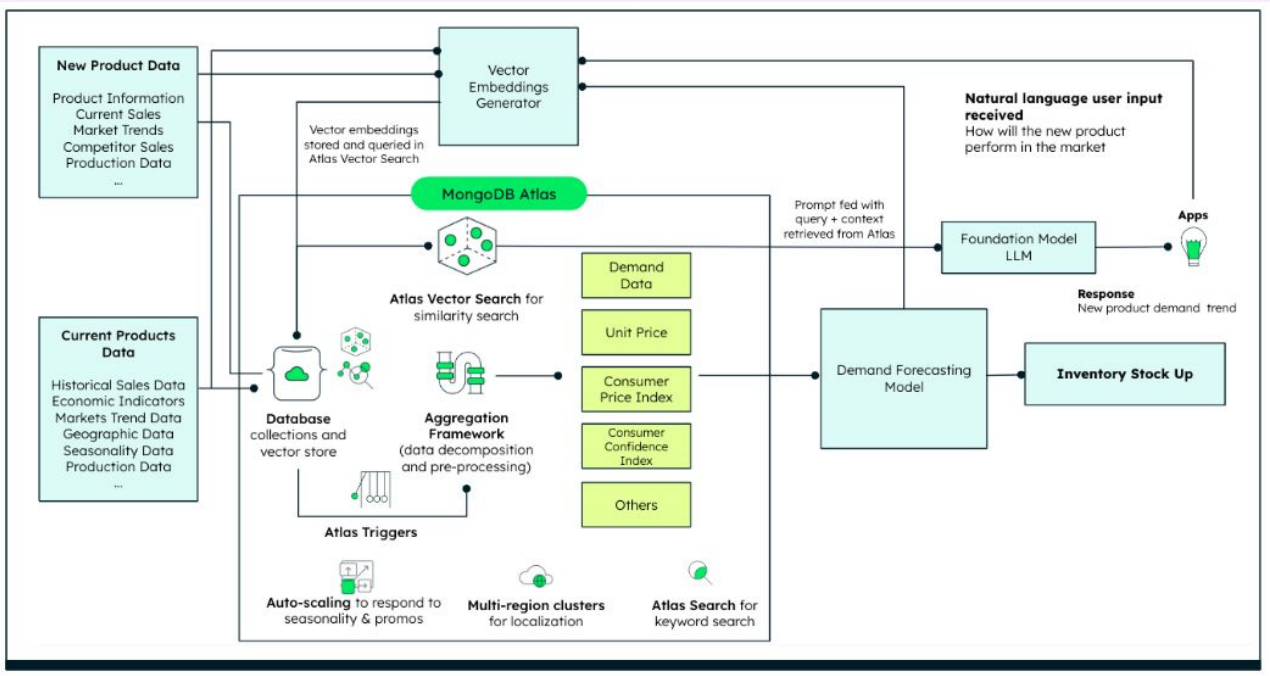


Figure 2. Gen AI-enabled demand forecasting with MongoDB Atlas

AI algorithms can be used to analyze complex datasets to predict future demand for products or parts. Improvement in demand forecasting accuracy is crucial for maintaining optimal inventory levels. AI-based time series forecasting can assist in adapting to rapid changes in customer demand. Once the demand is known, AI can play a pivotal role in stock optimization.

By analyzing historical sales data, and market trends, manufacturers can determine the most efficient stock levels and even reduce human error. On top of all this existing potential, Generative AI can help with generating synthetic inventory data and seasonally adjusted demand patterns. It can also help with creating scenarios to simulate supply chain disruptions.

MongoDB makes this process simple. At the warehouse, the inventory can be scanned using a mobile device. This data is persisted in Atlas Device SDK and synced with Atlas using Device Sync. Atlas Device Sync provides an offline-first seamless mobile experience for inventory tracking, making sure that inventory data is always accurate in Atlas. Once data is in Atlas, it can serve as the central repository for all inventory-related data. This repository becomes the source of data for inventory management AI applications, eliminating data silos and improving visibility into overall inventory levels and movements. Using Atlas Vector Search and Generative AI, manufacturers can easily categorize products based on their seasonal attributes, cluster products with similar seasonal demand patterns and provide context to the foundation model to improve the accuracy of synthetic inventory data generation.

Predictive Maintenance

The most basic approach to maintenance today is reactive – assets are deliberately allowed to operate until failures actually occur. The assets are maintained as needed, making it challenging to anticipate what's needed for repairs. Preventive maintenance, however, allows systems or components to be replaced based on a conservative schedule to prevent commonly occurring failures. Although predictive maintenance is expensive to implement due to frequent replacement of parts before end-of-life.

AI offers a chance to efficiently implement predictive maintenance using data collected from IoT sensors on machinery trained to detect anomalies. ML/AI algorithms like regression models or decision trees are trained on the preprocessed data, deployed on-site for inference, and continuously analyze sensor data. When anomalies are detected, alerts are generated to notify maintenance personnel, enabling proactive planning and execution of maintenance actions to minimize downtime and optimize equipment reliability and performance. A Retrieval augmented generation (RAG) architecture can be deployed to generate or curate the data preprocessor removing the need for specialized data science knowledge. The domain expert can provide the right prompts for the LLM. Once the maintenance alert is generated by an AI model, Gen AI can come in again to generate a repair strategy, taking spare parts inventory data, maintenance budget and personal availability into consideration. Finally, the repair manuals can be vectorized and used to power a chatbot application that guides the technician in performing the actual repair.

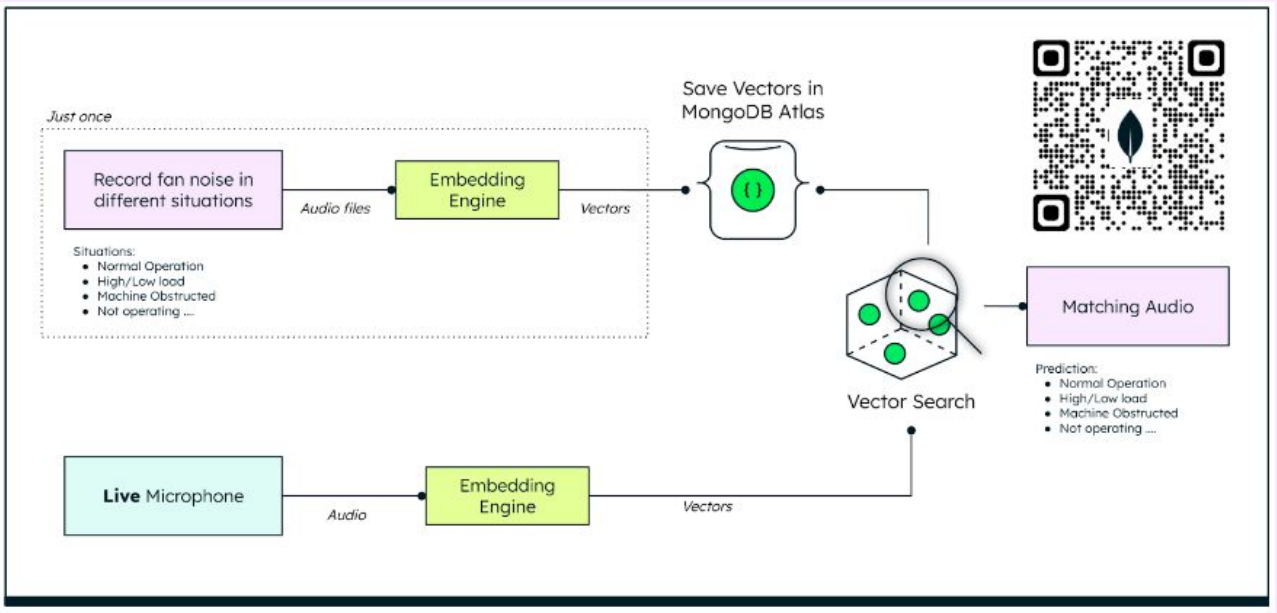


Figure 3. Audio-based anomaly detection with MongoDB Atlas.
Scan the QR code to try it out yourself!

MongoDB documents are inherently flexible while allowing data governance when required. Since machine health prediction models require not just sensor data but also maintenance history and inventory data, the document model is a perfect fit to model such disparate data sources. During the maintenance and support process of a physical product, information such as product information, replacement parts documentation, etc., needs to be available and easily accessible to support staff. Full-text search capabilities provided by Atlas can be integrated with the support portal and help staff retrieve information from Atlas clusters with ease. Atlas Vector Search is a foundational element for effective and efficient -powered predictive maintenance models. Manufacturers can use MongoDB Atlas to explore ways of simplifying machine diagnostics. Audio files can be recorded from machines which can then be vectorized and searched to retrieve similar cases. Once the cause is identified, they can leverage RAG to implement a chatbot interface that the technician can interact with and get context-aware step-by-step guidance on how to perform the repair.

Autonomous Driving

With the rise of connected vehicles, automotive manufacturers are compelled to transform their business models into software-first organizations. The data generated by connected vehicles is used to create better driver assistance systems and paves the way for autonomous driving applications.

However, it is challenging to create fully autonomous vehicles that can drive safer than humans. Some experts estimate that the technology to achieve level 5 autonomy is about 80% developed but the last 20% will be extremely hard to achieve and will take a lot of time to perfect.

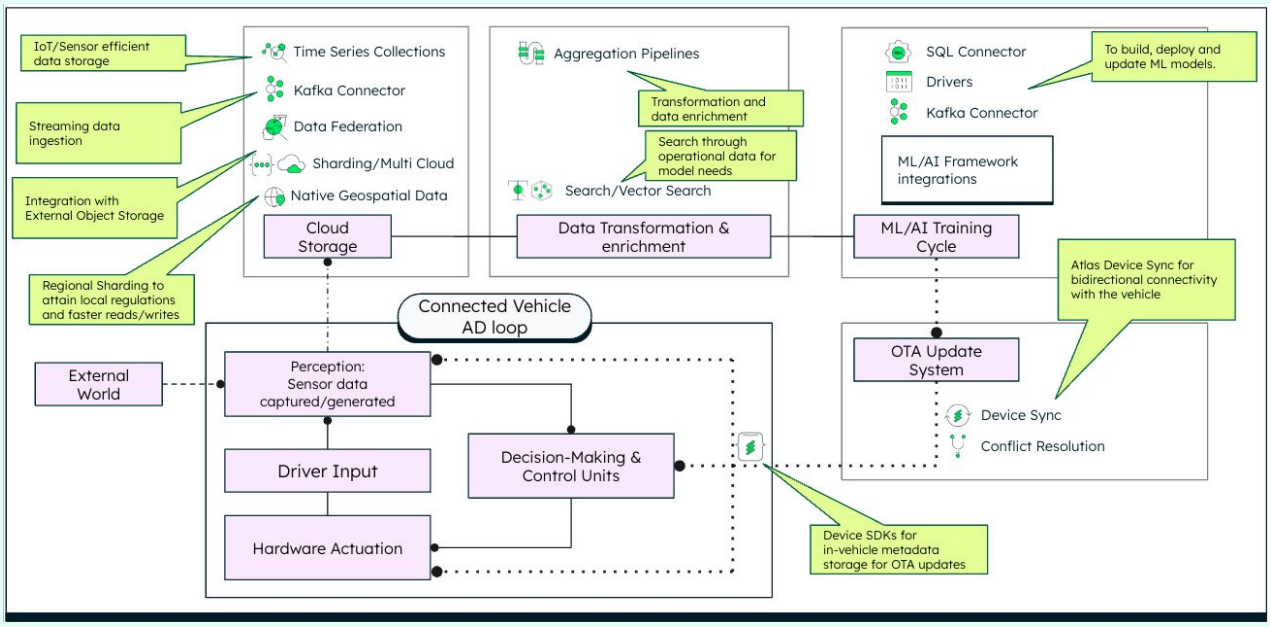


Figure 4. MongoDB Atlas's Role in Autonomous Driving

AI-based image and object recognition in automotive applications face uncertainties, but manufacturers must utilize data from radar, LiDAR, cameras, and vehicle telemetry to improve AI model training. Modern vehicles act as data powerhouses, constantly gathering and processing information from onboard sensors and cameras, generating significant Big Data. Robust storage and analysis capabilities are essential to manage this data, while real-time analysis is crucial for making instantaneous decisions to ensure safe navigation.

MongoDB can play a significant role in addressing these challenges. The document model is an excellent way to accommodate diverse data types such as sensor readings, telematics, maps, and model results. New fields to the documents can be added at run time, enabling the developers to easily add context to the raw telemetry data.

MongoDB's ability to handle large volumes of unstructured data makes it suitable for the constant influx of vehicle-generated information. Atlas Search provides a performant search engine to allow data scientists to iterate their perception AI models. Finally, Atlas Device Sync can be used to send configuration updates to the vehicle's advanced driving assistance system.

Other Notable Use Cases

AI plays a critical role in fulfilling the promise of Industry 4.0. There are numerous other use cases of AI that can be enabled by MongoDB Atlas. Some of them are listed below.

Logistics Optimization

AI can help optimize routes resulting in reduced delays, and enhanced efficiency in day-to-day delivery operations.

Quality Control and Defect Detection

Computer or machine vision can be used to identify irregularities in the products as they are manufactured. This ensures that product standards are met with precision.

Production Optimization

By analyzing time series data from sensors installed on production lines, waste can be identified and reduced, thereby improving throughput and efficiency.

Smart After Sales Support

Manufacturers can utilize AI-driven chatbots and predictive analytics to offer proactive maintenance, troubleshooting, and personalized assistance to customers.

Personalized Product Recommendations

AI can be used to analyze user behavior and preferences to deliver personalized product recommendations via a mobile or a web app, enhancing customer satisfaction and driving sales.

FOR MORE INFORMATION AND RESOURCES

Visit MongoDB Atlas for Manufacturing and Motion

Contact Information

Dr. Humza Akhtar

Manufacturing & Motion
Industry Solutions Principal

humza.akhtar@mongodb.com

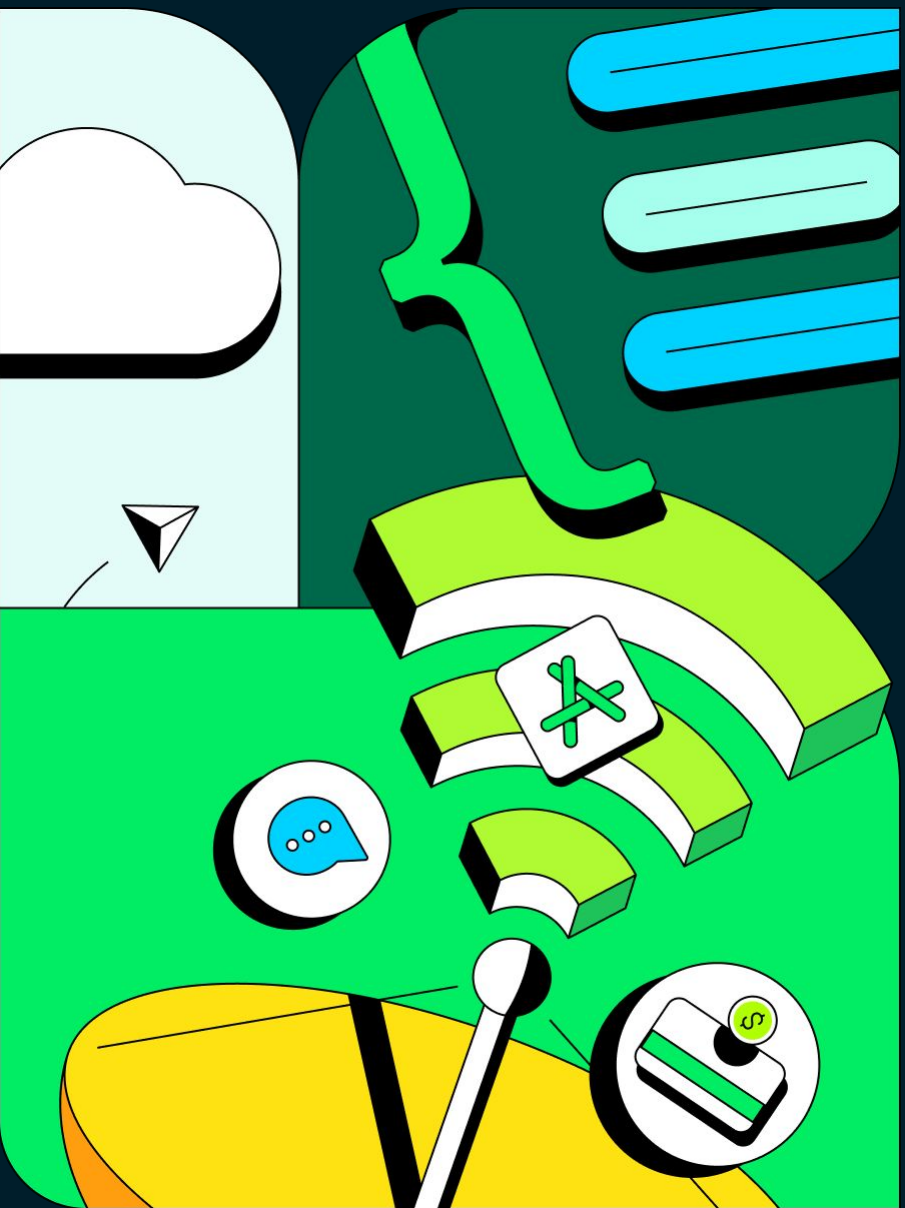
Raphael Schor

Manufacturing & Motion
Industry Solutions Principal

raphael.schor@mongodb.com

Atlas for Industries

Telecommunications and Media



Faced with high operational costs and low margins, the telecommunications and media industries are exploring new ways to create value and enhance revenue streams with AI.

The telecommunications industry operates in a landscape characterized by tight profit margins, particularly in commoditized communication and connectivity services where differentiation is minimal.

With offerings such as voice, data, and internet access being largely homogeneous, telecom companies need to differentiate and diversify revenue streams to create value and stand out in the market.

As digital natives disrupt traditional business models with agile and innovative approaches, established companies are not only competing among themselves but also with newcomers to deliver enhanced customer experiences and adapt to evolving consumer demands.

To thrive in an environment where advanced connectivity is increasingly expected, telecom operators must prioritize cost efficiency in their Operations Support Systems (OSS) and Business Support Systems (BSS), elevate customer service standards, and enhance overall customer experiences to secure market share and gain a competitive edge. They're not alone — media publishers, too, must streamline operations through automation while strengthening reader relationships to foster a willingness to pay for personalized and relevant content.

Service Assurance

Telecommunications providers need to deliver network services at optimal quality and performance levels to meet customer expectations and service level agreements. Key aspects of service assurance include performance monitoring, quality of service (QoS) management, and predictive analytics to anticipate potential service degradation or network failures before they occur. With the increasing complexity of telecommunications networks and the growing expectations of customers for high-quality, always-on services, a new bar has been set for service assurance, requiring companies to invest heavily in solutions that can automate and optimize these processes and maintain a competitive edge.

Service assurance is revolutionized by AI through several key capabilities: ML can be the powerful foundation for predictive maintenance, analyzing patterns and predicting network failures before they occur, allowing for preemptive maintenance and significantly reducing downtime. AI techniques can also sift through complex network systems to accurately identify the root causes of issues, improving the effectiveness of troubleshooting efforts. Also, with network optimization, analyzing log data to identify opportunities for improvement, raising efficiency and thus reducing operational costs and optimizing network performance in real-time.

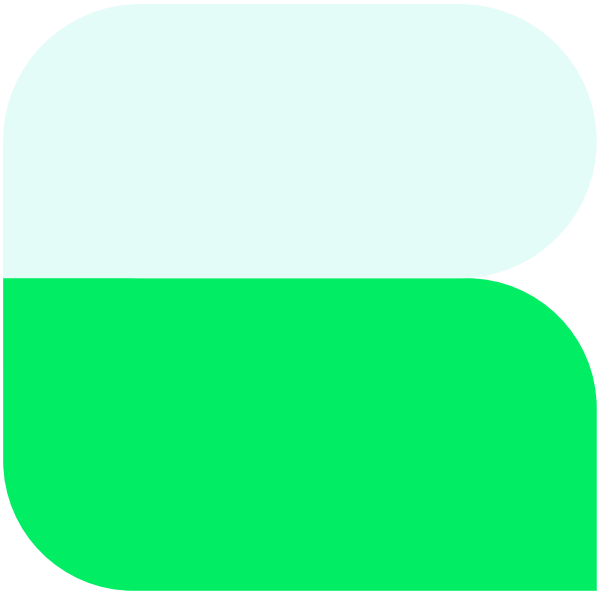
MongoDB Atlas's JSON-based document model is the ideal data foundation to underpin intelligent applications. Store log data from various systems without the need for time-intensive upfront data normalization efforts and with the flexibility to deal with a wide variety of different data structures, as well as with their potential change over time.

By vectorizing the data with an appropriate ML model, it will be possible to reflect the healthy system state and to identify log information that shows abnormal system behavior. Atlas Vector Search allows for conducting the required kNN search in an effective way and as a fully included service of the MongoDB Cloud Data Platform. Finally, using LLM, information about the error, including the analysis of the root cause, can be expressed in natural language, making the job of understanding and fixing the problem much easier for the staff who are in charge of maintenance.

Fraud Detection and Prevention

Telecom providers today are utilizing an advanced array of techniques for detecting and preventing fraud, constantly adjusting to the dynamic nature of telecom fraud. Routine activities for detecting fraud consist of tracking unusual call trends and data usage, along with safeguarding against SIM Swap incidents, a method frequently used for identity theft. To prevent fraud, strategies are applied at various levels, starting with stringent verification for new customers, during SIM swaps, or for transactions with elevated risk, taking into account the unique risk profile of each customer.

Machine learning offers telecommunications companies a powerful tool to enhance their fraud detection and prevention capabilities by training ML models on historical data like Call Detail Records (CDR). Moreover, these algorithms can assess the individual risk profile of each customer, tailoring detection and prevention strategies to their specific patterns of use.



The models can adapt over time, learning from new data and emerging fraud tactics, thus enabling real-time detection and the automation of fraud prevention measures, reducing manual checks, and speeding up response times

To deal with fraud successfully, a multitude of data dimensions need to be put into consideration, making the reaction time a critical factor in preventing the worst things to happen, so the solution must also support fast, sub-second decisions. By vectorizing the data with an appropriate ML model, normal (healthy) business can be defined, and in turn, deviations from the norm be identified, like for instance, suspicious user activities. In addition to Atlas Vector Search, the MongoDB Query API supports stream processing, simplifying data ingestion from various sources and detecting fraud in real time.

Content Discovery

Today's media organizations are expected to offer a degree of content personalization, from streaming services to online publications and more. Viewers want intelligently selected and suggested content tailored to their interests.

Leveraging AI can significantly enhance the process of suggesting the next best article to read or show to stream. The most powerful implementations of content personalization track behavior of the user like which content was searched for, how long was content displayed before the next click happened, what categories, etc. Based on these parameters, similar content can be presented, or, as an alternative strategy, content from unseen areas of the portal be presented, to have the user discover new types of media and check their appetite for consuming it.

To bring the right content to the right people at the right time, an automated system needs to maintain a multitude of information facets, which will lay the foundation for proper suggestions. With MongoDB and its document model, all required data points can be easily and flexibly stored in a user's profile, in content, and media.

Ultimately, by vectorizing the content, an even more powerful system of content suggestions can be built with Atlas Vector Search, which allows for similarity search that goes well beyond comparing just keywords or a list of attributes.

Other Notable Use Cases

Differential Pricing

Gather insights into what customers are willing to spend on content or a service by conducting A/B tests and analyzing the data with a ML algorithm. This method facilitates the adoption of dynamic pricing models instead of sticking to a standard price list, thereby potentially enhancing revenue and increasing the paying customer base.

Search Generative Experiences (SGE)

Provide more dynamic, personalized, and contextually relevant search results, thus making information retrieval not only more efficient but also more engaging and useful. This can include personalization and summarization elements, as well.

Content Summarization & Reformatting

Design a smart assistant tailored for writers, capable of providing automatic suggestions for content summaries, identifying suitable SEO keywords, and adapting articles for various specific audiences.

Contact Information

Benjamin Lorenz

Telco & Media Industry
Solutions Principal

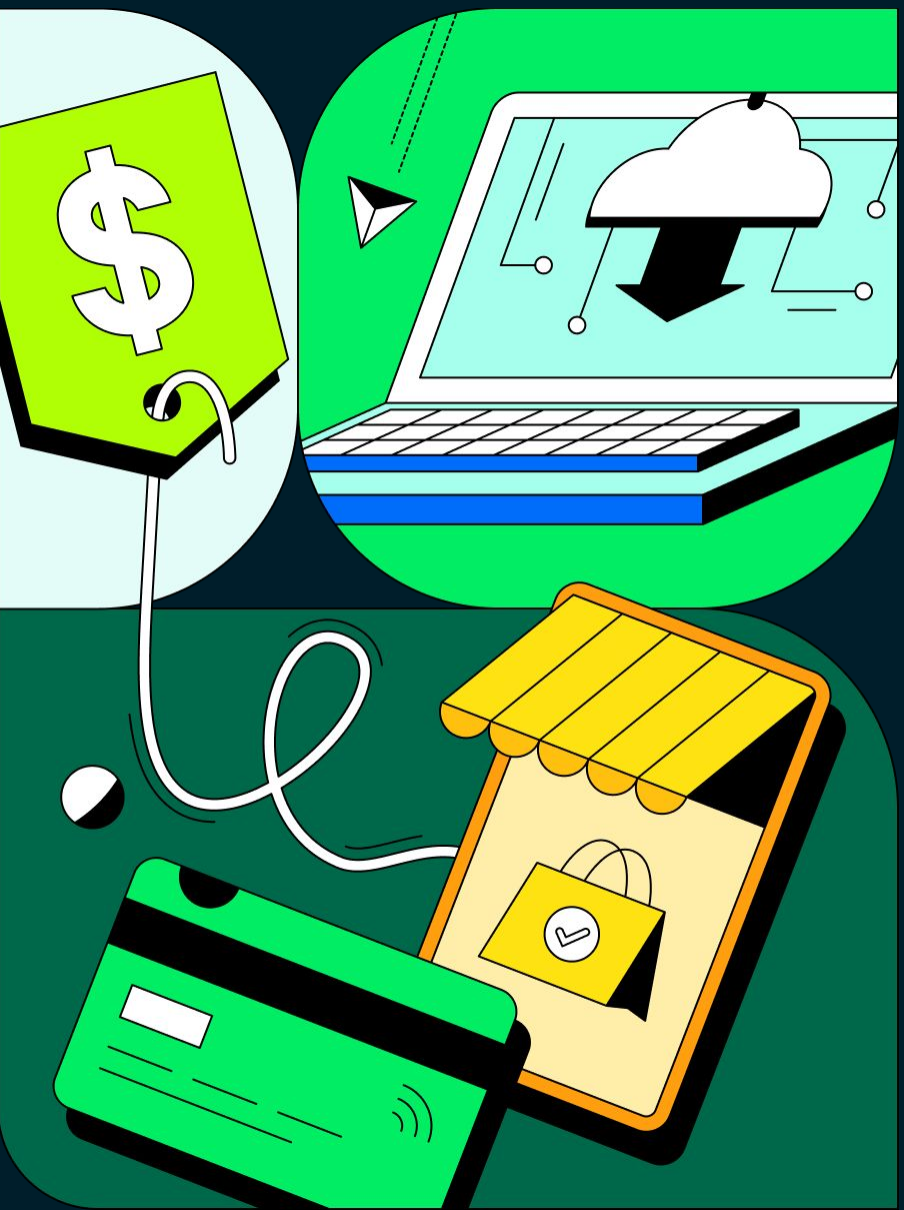
benjamin.lorenz@mongodb.com

FOR MORE INFORMATION AND RESOURCES

Visit MongoDB for Telecommunications

Atlas for Industries

Retail



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.

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 in order 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, and 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 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. 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, and 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 clickstream 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 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. 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 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.

READ MORE

AI-Enhanced Search in Ecommerce With MongoDB

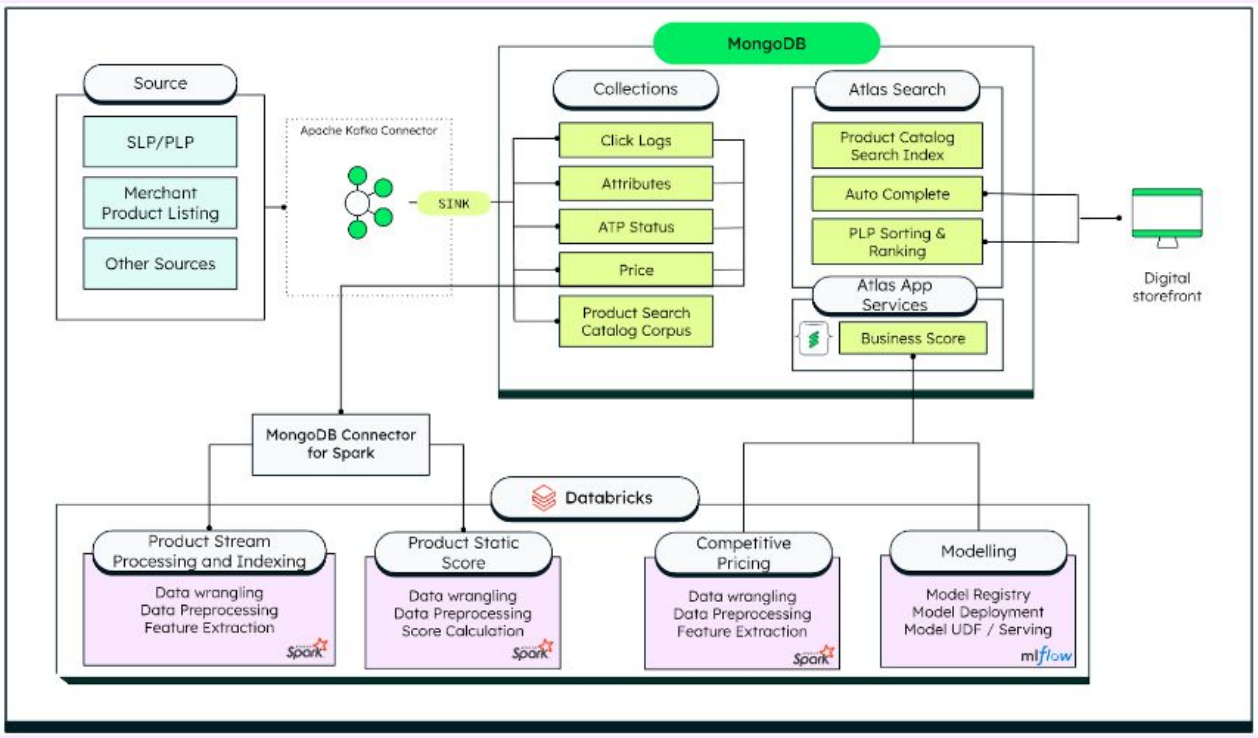


Figure 5. Architecture of an AI-enhanced search engine explaining the different MongoDB Atlas components and Databricks notebooks and workflows used for data cleaning and preparation, product scoring, dynamic pricing, and vector search

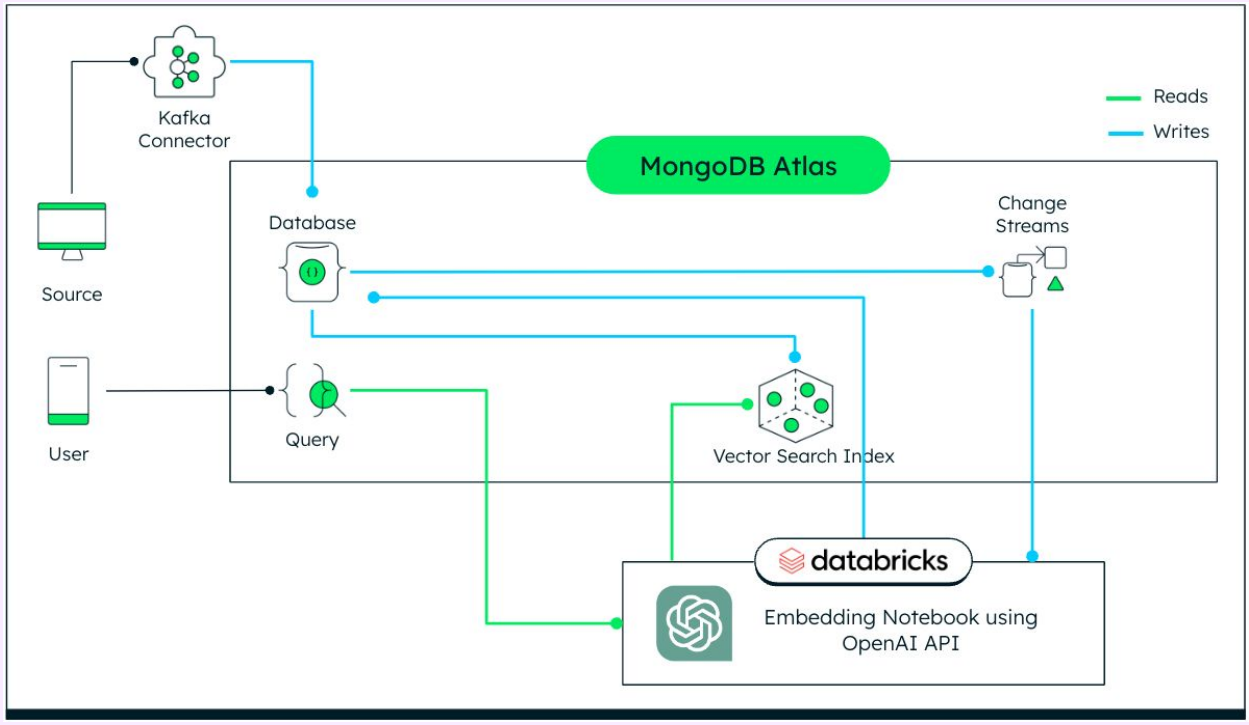


Figure 6. Architecture of a vector search solution showcasing how the data flows through the different integrated components of MongoDB Atlas and Databricks

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 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 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) 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 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 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

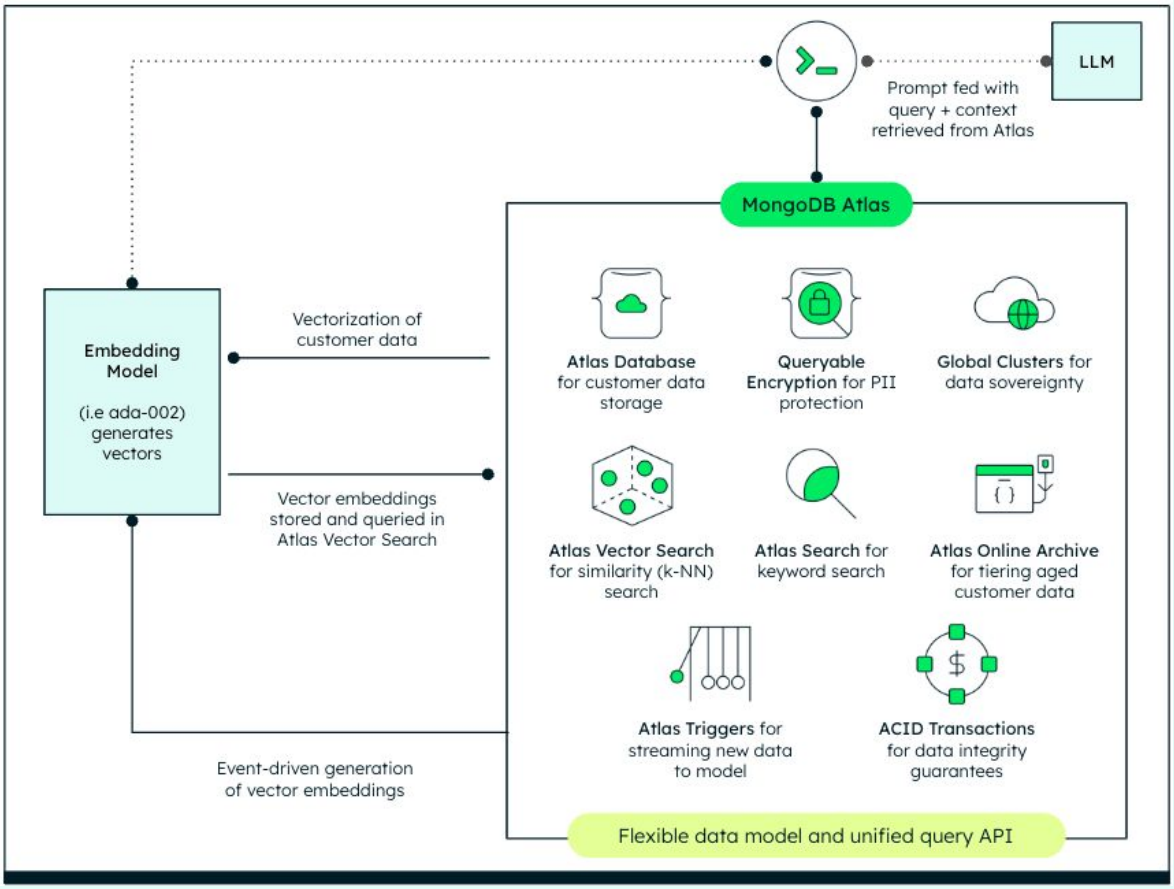


Figure 7. Architecture showing how MongoDB can be used with a vector embedding model. Data in MongoDB e.g. customer data, will be fed into the model (via Spark or other connector) and the generated vector embeddings will be added to each document in the collection. Then an Atlas Vector Search index can 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.

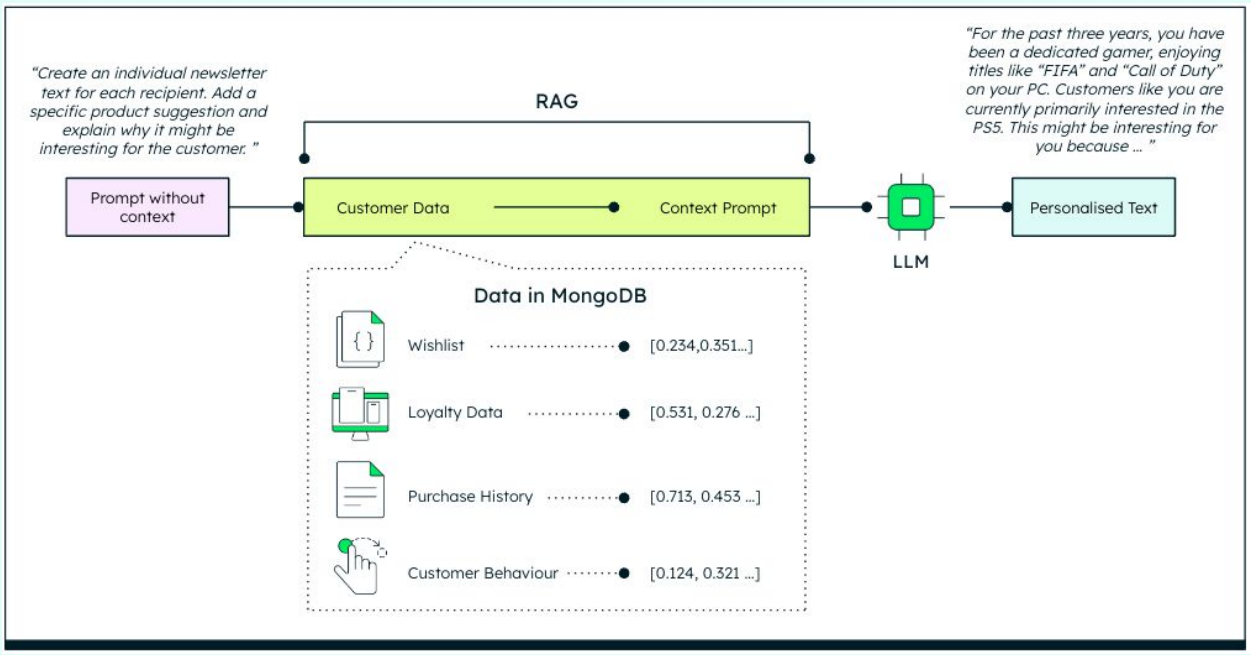


Figure 8. Example of the data flow for an AI-generated personalized newsletter. The prompt is entered by a user on the left hand side and context is added via the vectorised data in MongoDB- wishlist, loyalty data, purchase history, and customer behavior. Using RAG, the LLM can produce a personalized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

Demand Forecasting & 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 optimal performance and provide 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

Contact Information

Genevieve Broadhead

MongoDB Global Lead,
Retail Solutions

genevieve.broadhead@mongodb.com

Prashant Juttukonda

Retail Industry
Solutions Principal

prashant.juttukonda@mongodb.com

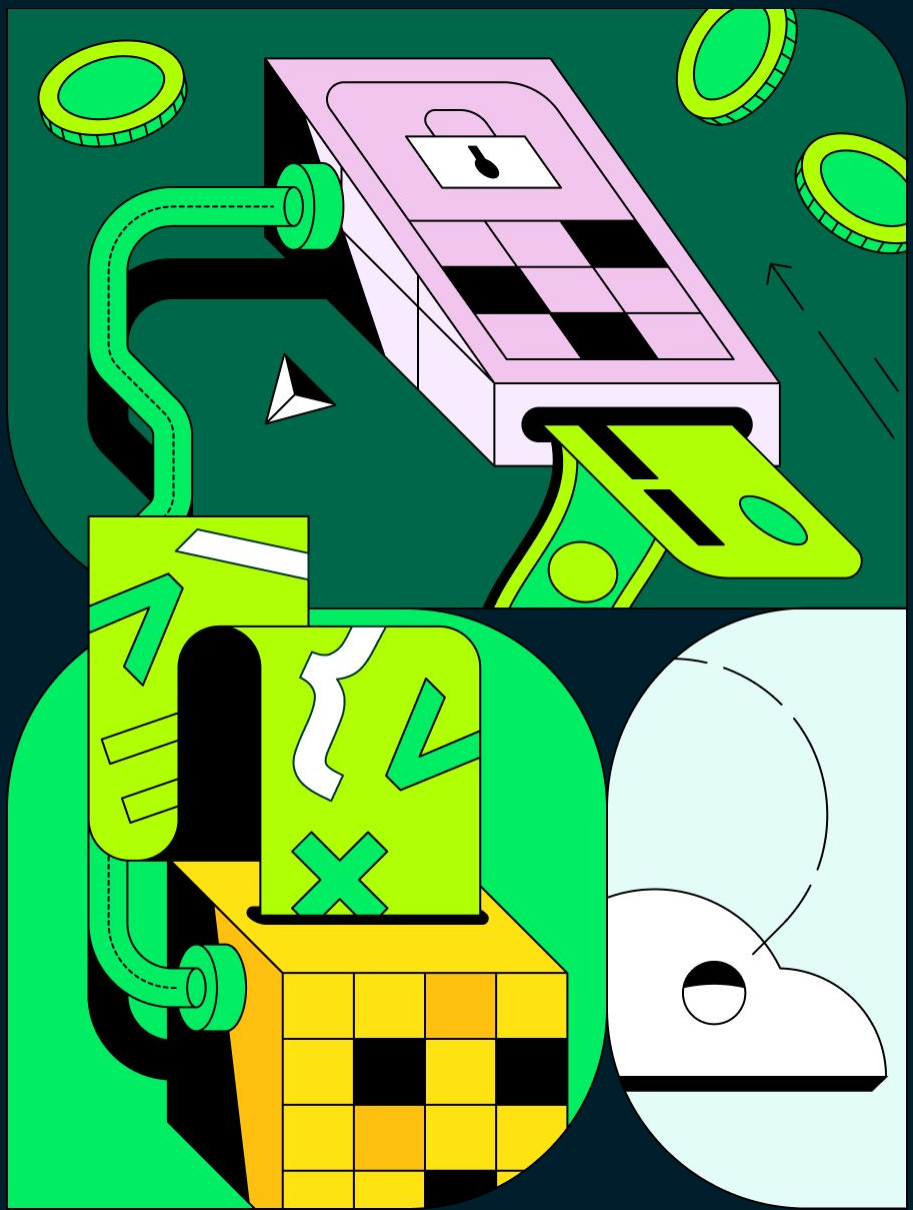
FOR MORE INFORMATION AND RESOURCES

Visit MongoDB for Retail

Atlas for Industries

Financial Services

Celent and McKinsey see Financial Services as one of the most impacted industries by the AI revolution. However, it is to be expected that the benefits will being in internal process optimizations and will only slowly be adopted in direct customer interactions.



One key service that relationship managers provide to their private banking customers is aggregating and condensing information.

Because banks typically operate on fragmented infrastructure, this can require a lot of detailed knowledge about this infrastructure and how to source information such as:

- When are the next coupon dates for bonds in the portfolio?
- What has been the cost of transactions for a given portfolio?
- What would be a summary of our latest research?
- Please generate a summary of my conversation with the client.

Until now, these activities would be highly manual and exploratory. For example, a relationship manager (RM) looking for the next coupon dates would likely have to go into each of the clients' individual positions and manually look up the coupon dates. If this is a frequent enough activity, the RM could raise a request for change with the product manager of the portfolio management software to add this as a standardized report. But even if such a standardized report existed, the RM might struggle to find the report quickly. Overall, the process is time-consuming.

Generative AI systems have been shown to be able to facilitate such tasks. Even without specifically trained models, RAG can be used to have the AI generate the correct answers, provide the inquirer with a detailed explanation of how to get to the data, and, in the same cases directly execute the query against the system and report back the results. Similar to a human, it is critical that the algorithm have access to not only the primary business data, e.g. the portfolio data of the customer, but also user manuals and static data. Detailed customer data, in machine-readable

format but also as text documents, is used to personalize the output for the individual customer.

In an interactive process, the RM can instruct the AI to add more information about specific topics, tweak the text, or make any other necessary changes. Ultimately, the RM will be the quality control for the AI's output to mitigate hallucinations or information gaps

As outlined above, not only will the AI need highly heterogeneous data from highly .

structured portfolio information to text documents and system manuals to provide a flexible natural language interface for the RMs, it will also have to have timely processing information about a customer's transactions, positions, and investment objectives. Providing transactional database capabilities as well as vector search makes it easy to build RAG-based applications using MongoDB's developer data platform.

Risk Management and Regulatory Compliance

Risk & Fraud Prevention

Banks are tasked not only with safeguarding customer assets but also with detecting fraud, verifying customer identities (KYC), supporting sanctions regimes (Sanctions), and preventing various illegal activities (AML). The challenge is magnified by the sheer volume and complexity of regulations, making the integration of new rules into bank infrastructure costly, time-consuming, and often inadequate. ***For instance, when the EU's Fifth Anti-Money Laundering Directive (AML) was implemented, it broadened regulations to cover virtual currencies and prepaid cards.*** Banks had to swiftly update their onboarding processes, and software, train staff, and possibly update their customer interfaces to comply with these new requirements.

AI offers a transformative approach to fraud detection and risk management by automating the interpretation of regulations, supporting data cleansing, and enhancing the efficacy of surveillance systems. Unlike static, rules-based frameworks that may miss or misidentify fraud due to narrow scope or limited data, AI can adaptively learn and analyze vast datasets to identify suspicious activities more accurately. Machine learning, in particular, has shown promise in trade surveillance, offering a more dynamic and comprehensive approach to fraud prevention.

Regulatory Compliance and Code Change Assistance

The regulatory landscape for banks has grown increasingly complex, demanding significant resources for the implementation of numerous regulations. Traditionally, adapting to new regulations has required the manual translation of legal text into code, provisioning of data, and thorough quality control—a process that is both costly and time-consuming, often leading to incomplete or insufficient compliance. For instance, to comply with the ***Basel III international banking regulations***, *developers must undertake extensive coding changes to accommodate the requirements laid out in thousands of pages of documentation.*

AI has the capacity to revolutionize compliance by automating the translation of regulatory texts into actionable data requirements and validating compliance through intelligent analysis. This approach is not without its challenges, as AI-based systems may produce non-deterministic outcomes and unexpected errors. However, the ability to rapidly adapt to new regulations and provide detailed records of compliance processes can significantly enhance regulatory adherence

Financial Document Search and Summarization

Financial institutions, encompassing both retail banks and capital market firms, handle a broad spectrum of documents critical to their operations. Retail banks focus on contracts, policies, credit memos, underwriting documents, and regulatory filings, which are pivotal for daily banking services. On the other hand, capital market firms delve into company filings, transcripts, reports, and intricate data sets to grasp global market dynamics and risk assessments.

These documents often arrive in unstructured formats, presenting challenges in efficiently locating and synthesizing the necessary information. While retail banks aim to streamline customer and internal operations, capital market firms prioritize the rapid and effective analysis of diverse data to inform their investment strategies. Both retail banks and capital market firms allocate considerable time to searching for and condensing information from documents internally, resulting in reduced direct engagement with their clients.

Generative AI can streamline the process of finding and integrating information from documents by using NLP and machine learning to understand and summarize content. This reduces the need for manual searches, allowing bank staff to access relevant information more quickly.

MongoDB can store vast amounts of both live and historical data, regardless of its format which is typically needed for AI applications. It offers Vector Search capabilities essential for Retrieval Augmented Generation (RAG) in. MongoDB supports transactions, ensuring data accuracy and consistency for AI model retraining with live data. It facilitates data access for both deterministic algorithms and AI-driven rules through a single interface. MongoDB boasts a strong partnership ecosystem, including companies like Radiant AI and Mistral LLM, to speed up solution development

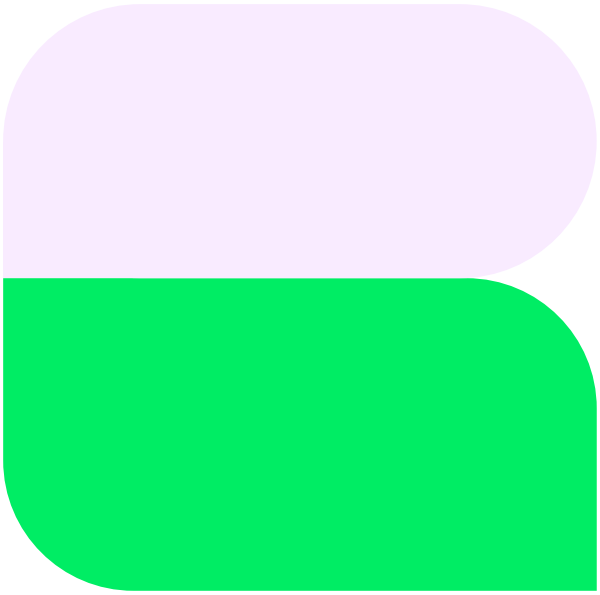
ESG Analysis

The profound impact of environmental, social, and governance (ESG) is evident, driven by regulatory changes, especially in Europe, compelling financial institutions to integrate ESG into investment and lending decisions. Regulations such as the EU Sustainable Finance Disclosure Regulation (SFDR) and the EU Taxonomy Regulation are examples of such directives that require financial institutions to consider environmental sustainability in their operations and investment products. Investors' demand for sustainable options has surged, leading to increased ESG-focused funds. The regulatory and commercial requirements in turn, drive banks to also improve their green lending practices. This shift is strategic for financial institutions, attracting clients, managing risks, and creating long-term value.

However, financial institutions face many challenges in managing different aspects of improving their ESG analysis. The key challenges include defining and aligning standards, and processes and managing the flood of rapidly changing and varied data to be included for ESG analysis purposes.

AI can help to address these key challenges in not only an automatic but also adaptive manner via techniques like machine learning. Financial institutions and ESG solution providers have already leveraged AI to extract insights from corporate reports, social media, and environmental data, improving the accuracy and depth of ESG analysis. As the market demands a more sustainable and equitable society, predictive AI combined with generative AI can also help to reduce bias in lending to create a fairer and more inclusive financing while improving the predictive powers. The power of AI can help facilitate the development of sophisticated sustainability models and strategies, marking a leap forward in integrating ESG into broader financial and corporate practices.

MongoDB's dynamic architecture revolutionizes ESG data management, handling semi-structured and unstructured data. Its flexible schema nature allows the adaptation of data models as ESG strategies evolve. Advanced text search capabilities efficiently analyze vast semi-structured data for informed ESG reporting. Support for vector search enriches ESG analysis with multimedia content insights.



Incorporating Large Language Models (LLMs) enhances MongoDB's capacity to process ESG textual content, automating sentiment extraction, summarization, and trend identification. Combining LLMs with vector data management capabilities, generative AI applications can be created to interpret the complex and evolving sustainability taxonomy and guide the investment and financing processes in a compliant manner. This AI-driven approach, supported by MongoDB's robust data management, offers a sophisticated means of analyzing extensive narrative data in ESG reporting.

Furthermore, MongoDB supports geospatial and network graph analytics, providing a powerful combination of analytics to identify the physical risks associated with climate change (e.g., floods, wildfires) to assets financed by banks or investment firms and for assessing supply chain impacts of the climate risks. The risk analytics can then enable targeted strategies for risk mitigation and supply chain resilience.

MongoDB's value extends beyond ESG data management, accelerating productivity for developers and data science teams. Its intuitive data model, analytical tools, and AI integrations streamline the development and deployment of data-driven applications, making MongoDB pivotal for organizations advancing their ESG agendas efficiently.

Below is a diagram of an enterprise ESG solution architecture with the boxes labeled with the green leaf where MongoDB can be deployed to support the ESG data analytics related services.

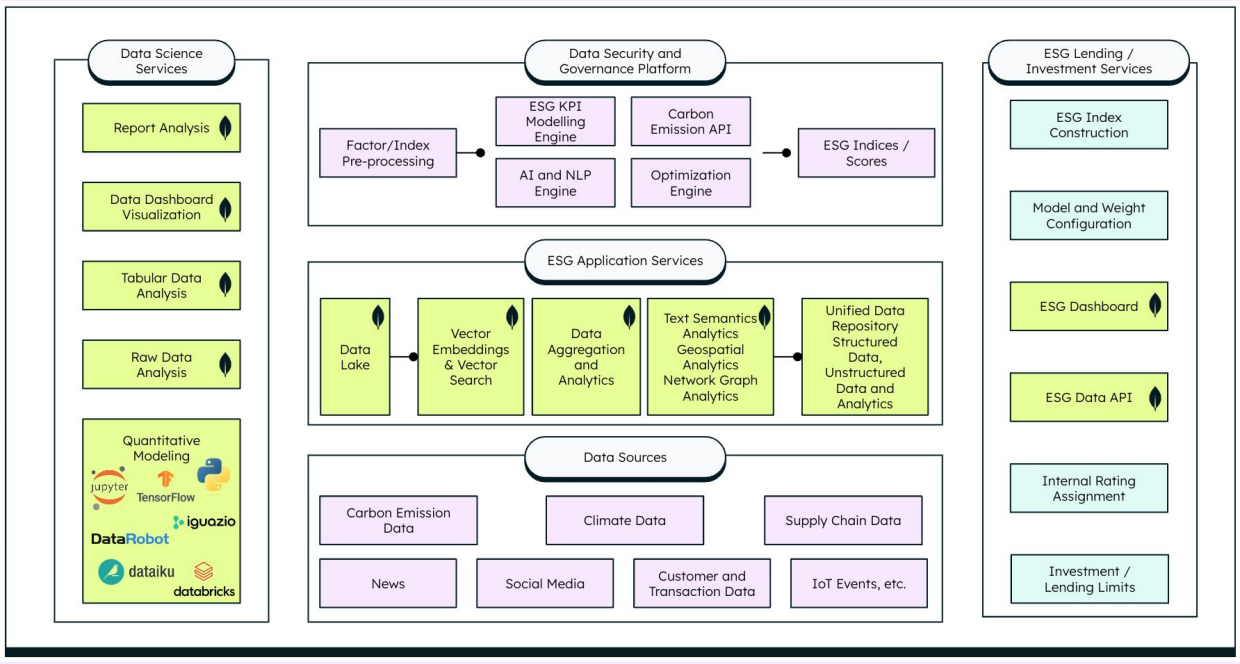


Figure 9. Blueprint for Enterprise ESG Solution Architecture Using MongoDB

Other Notable Use Cases

Risk Modeling

AI can be used to create synthetic scenarios and data that can be used to stress test financial systems and models

Investment Portfolio Optimization

AI can be used to generate an optimized portfolio tailored to customer preferences.

Contact Information

Joerg Schmuecker

MD, Financial Services Industry

j.schmuec@mongodb.com

Wei You Pan

Director, Financial Services Industry

weiyou.pan@mongodb.com

Shiv Pullepu

Principal, Financial Services Industry

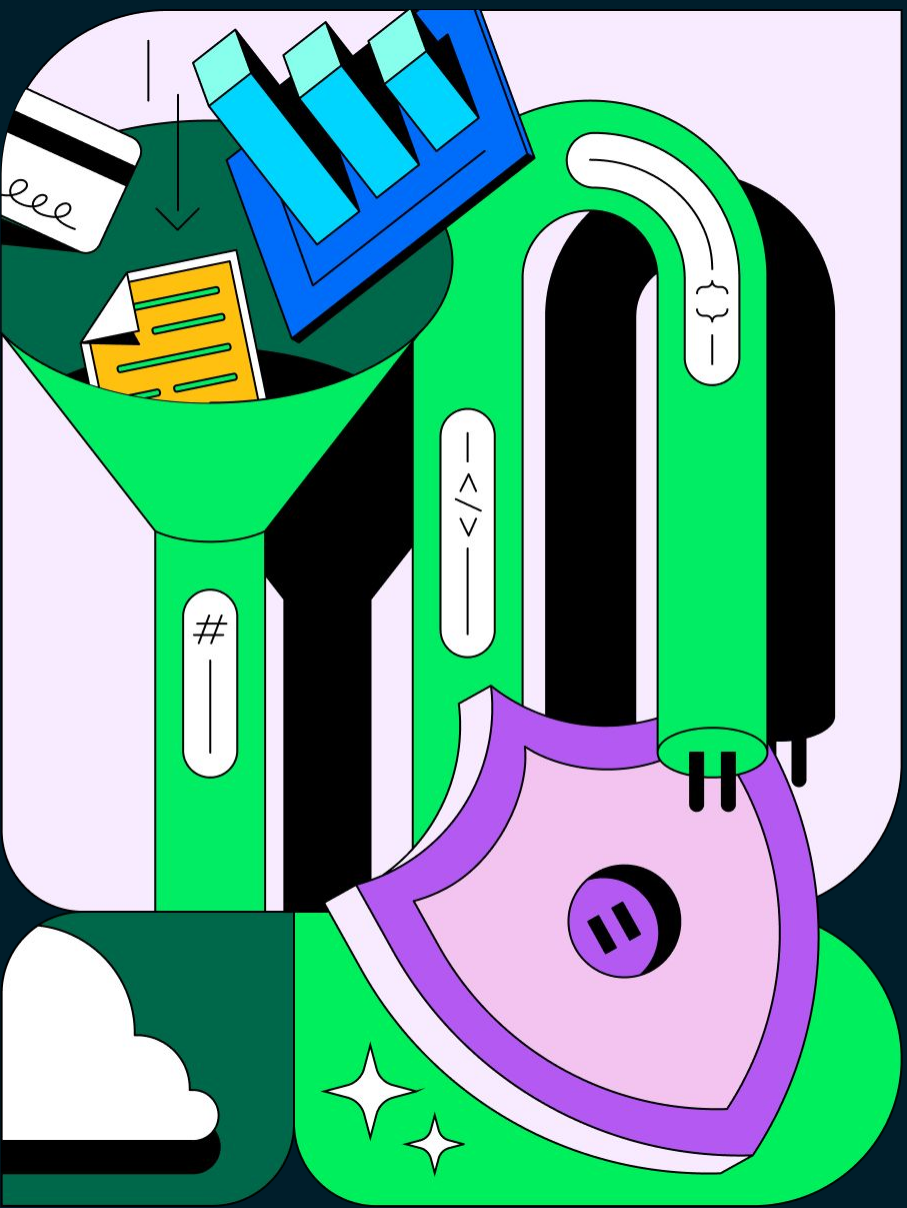
shiva.pullepu@mongodb.com

FOR MORE INFORMATION AND RESOURCES

Visit MongoDB for Financial Services

Atlas for Industries

Insurance



With its ability to streamline process, enhance decision-making, and improve customer experience with far less time, resources and staff than traditional IT systems, AI offers insurers great promise.

In an inherently information-driven industry, insurance companies ingest, analyze, and process massive amounts of data. Whether it's agents and brokers selling more policies, underwriters adequately pricing, renewing and steering product portfolios, claim handlers adjudicating claims, or service representatives providing assurance and support, data is at the heart of it all.

Given the volumes of data, and the amount of decision-making that needs to occur based on it, insurance companies have a myriad of technologies and IT support staff within their technology investment portfolios. It's no surprise that AI is at the top of the list when it comes to current or prospective IT investments. With its ability to streamline processes, enhance decision-making, and improve customer experiences with far less time, resources and staff than traditional IT systems, AI offers insurers great promise.

Underwriting & Risk Management

Few roles within insurance are as important as that of the underwriters who strike the right balance between profit and risk, bring real-world variables to the actuarial models at the heart of the insurer, and help steer product portfolios, markets, pricing, and coverages. Achieving equilibrium between exposures and premiums means constantly gathering and analyzing information from a myriad of sources to build a risk profile sufficient and detailed enough to make effective policy decisions.

While many well-established insurers have access to a wealth of their own underwriting and claims experience data, integrating newer and real-time sources of information, keeping up with regulatory changes, and modeling out what-if risk scenarios still involve significant manual effort.

Perhaps the single greatest advantage of AI will be its ability to quickly analyze more information with fewer people and resources. The long-term impact will likely be profound, and there is tremendous promise within underwriting.

Advanced Analytics

Traditional IT systems are slow to respond to changing formats and requirements surrounding data retrieval. The burden falls on the underwriter to summarize data and turn that into information and insight. Large Language Models are now being leveraged to help speed up the process of wrangling data sources and summarizing the results, helping underwriting teams make quicker decisions from that data.

Workload and Triage Assistance

AI models are mitigating seasonal demands, market shifts, and even staff availability that impact the workload and productivity of underwriting teams, saving underwriting time for the high-value accounts and customers where their expertise is truly needed. Amid high volumes for new and renewal underwriting, traditional AI models can help classify and triage risk, sending very low-risk policies to ‘touchless’ automated workflows, low to moderate risk to trained service center staff, and high-risk and high value accounts to dedicated underwriters.

Decision-Making Support

Determining if a quoted rate needs adjustment before binding and issuing can take significant time and manual effort. So can preparing and issuing renewals of existing policies, another large portion of the underwriters’ day-to-day responsibilities. Automated underwriting workflows leveraging AI are being employed to analyze and classify risk with far less manual effort. This frees up significant time and intellectual capital for the underwriter.

Vast amounts of data analyzed by underwriters are kept on the underwriters desktop rather than IT-managed databases. MongoDB offers an unparalleled ability to store data from a vast amount of sources and formats, and deliver the ability to respond quickly to requests to ingest new data. As data and requirements change, the Document Model allows insurers to simply add more data and fields without the costly change-cycle associated with databases that rely on single, fixed structures.

For every major business entity found within the underwriting process, such as broker, policy, account and claim, there is a wealth of unstructured data sources, waiting to be leveraged by generative AI. MongoDB offers insurers a platform that consolidates complex data from legacy systems, builds new applications, and extends those same data assets to AI augmented workflows. By eliminating the need for niche databases for these AI-specific workloads, MongoDB reduces technology evaluation and on-boarding time, development time, and developer friction.

LEARN MORE

Automating Digital Underwriting with ML

Claim Processing

Efficient claim processing is critical for an insurer. Timely resolution of a claim and good communication and information transparency throughout the process are key to maintaining positive relationships and customer satisfaction. In addition, insurers are on the hook to pay and process claims according to jurisdictional regulations and requirements, which may include penalties for failing to comply with specific timelines and stipulations.

In order to process a claim accurately, a wealth of information is needed. A typical automobile accident may include not only verbal and written descriptions from claimants and damage appraisers but also unstructured content from police reports, traffic and vehicle dashboard cameras, photos, and even vehicle telemetry data. Aligning the right technology and the right amount of your workforce in either single or multi-claimant scenarios is crucial to meeting the high demands of claim processing.

Taming the Flood of Data

AI is helping insurers accelerate the process of making sense of a trove of data and allowing insurers to do so in real time. From Natural Language Processing to image classification and vector embedding, all the pieces of the puzzle are now on the board for insurers to make a generation leap forward when it comes to transforming their IT systems and business workflows for faster information processing.

Claims Experience

Generating accurate impact assessments for catastrophic events in a timely fashion in order to inform the market of your exposure can now be done with far less time, and with far more accuracy, by cross-referencing real-time and historical claims experience data, thanks to the power of Generative AI and vector-embedding of unstructured data.

Claim Expediter

Using vector-embeddings from photo, text, and voice sources, insurers are now able to decorate inbound claims with richer and more insightful metadata so that they can more quickly classify, triage, and route work. In addition, real-time insight into workload and staff skills and availability is allowing insurers to be even more prescriptive when it comes to work assignments, driving towards higher output and higher customer satisfaction.

Litigation Assistance

Claims details are not always black and white, parties do not always act in good faith, and insurers expend significant resources in the pursuit of resolving matters. AI is helping insurers drive to resolution faster and even avoid litigation and subrogation altogether, thanks to its ability to help us analyze more data more effectively and more quickly.

Risk Prevention

Many insurers provide risk-assessment services to customers using drones, sensors, or cameras, to capture and analyze data. This data offers the promise of preventing losses altogether for customers and lowering exposures, liability, and expenses for the insurer. This is possible thanks to a combination of vector-embedding, traditional and generative AI models.

LEARN MORE

AI-Enhanced Claim Adjustment for Auto Insurance

Customer Experience

Accessing information consistently during a customer service interaction, and expecting the representative to quickly interpret it, are perennial challenges with any customer service desk. Add in the volume, variety, and complexity of information within insurance, and it's easy to understand why many insurers are investing heavily in transformation of their customer experience call center systems and processes.

24/7 Virtual Assistance

As with many AI-based chat agents, the advantage is that it can free up your call center staff to work on more complex and high-touch cases. Handling routine inquiries can now include far more complex scenarios than before, thanks to the power of vector-embedded content and Large Language Models.

Customer Profiles

Every interaction is an opportunity to learn more about your customers. Technologies such as voice-to-text streaming, vector embedding, and generative AI help insurers build out a more robust 'social profile' of their customers in near real-time.

Claims Assistance

Generative AI can deliver specific claim-handling guidelines to claim-handling staff in real-time, while traditional ML models can interrogate real-time streams of collected information to alert either the customer or the claim-handler to issues with quality, content, or compliance. AI capabilities allow insurers to process more claims more quickly and significantly reduce errors or incomplete information.

Real-time Fraud Detection

According to estimates from the Coalition Against Insurance Fraud, the U.S. insurance industry lost over $308 billion to fraud in 2022. With vector-embedding of unstructured data sources, semantic and similarity searches across both vector and structured metadata, and traditional machine learning models, insurers can detect and prevent fraud in ways that were simply not ever before possible.

Other Notable Use Cases

Predictive Analytics

AI-powered predictive analytics can anticipate customer needs, preferences, and behaviors based on historical data and trends. By leveraging predictive models, insurers can identify at-risk customers, anticipate churn, and proactively engage with customers to prevent issues and enhance satisfaction.

Crop Insurance and Precision Farming

AI is being used in agricultural insurance to assess crop health, predict yields, and mitigate risks associated with weather events and crop diseases, which helps insurers offer more accurate and tailored crop insurance products to farmers.

Predictive Maintenance for Property Insurance

AI-powered predictive maintenance solutions, leveraging IoT sensors installed in buildings and infrastructure, are used in property insurance to prevent losses and minimize damage to insured properties.

Usage-Based Insurance (UBI) for Commercial Fleets

AI-enabled telematics devices installed in commercial vehicles collect data on driving behavior, including speed, acceleration, braking, and location. Machine learning algorithms analyze this data to assess risk and determine insurance premiums for commercial fleets to help promote safer driving practices, reduce accidents, and lower insurance costs for businesses.

Contact Information

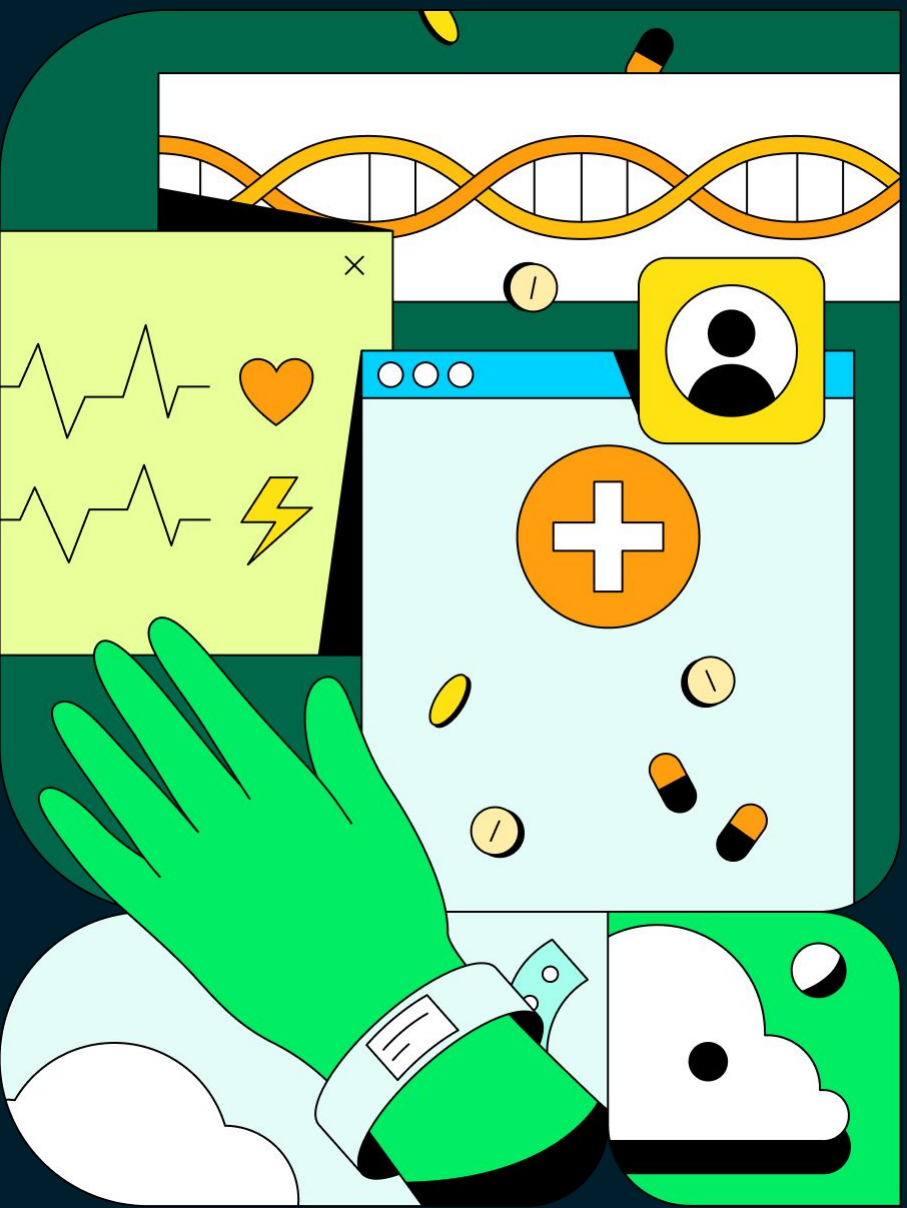
Jeff Needham

Insurance Industry
Solutions Principal

jeff.needham@mongodb.com

Atlas for Industries

Healthcare and Life Sciences



AI emerges as a pivotal technology, with the potential to enhance decision-making, improve patient experiences, and streamline operations – and to do so more efficiently than traditional systems.

In healthcare, transforming data into actionable insights is vital for enhancing clinical outcomes and advancing patient care. From medical professionals improving care delivery to administrators optimizing workflows and researchers advancing knowledge, data is the lifeblood of the healthcare ecosystem. Today, AI emerges as a pivotal technology, with the potential to enhance decision-making, improve patient experiences, and streamline operations — and to do so more efficiently than traditional systems.

Patient Experience & Engagement

While they may not expect it based on past experiences, patients crave a seamless experience with healthcare providers. Ideally, patient data from healthcare services, including telehealth platforms, patient portals, wearable devices, and EHR, can be shared – securely – across interoperable channels. Unfortunately, disparate data sources, burdensome and time-consuming administrative work

for providers, and overly complex and bloated solution stacks at the health system level all stand in the way of that friction-free experience.

AI can synthesize vast amounts of data and provide actionable insights, leading to personalized and proactive patient care, automated administrative processes, and real-time health insights. AI technologies, such as machine learning algorithms, natural language processing, and chatbots, are being used to enhance and quantify interactions. Additionally, AI-powered systems can automatically schedule appointments, send notifications, and optimize clinic schedules, all reducing wait times for patients. AI-enabled chatbots and virtual health assistants provide 24/7 support, offering instant responses, medication reminders, and personalized health education. AI can even identify trends and predict health events, allowing for early intervention and reduction in adverse outcomes.

MongoDB's flexible data model can unify disparate data sources, providing a single view of the patient that integrates EHRs, wearable data, and patient-generated health data for personalized care and better patient outcomes. For wearables and medical devices, MongoDB is the ideal underlying data platform to house time series data, significantly cutting down on storage costs while enhancing performance. With Atlas for the Edge, synchronization with edge applications, including hospital-at-home setups, becomes seamless.

On the patient care front, MongoDB can support AI-driven recommendations for personalized patient education and engagement based on the analysis of individual health records and engagement patterns, and Vector Search can power search functionalities within patient portals, allowing patients to easily find relevant information and resources, thereby improving the self-service experience.

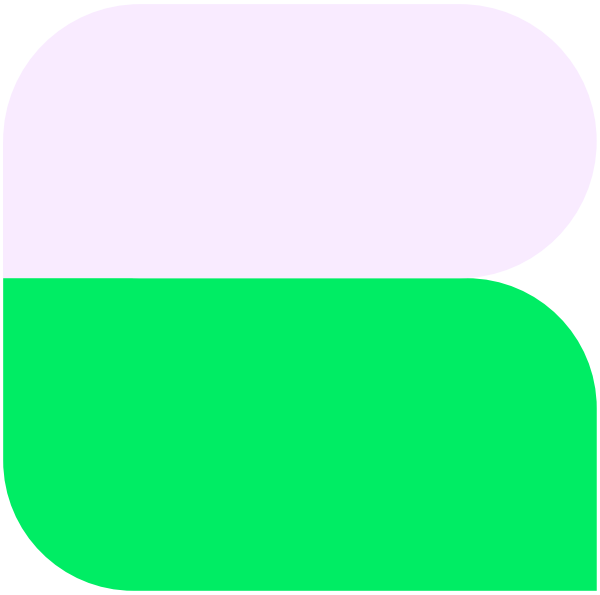
Enhanced Clinical Decision Making

Healthcare decision-making is critically dependent on the ability to aggregate, analyze, and act on an exponentially growing volume of data. From EHRs and imaging studies to genomic data and wearable device data, the challenge is not just the sheer volume but the diversity and complexity of data. Healthcare professionals need to synthesize information across various dimensions to make informed, real-time, accurate decisions. Interoperability issues, data silos, lack of data quality, and the manual effort required to integrate and interpret this data all stand in the way of better decision-making processes.

The advent of AI technologies, particularly NLP and LLMs, offers transformative potential for healthcare decision-making by automating the extraction and analysis of data from disparate sources, including structured data in EHRs and unstructured text in medical literature or patient notes.

By enabling the querying of databases using natural language, clinicians can access and integrate patient information more rapidly and accurately, enhancing diagnostic precision and personalizing treatment approaches. Moreover, AI can support real-time decision-making by analyzing streaming data from wearable devices, alerting healthcare providers to changes in patient conditions that require immediate attention.

MongoDB, with its flexible data model and powerful data development platform, is uniquely positioned to support the complex data needs of healthcare decision-making applications. It can seamlessly integrate diverse data types, from FHIR-formatted clinical data to unstructured text and real-time sensor data, in a single platform. By integrating MongoDB with Large Language Models (LLMs), healthcare organizations can create intuitive, AI-enhanced interfaces for data retrieval and analysis. This integration not only reduces the cognitive load on clinicians but also enables them to access and interpret patient data more efficiently, focusing their efforts on patient care rather than navigating complex data systems. MongoDB's scalability



ensures that healthcare organizations can manage growing data volumes efficiently, supporting the implementation of AI-driven decision support systems. These systems analyze patient data in real-time against extensive medical knowledge bases, providing clinicians with actionable insights and recommendations, thereby enhancing the quality and timeliness of care provided.

MongoDB's Vector Search further enriches decision-making processes by enabling semantic search across vast datasets directly within the database. This integrated approach enables the application of pre-filters based on extensive metadata, enhancing the efficiency and relevance of search results without the need to synchronize with dedicated search engines or vector stores, meaning healthcare professionals can utilize previously undiscoverable insights, streamlining the identification of relevant information and patterns.

Clinical Trials and Precision Medicine

The need for innovation and transformation isn't just limited to the patient-provider-healthcare system experience. The challenges of conducting clinical trials and advancing precision medicine are significant, from identifying and enrolling suitable participants to data management practices are fraught with the potential for errors, compromising the accuracy and reliability of trial outcomes. Moreover, the traditional one-size-fits-all approach to treatment development fails to address the unique genetic makeup of individual patients, limiting the effectiveness of therapeutic interventions.

AI can make clinical trials faster and treatments more personalized. It's like having a super-smart assistant that can quickly find the right people for studies, keep track of all the data without making mistakes, and even predict which medicines will work best for different people. This means doctors can create safe, efficient treatments that fit you perfectly, just like a tailor-made suit. Plus, with AI's help, these custom treatments can be developed quicker and be more affordable, bringing us closer to a future where everyone gets the care they need, designed just for them. It's a big step towards making medicine not just about treating sickness but about creating health plans that are as unique as patients are.

MongoDB plays a pivotal role in modernizing clinical trials and advancing precision medicine by addressing complex data challenges. Its flexible data model excels in integrating diverse data types, from EHRs and genomic data to real-time patient monitoring streams. This capability is crucial for clinical trials and precision medicine, where combining various data sources is necessary, sometimes through a project purpose ODL, to develop a comprehensive understanding of patient health and treatment responses.

For clinical trials, MongoDB can streamline participant selection by efficiently managing and querying vast datasets to identify candidates who meet specific criteria, significantly reducing the recruitment time. Its ability to handle large-scale, complex datasets in real time also facilitates the dynamic monitoring of trial participants, enhancing the safety and accuracy of trials.

Other Notable Use Cases

Patient Flow Optimization and Emergency Department Efficiency

AI algorithms can process historical and real-time data to forecast patient volumes, predict bed availability, and identify optimal staffing levels, enabling proactive resource allocation and patient routing.

AI-Enhanced Digital Pathology and Medical Imaging

Build modern VNA (Vendor Neutral Archive and Digital pathology solutions with innovative approaches, dealing with interoperable data, and manage extensive metadata associated with all your resources enabling fast findings and automated annotations.

Virtual Health Assistants for Chronic Disease Management

Utilizing AI-powered virtual assistants to monitor patients' health status, provide personalized advice, and support medication adherence for chronic conditions such as diabetes and hypertension.

Operational Efficiency in Hospital Resource Management

Implementing AI to optimize hospital operations, from staff scheduling to inventory management, ensuring resources are used efficiently and patient care is prioritized.

Contact Information

Francesc Mateu Amengual

Healthcare Industry
Solutions Principal

francesc.mateu@mongodb.com

FOR MORE INFORMATION AND RESOURCES

Visit MongoDB Atlas for Healthcare

AI Ecosystem and Partnerships

How AI companies are leveraging MongoDB to build SaaS and Component-based AI solutions

Unlocking the Power of AI With SaaS

AI and SaaS are each powerful forces in tech. But when fused together, their impact can be even greater than expected. Let's delve into MongoDB SaaS AI partners that you leverage to save building the solution yourself.

Iguazio (acquired by McKinsey) & MongoDB: Building & scaling gen AI apps for enterprises efficiently, effectively and responsibly.

Iguazio (acquired by McKinsey & company) is a Gen AI Factory & MLOps tech stack that accelerates the development, deployment and management of ML and Gen AI applications. Trusted by large Financial Services, Manufacturing, Transportation and Retail clients, including Fortune 500 companies, Iguazio ensures AI and gen AI applications don't remain in the lab, but create real impact in live business environments.

From building your first Gen AI app, to a full blown Gen AI Factory

By automating and streamlining AI, Iguazio accelerates time-to-market, lowers operating costs, de-risks, provides guardrails and enhances business impact and profitability. This enables Iguazio to support enterprise needs, either in a self-serve or managed services model, with an open and flexible architecture.

Iguazio provides you with the latest capabilities for:

1. **Gen AI Ops:** Operationalizing AI / Gen AI apps efficiently at scale to create real business impact.
2. **Gen AI Guardrails:** De-risking Gen AI to meet compliance, regulations and controls relevant to your industry while ensuring peak performance.

Iguazio supports data management, training and fine-tuning LLMs, application deployment and LiveOps that enables monitoring models and data for feedback.

Accelerated and De-risked AI & Gen AI Deployment

- **AI / Gen AI Operationalization with minimal engineering**: MongoDB and Iguazio offer a unified, scalable data solution from prototype to production.
- **Hybrid environments:** MongoDB and Iguazio offer flexible deployment options: cloud, on-premises, or hybrid, tailored to meet MLOps/LLMOps and DataOps needs.
- MongoDB and Iguazio **unify all data management needs** (logging, auditing, etc.) in a single solution to ensure consistency, faster performance and significantly less overhead.
- **Scalability and performance** enable effortless management of large data volumes and intricate transformations, ensuring high reliability and accuracy.
- **Security and compliance**: MongoDB and Iguazio ensure top security and compliance for finance and other regulated sectors, safeguarding sensitive data with encryption and access controls.
- Customers build **diverse applications** and derive actionable insights from their data so they can **drive innovation across use cases.**

MongoDB | iguazio

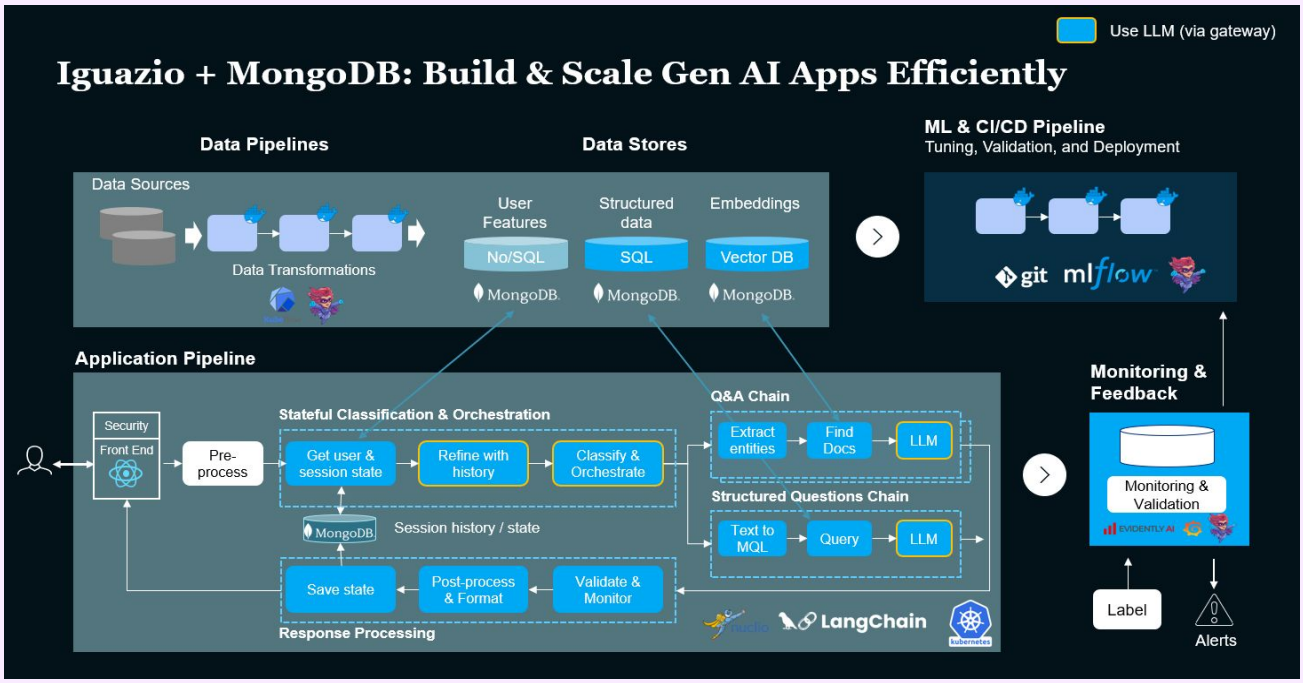


Figure 10: How to build and scale GenAI applications efficiently with MongoDB and Iguazio

MongoDB and Iguazio can be used for creating a smart customer care agent that documents call details, provides live contextual recommendations as a co-pilot, provides live agent support, customizes offers and recommendations and more.

First, the joint architecture processes and analyzes raw data (e.g., web pages, PDFs, images) inputted by the customer or the enterprise.

Then, the data is processed in a batch pipeline for analyzing customer logs and a stream pipeline for live interactions.

Finally, results are stored in MongoDB, leveraging its capabilities for managing unstructured data like user age, preferences and historical transactions, together with structured data like account balance and product lists.

Guardrails for Protecting Against LLM Risks

Iguazio eliminates LLM risks with guardrails that ensure:

- Fair and unbiased outputs
- Intellectual property protection
- PII elimination to safeguard user privacy
- Improved LLM accuracy and performance for minimizing AI hallucinations
- Filtering of offensive or harmful content
- Alignment with legal and regulatory standards
- Ethical use of LLMs

Basikon's credit and leasing platform that enables you to nurture your customer, employee and partner experiences

Basikon, a leader in financial technology, presents a powerful Software as a Service (SaaS) platform, transforming how financial institutions oversee loans, leases, guarantees, and wholesale financing. Leveraging cutting-edge technology and universal datacenters, Basikon processes millions of contracts daily, providing innovative solutions for financial institutions to improve collaboration with partners and customers.

Efficient Banking

The SaaS platform transforms financial management, enabling a front-to-back loan and lease digital system in months, not years.

Orchestrating and Enabling Software

The platform orchestrates digital journeys across front and back offices, managing financing product distribution directly or through partners, boosting agility and productivity.

Modern Technology Utilization

The company uses modern tech to streamline processes, cut approval wait times, and prevent errors, enhancing the customer experience with automated end-to-end processes.

Rapid Deployment

The model enables rapid system deployment. Despite the complexity, over 15 systems have been deployed since 2019, setting it apart in fintech.

Building a Scalable and Agile Financial Services Platform

- **Unmatched Scalability**: Basikon leverages MongoDB's horizontal scaling capabilities to efficiently handle massive volumes of data, crucial for their lending and leasing solutions.
- **Adaptable Data Model**: MongoDB's flexible data model provides the agility to adapt to ever-changing financial service requirements, enabling Basikon to innovate and evolve rapidly.
- **High-Speed Operations**: MongoDB's exceptional performance on big data makes it the ideal choice for Basikon, which manages millions of contracts daily.
- **Seamless Cloud Integration**: MongoDB's compatibility with cloud platforms aligns perfectly with Basikon's cloud-native approach, allowing for effortless integration and operation.
- **Streamlined Integrations**: MongoDB's robust API set simplifies integration with Basikon's pre-built integrations and external systems, ensuring a smooth flow of data.

MongoDB | Basikon

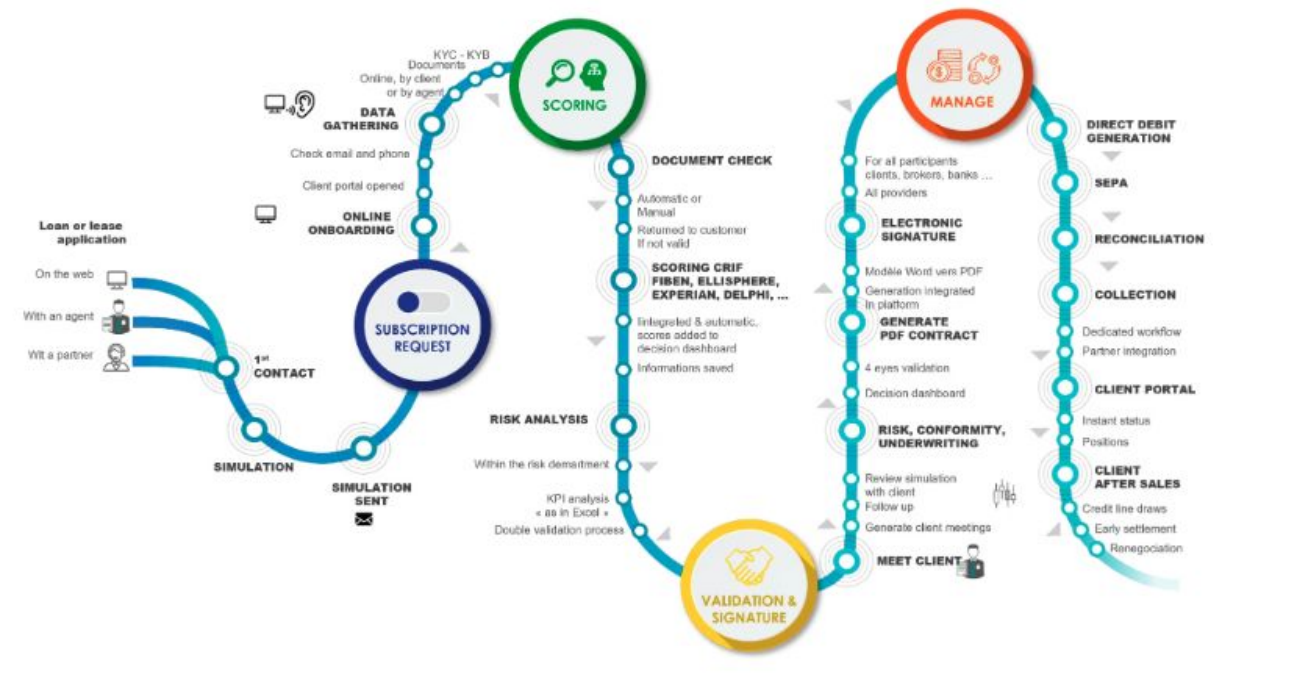


Figure 11: Basikon handling the complete customer life cycle

As the financing landscape evolves, there's a growing need for configurable, cloud-based software built on microservices. This type of software should be able to manage the entire life cycle of any financing product. Basikon, as illustrated in Figure 1, fulfills this need by orchestrating the digital journey across all stages, from initial customer interaction to loan approval and management. It also empowers financial institutions to manage their distribution networks and partner relationships directly through the platform.

Moreover, Basikon's cloud-based software architecture built on microservices ensures adaptability to changing market dynamics and regulatory requirements. By offering a comprehensive solution for managing the entire financing product life cycle, Basikon enables financial institutions to remain agile and responsive in today's competitive landscape. This integrated approach not only streamlines processes but also enhances transparency and accountability throughout the financing journey, ultimately driving greater customer satisfaction and loyalty.

"MongoDB Atlas is very stable – in 4 years, we did not experience a single interruption of service or find a single bug. Upgrades are done in seconds with just the press of a button, increasing our agility 10x. At Basikon, MongoDB has played a crucial role in our success and we wouldn't be where we are today without it."
Thomas Nokin, Founder and CEO at Basikon

Read the full customer story here.

Unlock Your Enterprise Content: Encore's AI-Enabled Cloud Solution Delivers Results

Remember all that messy, unavailable, unstructured data? Encore has turned that into a goldmine of knowledge. With Encore's AI-enabled platform, everything's ingested quickly, painlessly, and organized to meet business needs. Encore is a SaaS, cloud-native, enterprise content management platform which is a cost-effective solution for storing, retrieving & archiving business content. Encore's scalable repository leverages the power of MongoDB Atlas and the suite of AWS services making finding content easy. Layering in AI-enabled services opens up the possibilities of automation and efficiency to drive business growth.

AI-Enabled Services
Organizations are sitting on a treasure trove of data in the form of documents, emails, images, and other unstructured content. Encore enables businesses to tap into this data using AI technology.

Enterprise Scale and Performance
Seamless horizontal scaling & redundancy backed by the power and reliability of AWS & MongoDB. The Encore platform scales to meet whatever you need to optimally run your business when you need it.

Service and Event-Driven Architecture
Transparency and ease of integration are central to the Encore design principles. Tracking all platform events published and making them available through a suite of services provides insights that open up additional opportunities.

Secure, Complaint and Highly Available
Encore's model puts security & compliance in front of development & deployment. Their SOC2 and StateRAMP compliance demonstrate their commitment to security excellence. DR/HA removes any possibility of interruption to operations.

Choice of MongoDB as Developer Data Platform

Flexible Schema: Encore's is able to quickly adapt to unique requirements. With MongoDB's flexible schema, they can add changes to the platform without creating complexity.

Powerful Search: Businesses invest in content solutions with an expectation that users can quickly locate documents and get accurate answers to their open questions. MongoDB's Vector capabilities provide the foundation for all Semantic search and associated embeddings.

Enterprise Scale: Demand is met with vertical and horizontal scaling. The ease with which MongoDB provides scale-out to bring more nodes online or vertical scale to adjust processing power is ideal for Encore customers.

Ease of Use: MongoDB Atlas provides tools for the product team to efficiently build new features and operationalize the Encore solution, for example, aggregation pipelines for quick results to end users, or automatic failover in case of an outage.

MongoDB | encore

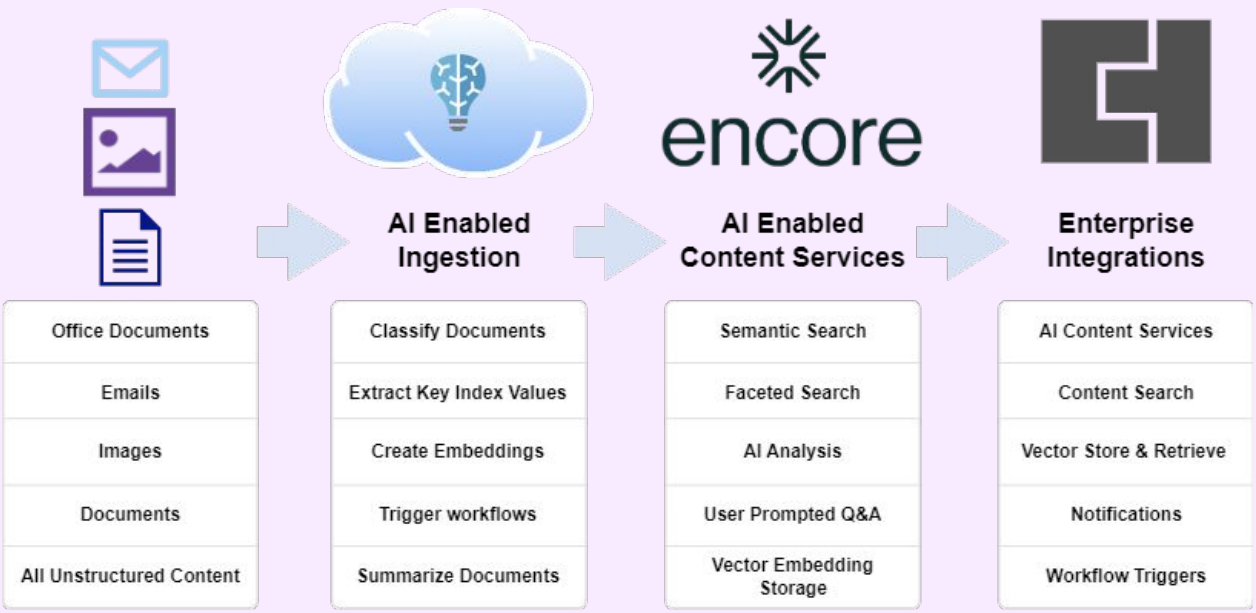


Figure 12: The value of Encore

Optimizing your Business with AI-enabled Content

With the foundation of Encore built on MongoDB Atlas, leveraging Semantic Search with Mongo's Vector database has never been so easy. Encore creates embeddings using the latest AI LLMs and stores them in MongoDB Atlas, reimagining the search capabilities on unstructured content. With AI-enabled services, customers can gain access to all the relevant data.

Document Summaries: Encore enables you to simply press a single button to get a summary of key points to help you quickly gain insight into the content of interest.

AI-Driven Analysis: Powering your organization with Encore translates to a number of operational efficiency wins. Organizations are regularly tasked with auditing archived content and expected to respond quickly. With Encore AI-enabled analysis, teams can use the collections feature to group content and provide prompts for the analysis. Encore's crawlers leverage AI and Vector search to locate the documents that meet the criteria and create summaries to simplify the findings.

Vector Search: With Encore's search, powered by MongoDB's Vector Database and AWS Bedrock, finding content with simple user-provided prompts will retrieve relevant content within seconds.

Workflow Automation: Leveraging the latest in AI technology with the Encore platform's API architecture opens up the opportunity for organizations to rethink expensive manual workflows. With Encore you can eliminate manual steps that require a review of specific content and instead automate utilizing our search to locate content, extract the data from the content that is relevant to your workflow criteria and expedite your customer requests.

How Cognigy Built a Leading Conversational AI Solution With MongoDB

Cognigy is a pioneering force in AI-driven customer service solutions on a global scale. They are at the forefront of revolutionizing the customer service industry by providing the most cutting-edge AI workforce on the market. Trusted by giants like Toyota, Bosch, and Lufthansa, their award-winning solution empowers businesses to deliver exceptional customer service: instant, personalized, in any language, and on any channel.

AI-Driven Customer Service

Their main product, Cognigy.AI, allows companies to create AI Agents, improving experiences through smart automation and natural language processing. This makes it easy for businesses to develop and deploy intelligent voice and chatbots.

Drag-and-Automate AI

Cognigy's low-code platform lets business users build virtual agents with drag-and-drop tools like Flows, Playbooks, and Lexicons.

Integration with Third-Party Platforms

Cognigy makes it simple to integrate with third-party platforms like Facebook Messenger, Line, and WhatsApp. This broadens the reach of customer service teams and helps businesses connect with their audience on various channels they use.

Enterprise-Level Security and Compliance

Cognigy prioritizes security by offering features that comply with industry standards like SOC 2, GDPR, CCPA, and HIPAA.

Seamless Integration & Peak Performance

- MongoDB's **JSON document storage** aligns perfectly with Cognigy's application language, facilitating seamless integration with Typescript and intuitive querying processes
- MongoDB's **scalability** via sharding aligns with Cognigy's growth vision, enabling expansion across cloud providers and on-premises setups.
- MongoDB's **developer data platform** empowered Cognigy to efficiently manage diverse data types, ensuring peak performance under high loads.
- MongoDB empowered Cognigy.AI to handle **expanding user interactions** while maintaining peak performance, ensuring scalability and responsiveness in scaling conversational agents.
- MongoDB's **document model flexibility** enables easy data model modifications, reducing concerns about data and schema migrations.

MongoDB | COGNIGY

Figure 13: Cognigy's replica-sets in production

Have you ever built a chatbot that struggled to keep up with user demands? Imagine a platform that can handle hundreds of queries per second, even during peak hours, all while storing massive amounts of data. That's the power of MongoDB at work for Cognigy.AI!

Cognigy constructed the platform by employing a composable architecture model with over 30 specialized microservices, which they adeptly orchestrated through Kubernetes. These microservices were strategically fortified with MongoDB's replica-sets, spanning across three availability zones, a move aimed at bolstering reliability and fault tolerance.

As you can see in Figure 1 "Cognigy's replica-sets in production", MongoDB's magic isn't just marketing hype. This tech allows Cognigy.AI to effortlessly manage a growing number of user interactions, processing all sorts of data without breaking a sweat. MongoDB's data model is like a chameleon, constantly adapting. Imagine your chatbot being able to learn and improve over time, just like a chameleon adapting to its surroundings. This is what MongoDB's flexible data model enables for Cognigy.AI. As new data and user interactions flow in, Cognigy.AI can continuously update and refine its understanding of how to best serve your customers. This collaboration is a prime example of how powerful technology can be the driving force behind groundbreaking products like Cognigy.AI. Imagine the possibilities: chatbots that can provide personalized recommendations, troubleshoot complex issues, or even have engaging conversations.

How Devnagri Brings the Internet to 1.3 Billion People with Machine Translations

Devnagri is India's leading AI-powered translation engine, enabling brands to localize content five times faster and more accurately. As a SaaS platform, it focuses on translating Indian languages, utilizing a hybrid approach of 80% machine and 20% human effort to achieve 99% accuracy in translating millions of words daily.

Customizable AI models

Devnagri trains machine translation models with data stored in MongoDB Atlas, achieving real-time translation.

Tackling the digital divide in India

This AI platform offers machine translation for non-English speakers, focusing on e-learning, banking, e-commerce, and media.

Adapting to future advancements

Devnagri considers using advanced models like OpenAI GPT-4 and Llama-2-7b, fine-tuned with their own translation data.

Human-in-the-loop approach

Devnagri integrates human feedback (RLHF) to enhance translation accuracy, emphasizing their dedication to both automation and human expertise for quality control.

Devnagri's Strategic Use of MongoDB for Machine Translation

- **Flexible data model**: MongoDB's document data model suits Devnagri's need to store diverse structured and unstructured content efficiently for training their machine translation models.
- **Faster time to market**: Efficiency aids Devnagri in faster training and improved translation quality, accelerating product launches.
- Supports **real-time needs**: Devnagri uses data stored in MongoDB to train models for real-time translation.
- Access to **expertise**: Being part of MongoDB's AI Innovators Program grants Devnagri technical guidance and best practices, aiding their development process.
- **Scalability for performance**: The distributed architecture of MongoDB allows Devnagri to parallelize tasks across multiple machines, improving training and translation speed.

MongoDB | devnagri

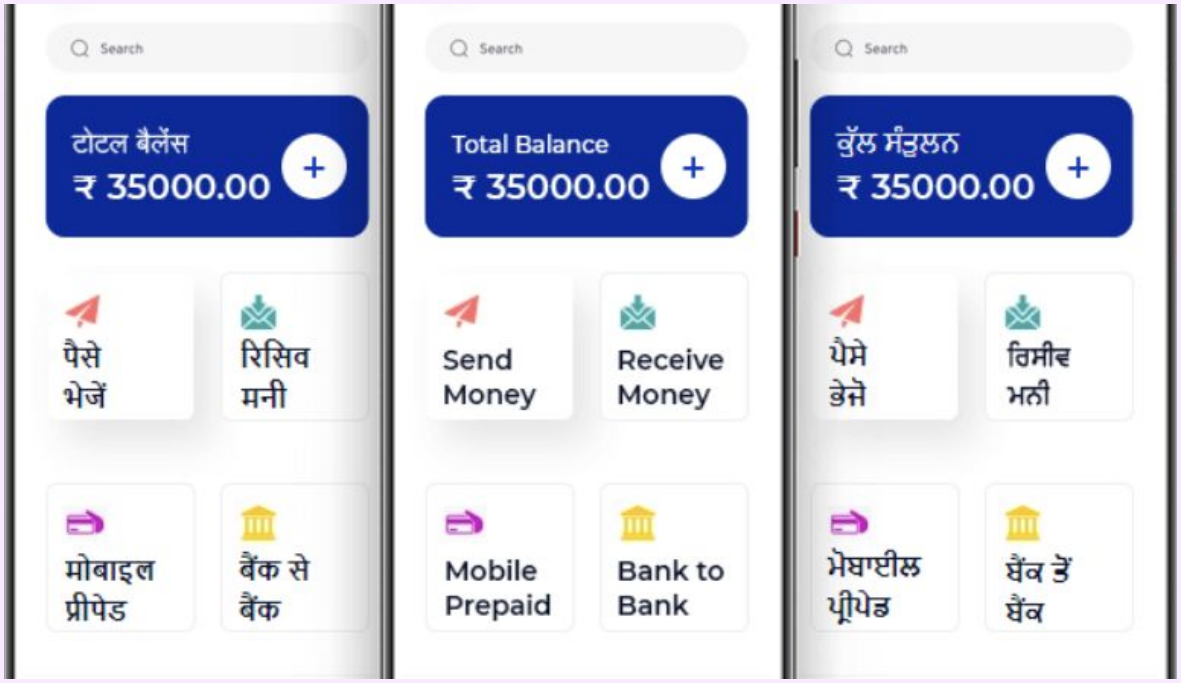


Figure 14: A visual of Devnagri's real time translation engine

Devnagri's real time translation engine helps over 100 Indian brands connect with their customers over digital channels for the first time

The real-time translation engine has helped over 100 Indian brands connect with their customers over digital channels for the first time. This achievement signifies a breakthrough in overcoming the language barrier in India, where 90% of the population are not fluent in English, and more than 22 Indian languages are in use.

The platform's focus spans diverse industries such as e-learning, banking, e-commerce, and media publishing, offering a tailored solution beyond a general consumer tool. Powered by custom transformer models and advancements like OpenAI GPT-4, Devnagri's technology strives to democratize internet access for India's non-English speakers.

WINN.AI: The virtual assistant tackling sales admin overhead

WINN.AI is more than just a tool; it's a productivity powerhouse designed to transform the way sales teams operate. By reducing administrative busywork, WINN.AI is helping organizations save time, money, and resources, enabling sales teams to better invest their working hours in serving customers.

AI-Powered Sales Assistant

An AI-powered real-time sales assistant joins virtual meetings, understands conversation context, and responds to customer queries, enabling salespeople to focus on selling rather than administrative tasks.

Sales Playbook Prompts

WINN.AI can provide prompts from a sales playbook, helping to guide the salesperson during customer interactions. It also ensures meetings stay on track and on time.

Contextual Understanding

The platform understands conversation context, providing real-time relevant information to the salesperson, including customer references and competitive data.

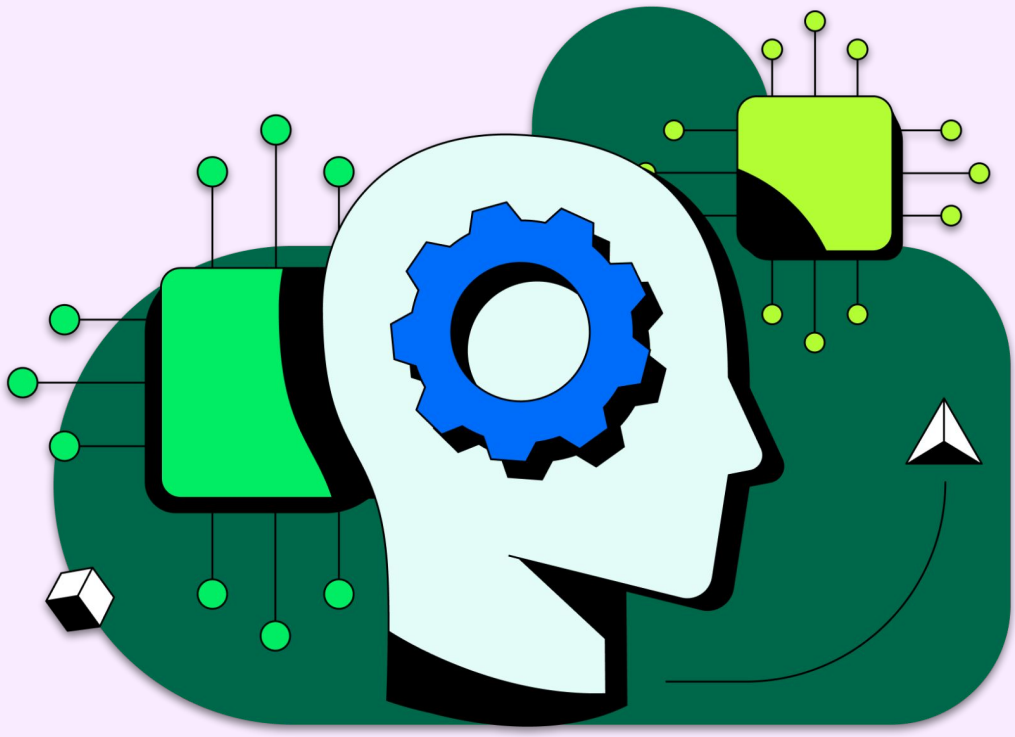
CRM Integration

After each meeting, WINN.AI extracts and summarizes relevant information, updating the CRM system with follow-on actions, eliminating manual data entry, saving time, and reducing errors.

Building a Strong Foundation for AI

- **Developer Familiarity:** The developers at WINN.AI are familiar with MongoDB, eliminating the need for database administrators or external experts and enabling the team to focus on building AI-powered products.
- **Flexibility**: MongoDB's flexibility allows WINN.AI to handle data of any form, offering agility surpassing traditional relational databases.
- **Managed Services**: MongoDB Atlas provides WINN.AI with managed services for running, scaling, securing, and backing up their data, simplifying the tech stack and ensuring data safety.
- **Cost Efficiency**: By using MongoDB, WINN.AI can invest the savings from not needing any DBA or external experts back into building great AI-powered products.
- **Stability**: In the ever-changing AI tech market, MongoDB serves as a stable anchor for WINN.AI. This allows the developers to freely create with AI while being able to maintain a reliable data infrastructure.

MongoDB | WINN AI



WINN.AI: The virtual assistant tackling sales admin overhead

Beyond simply attending meetings, WINN.AI empowers salespeople by automating tedious administrative tasks. After each virtual encounter, WINN.AI intelligently summarizes key points and automatically updates the CRM system with follow-up actions. This eliminates the need for manual data entry, saving salespeople valuable time and minimizing errors.

Furthermore, WINN.AI boasts a powerful AI architecture. Initially built on custom NLP algorithms, the system now utilizes the advanced capabilities of GPT 3.5 and 4 for superior entity extraction and summarization. This ensures salespeople have the most relevant information at their fingertips during crucial customer interactions.

Additionally, WINN.AI seamlessly integrates with leading sales tools like Zoom, HubSpot, and Salesforce, for a streamlined workflow.

MongoDB. | WINN AI

Ada: Revolutionizing customer service with AI-powered automations built on MongoDB Atlas

Since 2016, Ada has become a dominant force in AI, reshaping customer service with its intelligent automation engine. Their AI swiftly resolves complex inquiries across any channel, in any form. Backed by nearly $200 million in funding and a team of 300 passionate innovators, Ada empowers over 300 industry leaders, including tech titans like Meta, Verizon, and AT&T, to deliver exceptional customer experiences.

AI-Powered Automations

Ada's advancements in transformer models, LLMs, and RLHF have significantly enhanced their AI assistants, enabling advanced reasoning to solve customer problems rather than just searching for information.

Efficient Use of Unstructured Data

They can query unstructured data and use it to train other models, enabling them to automate queries and provide support that goes beyond just answering multi-step queries.

Rapid Product Development

Ada prioritizes rapid product development, measured by the speed of shipping products and features, as well as the pace of learning and iterating. They can deliver new products in just a few months.

Impressive Track Record

Since 2016, Ada has powered more than 4 billion automated customer interactions for brands like Wealthsimple, Verizon, AirAsia, Yeti, and Square.

Unmatched Performance and Support: Keeping Ada Ahead

- **Flexibility and Agility**: Ada can easily scale their database as their business grows and adapt to new channels and modalities without being restricted by their database infrastructure.
- **Performance and Support**: Ada has found that the performance of MongoDB Atlas meets their needs, and they appreciate the great support from the MongoDB team.
- **Less Dependency on One Central Cloud Vendor**: By using a cloud-agnostic solution, Ada avoids being locked into a single cloud provider. This gives them more freedom and flexibility.
- **Distributed Event Processing System:** Ada is using MongoDB Change Streams to build a distributed event processing system that powers bots and analytics.

MongoDB | ada

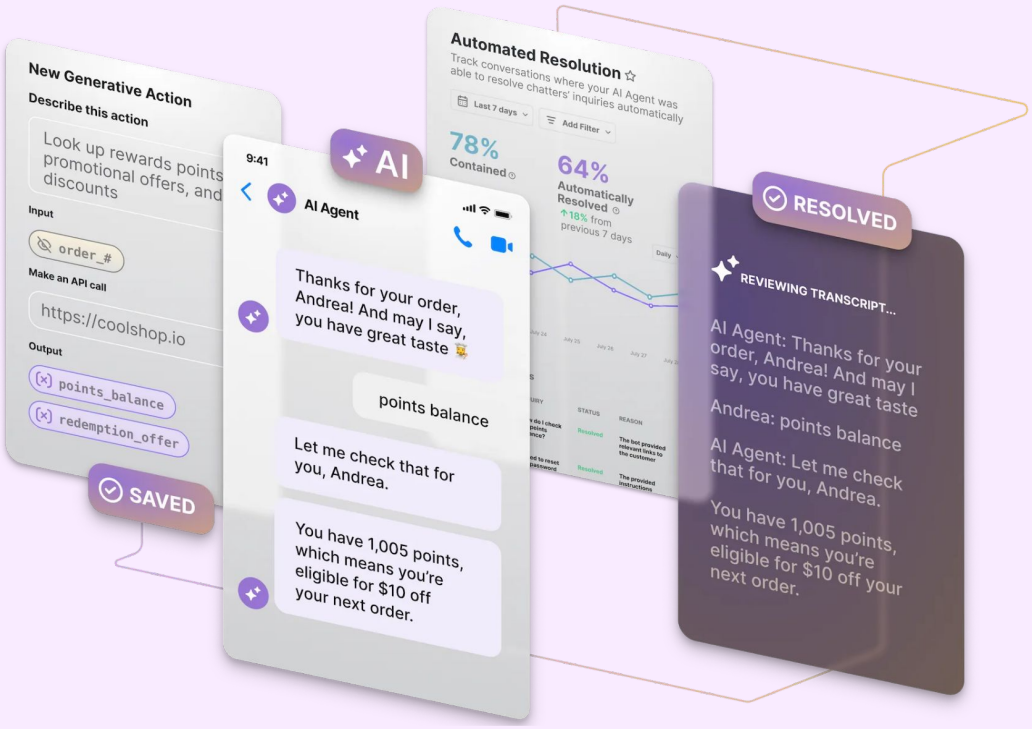


Figure 15: Ada's AI customer service

Beyond Automation, A Self-Learning AI for Superior Customer Service

Ada's focus on cutting-edge AI extends beyond just solving customer problems. They can automate tasks and provide advanced support by querying unstructured data, such as customer conversations. This allows them to train additional models that go beyond answering even complex, multi-step queries. This translates to a superior customer experience as Ada can automate more interactions and provide more comprehensive support.

Furthermore, Ada prioritizes rapid development, allowing them to deliver new features and products in just a few months. This agility ensures they stay ahead of the curve in customer service innovation. In essence, Ada is creating a self-learning AI loop that continuously improves customer service through automation and data-driven insights.

XOLTAR: GenAI companion for patient engagement and better clinical outcomes

XOLTAR is a pioneering conversational AI platform designed to foster long-lasting patient engagement. It provides an AI-powered accountability partner platform that mimics the one-on-one interactions nurses conduct with patients. Through personalized encounters, these AI companions guide patients toward adopting healthy habits necessary to manage their medical conditions.

AI Accountability Partner Platform

Xoltar provides an AI accountability partner platform that emulates the one-on-one interactions nurses conduct with patients.

Customizable AI Partners

Each AI partner can be customized by gender, race, and language, promote goals, monitor patients, collect and report video & audio RWE, and offer an emotionally engaged experience.

Hyper-Personalized Encounters

Through hyper-personalized encounters, the accountability partners lead patients to embrace the healthy habits required to manage their medical conditions.

Sensor Fusion Technology

Xoltar's sensor fusion technology interprets human emotion from facial expressions, voice patterns, context, and other non-verbal cues.

Powering Patient Care with Real-Time Data and Machine Learning

- **Long-term Memory and Model Training**: The data stored in MongoDB provides both long-term memory for each patient as well as input for ongoing model training and tuning.
- **Event-Driven Data Pipelines**: MongoDB powers XOLTAR's event-driven data pipelines. Follow-on actions generated from patient interactions are persisted in MongoDB.
- **Real-Time Notifications**: With Atlas Triggers, MongoDB notifies downstream consuming applications so they can react in real-time to new treatment recommendations and regimes.
- **Real-Time Interaction Management**: XOLTAR can manage patient interactions in real-time, thanks to the database. This is crucial for their omni-channel approach to patient care.
- **Support for Machine Learning Models:** MongoDB provides the necessary data for training and fine-tuning XOLTAR's sophisticated array of machine learning models.

MongoDB | XOLTAR

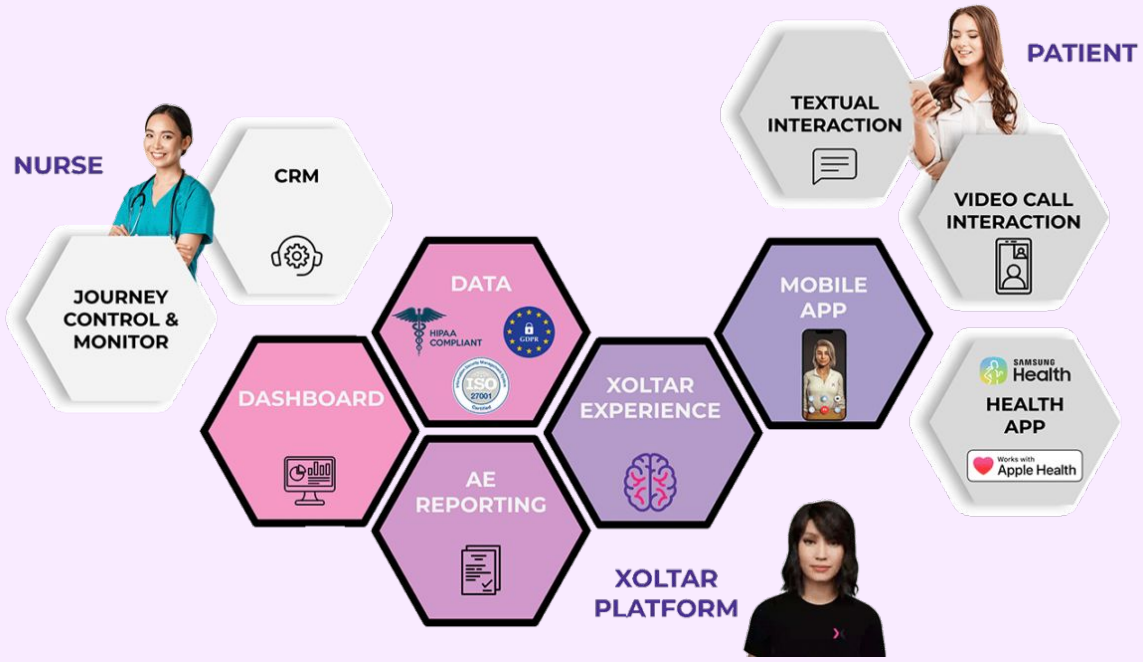


Figure 16: How the XOLTAR platform works

At the Heart of XOLTAR lies a sophisticated array of state-of-the-art machine learning models working across multiple modalities — voice and text, as well as vision for visual perception of micro-expressions and non-verbal communication. These custom multilingual models are trained and deployed to create a truthful, grounded, and aligned free-guided conversation, along with various transformers for real-time automatic speech recognition.

XOLTAR's models personalize each patient's experience by retrieving data stored in MongoDB Atlas. Taking advantage of the flexible document model, XOLTAR developers store both structured data, such as patient details and sensor measurements from wearables, alongside unstructured data, such as video transcripts. This data provides both long-term memory for each patient as well as input for ongoing model training and tuning.

MongoDB also powers XOLTAR'S event-driven data pipelines. Follow-on actions generated from patient interactions are persisted in MongoDB, with Atlas Triggers notifying downstream consuming applications so they can react in real-time to new treatment recommendations and regimes.

Through its participation in the MongoDB AI Innovators program, XOLTAR's development team receives access to free Atlas credits and expert technical support, helping them de-risk new feature development.

Conversation Intelligence with Observe.AI

Observe.AI, a California-based company funded by over $200 million, is the leading provider of live conversation intelligence for contact centers. Trusted by industry leaders like Accolade and Pearson, Observe.AI empowers businesses to transform the way they interact with customers. The company is focused on being the fastest way to boost contact center performance with live conversation intelligence.

Advanced AI Techniques
Observe.AI employs AI techniques, including transformers for NLP, for various tasks like text classification, intent recognition, summarization, and question-answering.

Speech Processing Expertise
Observe.AI goes beyond NLP into speech processing, using cutting-edge methods for tasks like automatic speech recognition and sentiment analysis to keep their language capabilities leading-edge.

Model Development and Training
Observe.AI uses TensorFlow and PyTorch to craft and fine-tune intricate natural language models, employing transfer learning and gradient-based optimization techniques.

Efficient Operationalization
Observe.AI optimizes MLOps with Docker and Kubernetes, enabling smooth model deployment, management, and scalability.

The role of MongoDB in Observe.AI technology stack

The MongoDB developer data platform gives the company's developers and data scientists a unified solution to build smarter AI applications.

"OBSERVE.AI processes and runs models on millions of support touchpoints daily to generate insights for our customers. Most of this rich, unstructured data is stored in MongoDB. We chose to build on MongoDB because it enables us to quickly innovate, scale to handle large and unpredictable workloads, and meet the security requirements of our largest enterprise customers."

Jithendra Vepa, Ph.D, Chief Scientist & India General Manager at Observe.AI

MongoDB | OBSERVE·AI

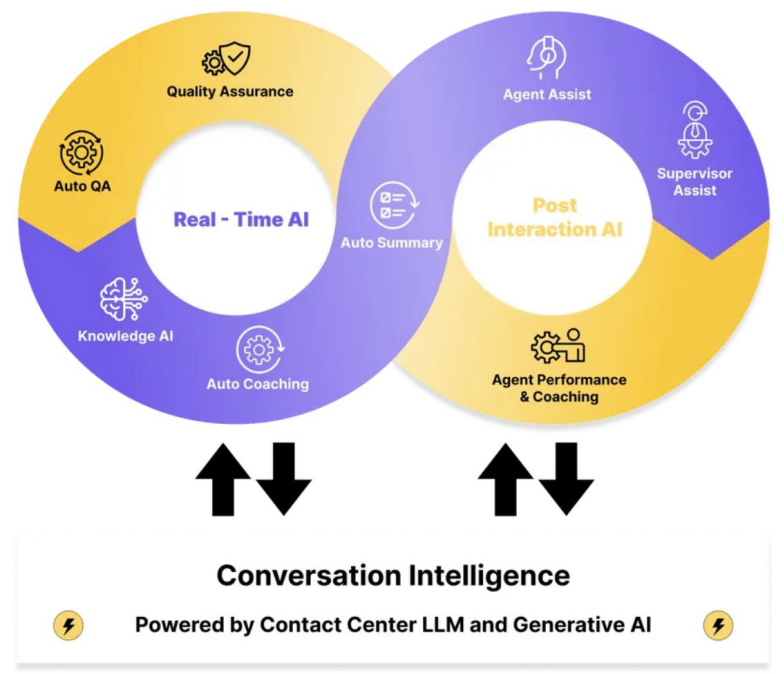


Figure 17: Observe.AI's conversation intelligence

Boost Sales and Support Teams: Data-Driven Insights from Observe.AI

The company has pioneered a 40 billion-parameter contact center large language model (LLM) and one of the industry's most accurate Generative AI engines. Through these innovations, Observe.AI provides analysis and coaching to maximize the performance of its customers' front-line support and sales teams.

Observe.AI's advanced AI tools analyze conversation data thoroughly, revealing key insights like emotions and sentiment. This helps businesses identify areas for improvement and provides targeted coaching for exceptional customer service.

"Our products employ a versatile range of AI and ML techniques, covering various domains. Within natural language processing (NLP), we rely on advanced algorithms and models such as transformers, including the likes of transformer-based in-house LLMs, for text classification, intent and entity recognition tasks, summarization, question-answering, and more. We embrace supervised, semi-supervised, and self-supervised learning approaches to enhance our models' accuracy and adaptability."

Jithendra Vepa, Ph.D, Chief Scientist & India General Manager at Observe.AI

How Flagler Health's AI-Powered Journey is Revolutionizing Patient Care

Flagler Health is dedicated to supporting patients with chronic diseases by matching them with the right physician for the right care. Typically, patients grappling with severe pain conditions face limited options, often relying on prolonged opioid use or exploring costly and invasive surgical interventions. Unfortunately, the latter approach is not only expensive but also has a long recovery period. Flagler finds these patients and triages them to the appropriate specialist for an advanced and comprehensive evaluation.

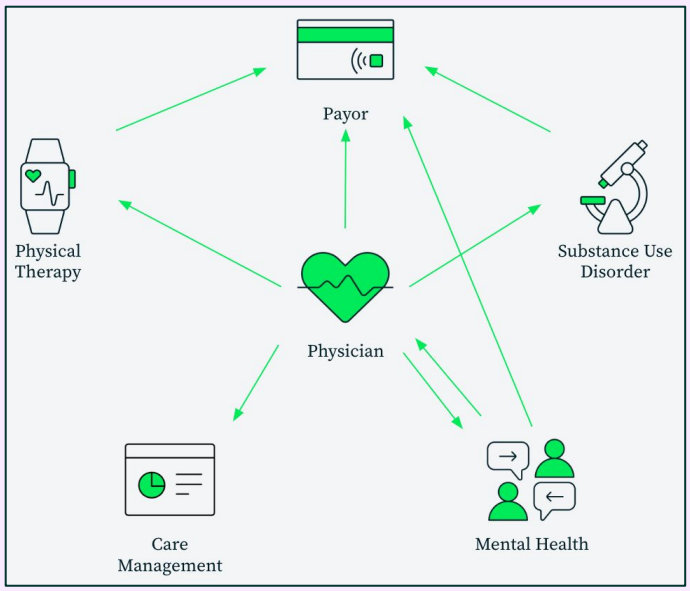
Flagler Health employs sophisticated AI techniques to rapidly process, synthesize, and analyze patient health records to aid physicians in treating patients with advanced pain conditions. This enables medical teams to make well-informed decisions, resulting in improved patient outcomes with an accuracy rate exceeding 90% in identifying and diagnosing patients.

As the company built out its offerings, it identified the need to perform similarity searches across patient records to match conditions. Flagler's engineers identified the need for a vector database but found standalone systems to be inefficient. They decided to use MongoDB Atlas Vector Search.

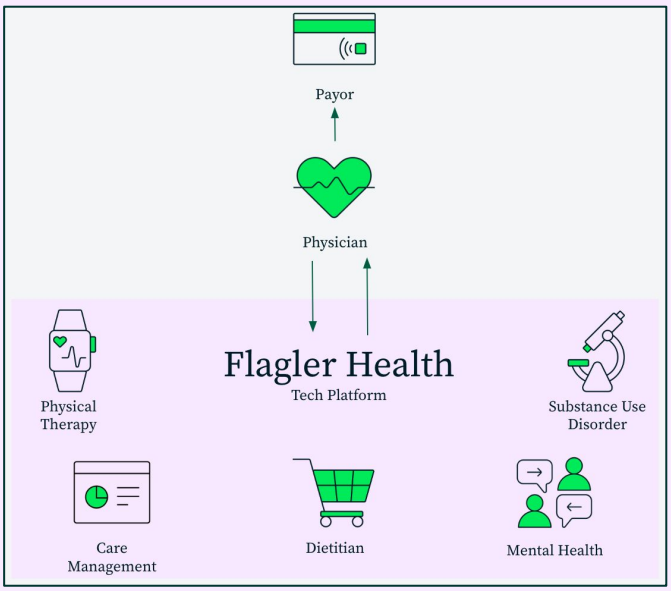
Creating an integrated platform to store all data in a single location with a unified interface, facilitating quick access and efficient data querying.

- Flagler Health emphasized the importance of a **flexible database that can evolve with the company's growth**. A relational model was deemed too rigid, leading the company to choose MongoDB's document model.
- MongoDB's flexibility allows for **easy customization of client configuration files**, streamlining data editing and evolution.
- The managed services provided on MongoDB's developer data platform **save time** and **offer reliability at scale** throughout the development cycle.
- With Atlas Vector Search, developers can **build AI-powered experiences** while accessing all the data they need through a **unified and consistent developer experience**..

MongoDB | FLAGLER HEALTH



Current state without Flagler Health



What Flagler Health can offer

Flagler Health collaborates with many clinics, first processing millions of electronic health record (EHR) files in Databricks and transforming PDFs into raw text. Using the MongoDB Spark Connector and Atlas Data Federation, the company seamlessly streams data from AWS S3 to MongoDB. Combined with the transformed data from Databricks, Flagler's real-time application data in MongoDB is used to generate accurate and personalized treatment plans for its users. MongoDB Atlas Search facilitates efficient data search across Flagler Health's extensive patient records. Beyond AI applications, MongoDB serves critical functions in Flagler Health's business, including its web application and patient engagement suite, fostering seamless communication between patients and clinics.

This comprehensive application architecture, consolidated on MongoDB's developer data platform, simplifies Flagler Health's operations, enabling efficient development and increased productivity. By preventing administrative loops, the platform ensures timely access to potentially life-saving care for patients.

Looking ahead, Flagler Health aims to enhance patient experiences by developing new features, such as a digital portal offering virtual therapy and mental health services, treatment and recovery tracking, and a repository of physical therapy videos. Leveraging MongoDB's AI Innovators program for technical support and free Atlas credits, Flagler Health is rapidly integrating new AI-backed functionalities on the MongoDB Atlas developer data platform to further aid patients in need.

MongoDB. | FLAGLER HEALTH

Dataworkz: Generate Faster Data Insights with GenAI Apps & Proprietary Data

The Dataworkz GenAI applications platform provides an all-in-one RAG as a Service to rapidly build, deploy, operationalize and scale GenAI applications, and eliminates the complexity involved in building reliable and scalable RAG applications. It includes advanced search and retrieval to provide relevant context to LLMs, and monitoring with traceability to observe and optimize application performance.

Visual RAG builder

No-code AI app development, with a knowledge graph, and lexical and semantic search, plus frictionless data wrangling, to create GenAI apps.

Composable AI stack

Configure with your existing or new technologies, with access to metrics, insights and elastic deployment.

End-to-end traceability

An integrated, highly performant platform with comprehensive visibility into the underlying instrumentation and transactions.

Expand GenAI app adoption

Implement additional use cases, connect new data sources and use RAG APIs to embed GenAI in workflows easily, efficiently and securely.

MongoDB + Dataworkz | The Power of Combined Innovation

- For Generative AI applications, Dataworkz argues that a company's key differentiator, or "superpower," lies in **enhancing underlying Large Language Models (LLMs)** with its own well-managed data.
- To easily access diverse internal data in MongoDB Atlas, Dataworkz is used—a comprehensive RAG development platform. Its **Composable AI stack**, **hybrid search**, **end-to-end traceability**, and **no-code data** transformation enhance GenAI applications.
- Companies use the state-of-the-art MongoDB Atlas technology to deliver their AI-enriched apps with the **right security** controls in place, and at the **scale** and **performance users expect**.

MongoDB | dataworkz

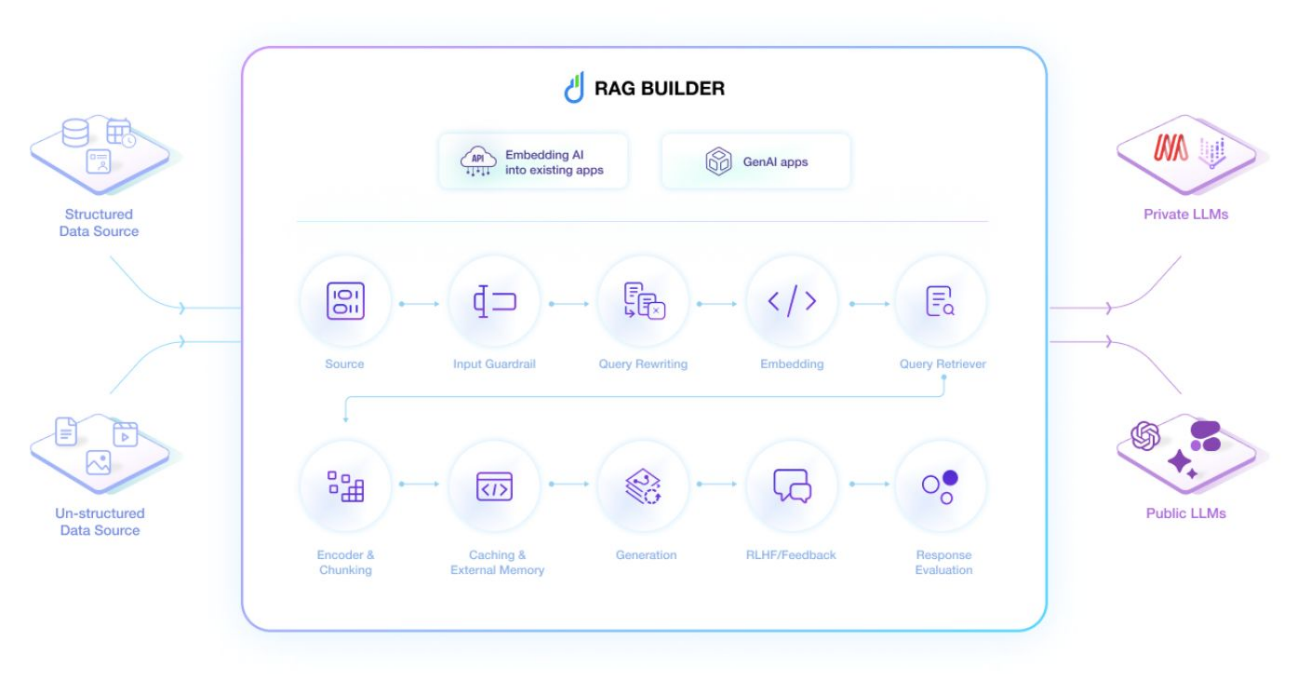


Figure 18: A leading-edge platform to rapidly build, deploy, operationalize and scale GenAI applications

Why Dataworkz: Unique capabilities for GenAI applications using RAG

Build

Visually Create GenAI Applications: Develop GenAI applications using a visual RAG builder. This eliminates the need to worry about the complexity of the underlying infrastructure.

Smart Routing with Knowledge Graph: Set up smart routing using a knowledge graph for lexical and semantic search.

Optimize

Comprehensive Visibility & Customization: Achieve complete AI stack transparency and easily customize data processing steps in a user-friendly, no-code interface.

Data-Driven Optimization: Conduct A/B testing on RAG pipelines using built-in evaluation metrics to determine the most effective configurations.

Observe

End-to-End Traceability: Get full visibility of your GenAI apps with end-to-end traceability for better performance optimization.

Centralized Monitoring: Track all system activity, like LLM calls, SLM calls, indexing, and retrieval, with one unified tool.

Scale

Rapid Application Development: Build diverse GenAI applications efficiently by utilizing pre-defined templates for various use cases.

Embeddable RAG: Integrate GenAI apps with Slack, Azure Studio, and HTML widgets via RAG APIs for enhanced accessibility in workflows.

MongoDB. | dataworkz

VISO TRUST: Transforming cyber risk intelligence

VISO TRUST is an AI-powered platform that helps companies quickly assess the cybersecurity risk of their vendors. It provides actionable security information in minutes, allowing businesses to make informed decisions with ease. VISO TRUST boasts a 90% reduction in workload and an 80% faster risk assessment process, with near-universal vendor adoption by their clients.

Automated Risk Management

VISO TRUST uses AI to streamline third-party risk assessments, enabling instant evaluation without extra analysts. It eliminates lengthy questionnaires and manual document analysis for a more efficient approach.

Artifact Intelligence

Curated AI extracts insights from source artifacts, automatically determining vendor security posture. This frictionless due diligence process simplifies assessing any number of third parties.

Risk Insights

On the platform, users can gain a comprehensive overview of their organization's cyber risk posture, enabling them to make data-driven decisions to reduce risk across all third-party relationships.

Compliance Excellence

Continuously exceeding ISO, NIST, AICPA, and other standards without impeding business operations is made possible by VISO TRUST. It empowers organizations to take control of their third-party security posture.

Empowering Customers with Faster Insights

- VISO TRUST deploys discriminator models that produce **high-confidence predictions** about features of the artifact.
- The artifacts undergo a process where their text content is extracted and integrated into MongoDB Atlas, thus becoming **integrated** into the dense retrieval system. This system executes Retrieval-Augmented Generation (RAG) by leveraging MongoDB functionalities such as Atlas Vector Search. Its aim is to furnish **ranked context to prompts** for large language models (LLMs).
- The outcomes of RAG serve as the foundation for seeding LLM prompts and linking their outputs in a chain, resulting in the generation of **highly precise factual details** regarding the artifact in the pipeline. This data facilitates the **swift delivery of intelligence** to customers, a task that previously required weeks to accomplish.

MongoDB | VISO TRUST

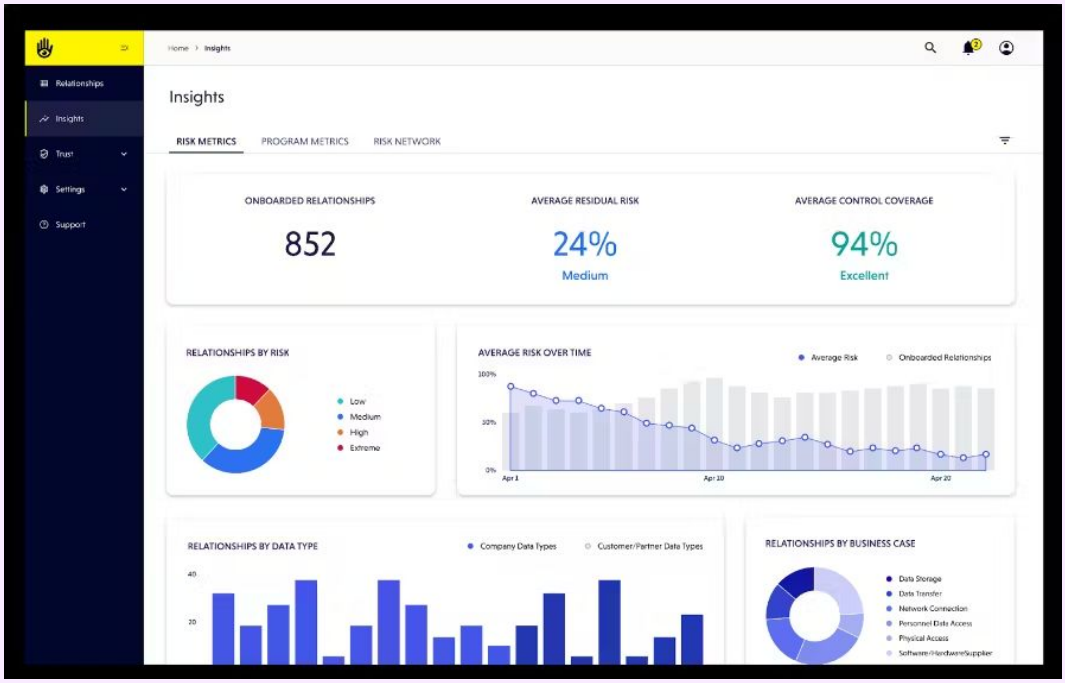


Figure 35: Insights dashboard

Streamlining Third-Party Cyber Risk Management

VISO TRUST is the only SaaS third-party cyber risk management platform that delivers the rapid security intelligence needed for modern companies to make critical risk decisions early in the procurement process.

VISO TRUST uses state-of-the-art models from OpenAI, Hugging Face, Anthropic, Google, and AWS, augmented by vector search and retrieval from MongoDB Atlas. Read our interview blog post with VISO TRUST to learn more.

How DevRev is Redefining CRM for Product-Led Growth

OneCRM from DevRev is purpose-built for Software-as-a-Service (SaaS) companies. It brings together previously separate customer relationship management (CRM) suites for product management, support, and software development. Built on a foundation of customizable large language models (LLMs), data engineering, analytics, and MongoDB Atlas, it connects end users, sellers, support, product owners, and developers. OneCRM converges multiple discrete business apps and teams onto a common platform.

The multi-cloud architecture of Atlas provides flexibility and choice that proprietary offerings from the hyperscalers can't match. While DevRev today runs on AWS, in the early days of the company, they evaluated multiple cloud vendors. Knowing that MongoDB Atlas could run anywhere gave them the confidence to make a choice on the platform, knowing they would not be locked into that choice in the future.

DevRev manages critical customer data, and so relies on MongoDB Atlas' native encryption and backup for data protection and regulatory compliance. The ability to provide multi-region databases in Atlas means global customers get further control over data residency, latency, and high availability requirements.

CRM + AI: Digging into the stack

DevRev's Support and Product CRM serve **over 4,500 customers**:

- Support CRM brings support staff, product managers, and developers onto an AI-native platform to automate Level 1 (L1), assist L2, and elevate L3 to become true collaborators.
- Product CRM brings product planning, software work management, and product 360 together so product teams can assimilate the voice of the customer in real-time.

AI is central to both the Support and Product CRMs. The company's engineers build and run their own neural networks, fine-tuned with application data managed by MongoDB Atlas.

This data is also encoded by open-source embedding models where it is used alongside OpenAI models for customer support chatbots and question-answering tasks orchestrated by autonomous agents. MongoDB partner LangChain is used to call the models, while also providing a layer of abstraction that frees DevRev engineers to effortlessly switch between different generative AI models as needed.

Data flows across DevRev's distributed microservices estate and into its AI models are powered by MongoDB change streams. Downstream services are notified in **real-time** of any data changes using a fully **reactive, event-driven architecture.**

MongoDB | DevRev

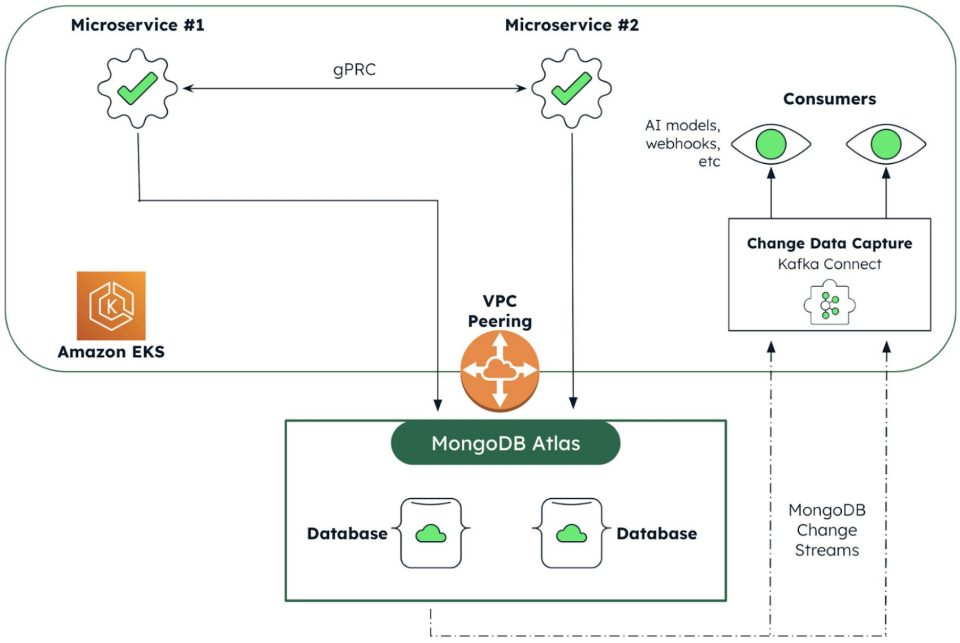


Figure 19: Event-driven microservices architecture for DevRev's AI-powered CRM platform

MongoDB Atlas: AI-powered CRM on an agile and trusted data platform

MongoDB is the primary database backing OneCRM, managing users, customer and product data, tickets, and more. DevRev selected MongoDB Atlas from the very outset of the company. The flexibility of its data model, freedom to run anywhere, reliability and compliance, and operational efficiency of the Atlas managed service all impact how quickly DevRev can build and ship high-quality features to its customers.

The flexibility of the document data model enables DevRev's engineers to handle the massive variety of data structures their microservices need to work with. Documents are large, and each can have many custom fields. To efficiently store, index, and query this data, developers use MongoDB's Attribute pattern and have the flexibility to add, modify, and remove fields at any time.

The freedom to run MongoDB anywhere helps the engineering team develop, test, and release faster. Developers can experiment locally, then move to integration testing, and then production — all running in different environments — without changing a single line of code. This is core to DevRev's velocity in handling over 4,000 pull requests per month:

- Developers can experiment and test with MongoDB on local instances — for example adding indexes or evaluating new query operators, enabling them to catch issues earlier in the development cycle.
- Once unit tests are complete, developers can move to temporary instances in Docker containers for end-to-end integration testing.
- When ready, teams can deploy to production in MongoDB Atlas.

Elevating the edge experience: Deploy AI anywhere with Cloneable and MongoDB

Cloneable provides the application layer that brings AI to any device at the edge of the network. The Cloneable platform empowers developers to craft dynamic applications using intuitive low/no-code tools, instantly deployable to a spectrum of devices - mobiles, IoT devices, robots, and beyond.

Component-Based Development

Cloneable apps are built using components, ranging from simple logic to complex data processing. These components allow you to construct applications that solve real-world problems by layering them together in the app builder.

Augmented Reality (AR)

Cloneable's AR component, empowers users to interact with field assets in real time. Whether navigating to a specific location or identifying an asset for inspection, AR enhances the user experience.

AI Object Detection

Cloneable provides an AI model for object detection. You can process input images from video previews or captured photos, and the model detects objects, outputting bounding boxes and relevant statistics based on business rules.

GIS Mapping

Cloneable leverages ESRI technology to enable smart, data-driven mapping styles. With intuitive analysis tools, you can gain location intelligence across field assets.

Real-Time Operational Tracking and Analysis

- Cloneable integrates seamlessly with MongoDB Atlas Device Sync, enabling the persistence of data locally on devices and its synchronized transfer to the cloud-based Atlas database. This **ensures** that enterprises can **track, measure, and respond** to events across their operations in r**eal-time**.
- Utilizing Cloneable and Atlas Vector Search to generate vector embeddings from images and device data enables users to e**fficiently search and analyze** field-collected events, thereby **enhancing decision-making and insights.**

MongoDB | CLONEABLE

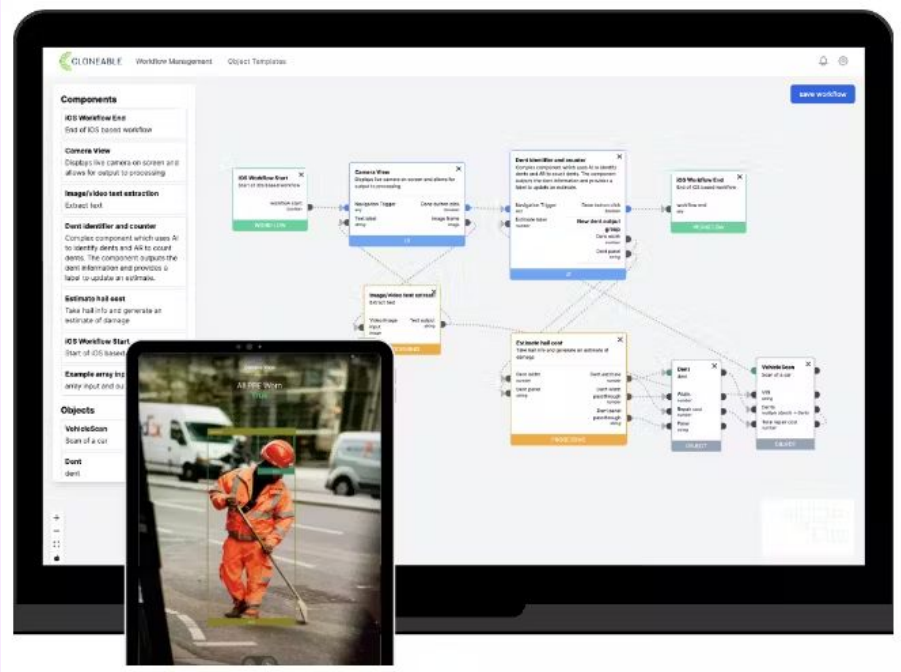


Figure 22: Cloneable components

Empowering Businesses with Efficiency, Personalization, and Growth

By harnessing machine learning models, a business can seamlessly leverage complex technologies across its operations. Models are pushed down to the device where they are converted to a native embedded format such as CoreML. From here, they are executed by the device's neural engine to provide low latency inference, computer vision, and augmented reality.

In addition to the operational efficiency gained through machine learning models, businesses also benefit from enhanced personalization and customer engagement. These models enable companies to analyze vast amounts of data to understand customer behavior, preferences, and trends, allowing for tailored recommendations, targeted marketing campaigns, and interactive experiences. By leveraging machine learning in this way, businesses can forge deeper connections with their customers, leading to increased satisfaction, loyalty, and ultimately, improved business outcomes.

How Patronus Automates LLM Evaluation to Boost Confidence in GenAI

Patronus AI is a company that develops tools to help businesses safely use large language models (LLMs). Their main product is an automated evaluation platform that can identify errors and unreliable outputs from LLMs. This is especially important for regulated industries where mistakes can have serious consequences.

Founded by machine learning experts from Meta AI and Meta Reality Labs, Patronus AI is on a mission to boost enterprise confidence in gen AI-powered apps, leading the way in shaping a trustworthy AI landscape.

"Our platform enables engineers to score and benchmark LLM performance on real-world scenarios, generate adversarial test cases, monitor hallucinations, and detect PII and other unexpected and unsafe behavior. Customers use Patronus AI to detect LLM mistakes at scale and deploy AI products safely and confidently."

Rebecca Qian, Co-founder and CTO at Patronus

Overcoming LLM hallucination

In recently published and widely cited research based on the FinanceBench question answering (QA) evaluation suite, Patronus made a startling discovery. Researchers found that a range of widely used state-of-the-art LLMs frequently hallucinated, incorrectly answering or refusing to answer up to 81% of financial analysts' questions! This error rate occurred despite the models' context windows being augmented with context retrieved from an external vector store.

While retrieval augmented generation (RAG) is a common way of feeding models with up-to-date, domain-specific context, a key question faced by app owners is how to test the reliability of model outputs in a scalable way. This is where Patronus comes in. The company has partnered with the leading technologies in the gen AI ecosystem — from model providers and frameworks to vector store and RAG solutions — to provide managed evaluation services, test suites, and adversarial data sets.

MongoDB. | Patronus AI

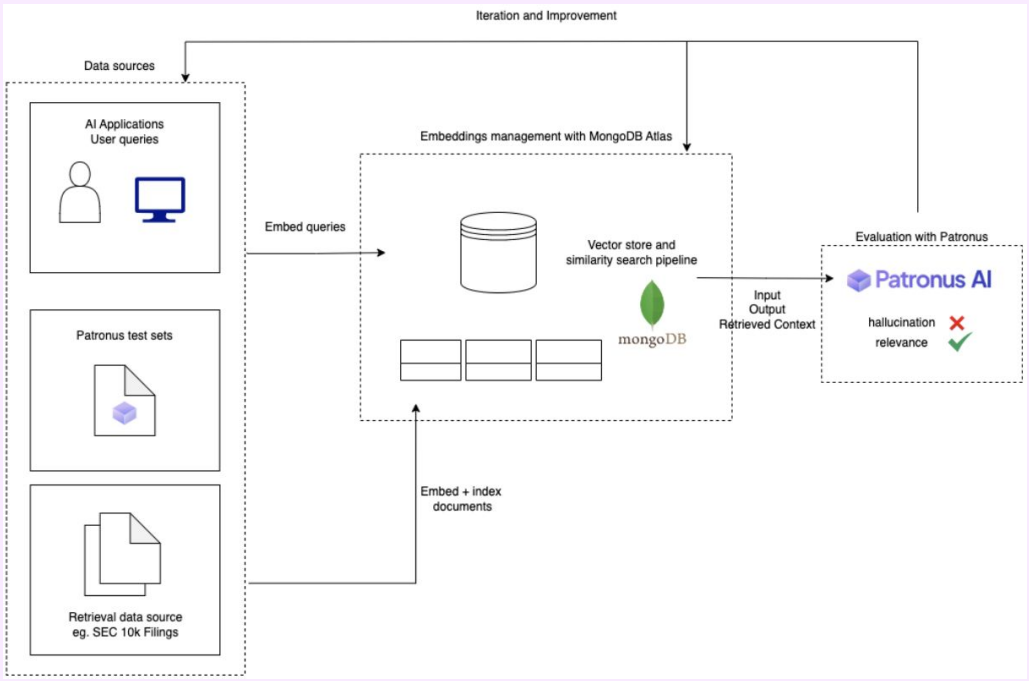


Figure 28: Reference architecture and workflow

Boosting confidence in LLMs with MongoDB

As Patronus assessed the landscape to prioritize which partners to work with, they saw massive demand from customers for MongoDB Atlas. Through the Patronus RAG evaluation API, they help customers verify that their RAG systems built on top of MongoDB Atlas consistently deliver top-tier, dependable information.

In its new 10-minute guide, Patronus takes developers through a workflow showcasing how to evaluate a MongoDB Atlas-based retrieval system. The guide focuses on evaluating hallucination and answers relevance against an SEC 10-K filing, simulating a financial analyst querying the document for analysis and insights. The workflow is built using:

- The LlamaIndex data framework to ingest and chunk the source pdf document
- Atlas Vector Search to store, index, and query the chunk's metadata and embeddings
- Patronus to score the model responses

Equipped with the results of an analysis, there are a number of steps developers can take to improve the performance of a RAG system. These include exploring different indexes, modifying document chunking sizes, re-engineering prompts, and for the most domain-specific apps, fine-tuning the embedding model itself. Review the 10-minute guide for a more detailed explanation of each of these steps.

How Gradient Accelerator Blocks Take You From Zero To AI in Seconds

Gradient, founded by AI experts from Google, Netflix, and Splunk, helps businesses build high-performing, cost-effective custom AI applications. It provides a platform for businesses to build, customize, and deploy bespoke AI solutions — starting with the fastest way to develop AI through the use of its Accelerator Blocks.

Fast Development with Pre-built Blocks

Gradient offers Accelerator Blocks - pre-built solutions for common AI tasks like entity extraction or document summarization. These blocks can be used directly or combined for more complex needs, reducing development time and effort.

Benefits for Regulated Industries

The platform empowers regulated industries such as finance and healthcare businesses with data and AI control for regulatory compliance, offering industry-specific models and performance/cost benefits.

"With MongoDB, developers can store data of any structure and then expose that data to OLTP, text search, and vector search processing using a single query API and driver. With this unification, developers have all of the core data services they need to build AI-powered apps that rely on working with live, operational data."

Tiffany Peng, VP of Engineering at Gradient

Simplified RAG with Powerful Tech

- **Simplified Infrastructure**: Gradient's Accelerator Block for retrieval augmented generation (RAG) leverages MongoDB Atlas Vector Search and LlamaIndex. By using these technologies, Gradient eliminates the need for complex infrastructure setup or deep knowledge of retrieval architectures.
- **Best-of-Breed Technologies**: Gradient partners with key vendors and communities in the AI ecosystem. MongoDB Atlas, included as a core part of the Gradient platform, provides operational databases and vector search capabilities in a unified, fully managed solution.
- **Seamless Data Handling**: With MongoDB, developers can store data of any structure and expose it to OLTP, text search, and vector search processing using a single query API and driver. This unification provides all the core data services needed to build AI-powered apps that work with live, operational data.

MongoDB | gradient

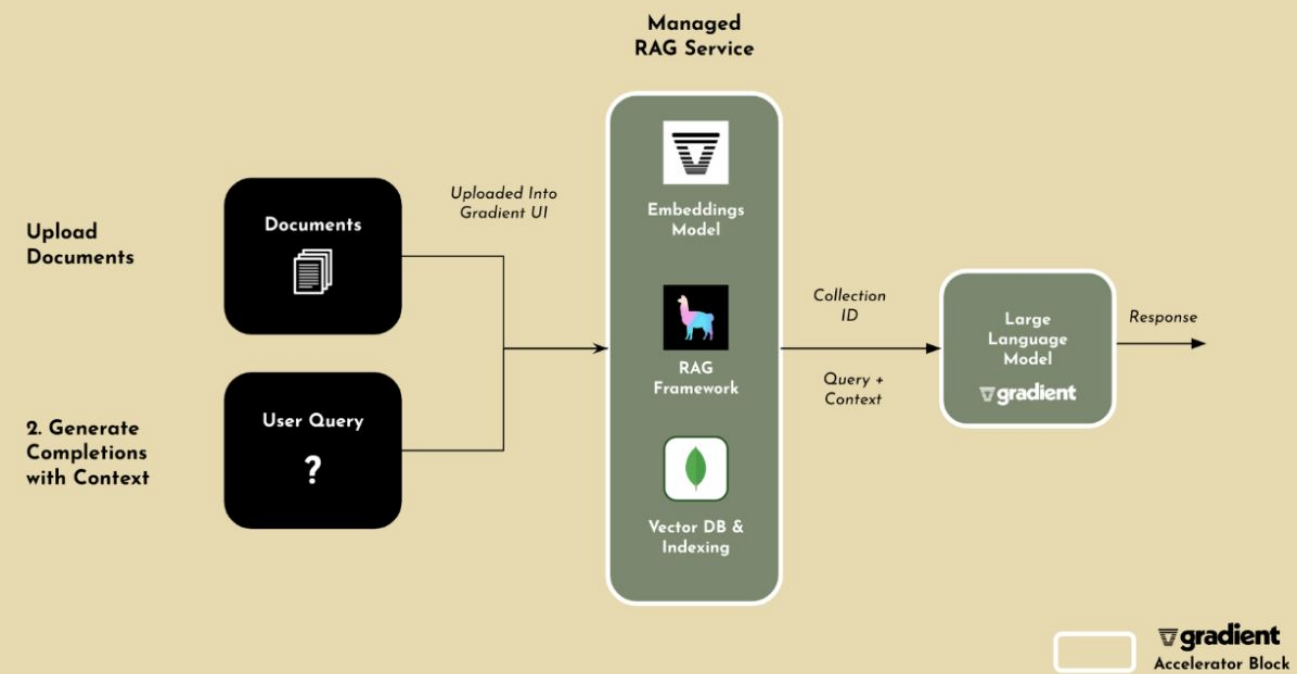


Figure 29: Managed RAG service

Gradient's Accelerator Block Boosts RAG Model Performance and Accuracy with Pre-Built Infrastructure

Gradient's newest Accelerator Block focuses on enhancing the performance and accuracy of a model through retrieval augmented generation (RAG). The Accelerator Block uses Gradient's state-of-the-art LLMs and embeddings, MongoDB Atlas Vector Search for storing, indexing, and retrieving high-dimensional vector data, and LlamaIndex for data integration.

Together, Atlas Vector Search and LlamaIndex feed foundation models with up-to-date, proprietary enterprise data in real-time. Gradient designed the Accelerator Block for RAG to improve development velocity up to 10x by removing the need for infrastructure, setup, or in-depth knowledge around retrieval architectures. It also incorporates best practices in document chunking, re-rankers, and advanced retrieval strategies.

One AI: Providing AI-as-a-Service to deliver solutions in days rather than months

One AI is a company that aims to democratize and deliver AI as a service for businesses. Their mission is to integrate AI into everyday life by transforming natural language into structured, actionable data. This is achieved through their easy-to-use APIs, which package leading AI capabilities from across the ecosystem.

AI-as-a-Service

One AI provides AI-as-a-Service, delivering solutions in days rather than months. This allows businesses to deploy tailored AI solutions quickly and efficiently.

Diverse Use Cases

One AI's customers span multiple domains, utilizing their service for a variety of use cases, from analyzing financial documents to AI-automated video editing.

API's for Developers

The One AI APIs allow developers to analyze, process, and transform language input in their code, without requiring any training data or NLP/ML knowledge.

Flexible Data Infrastructure

One AI works with over 20 different AI/ML models and leverages a flexible data infrastructure, specifically the MongoDB document model, to continuously explore and add new capabilities for the AI.

Choice of MongoDB as Developer Data Platform

- **Focus on Core Mission**: MongoDB allows One AI to focus on their core mission of using AI to derive meaning from large volumes of unstructured text1. Dealing with database requirements and services, such as managing the pipeline, storage, and backups, involves a lot of time, effort, and hassle. MongoDB handles these tasks, allowing One AI to concentrate on their main objective.
- **Flexible Data Infrastructure**: With MongoDB, One AI can add, expand, and explore new capabilities on a continuous basis.
- **Regular New Releases**: One AI benefits from regular new releases from MongoDB, such as Atlas Vector Search. This feature allows One AI to have vectorized language representation in the same database as other representations, which can be accessed via a single query interface. This solves a core problem for One AI as an API company.

MongoDB | ONE AI

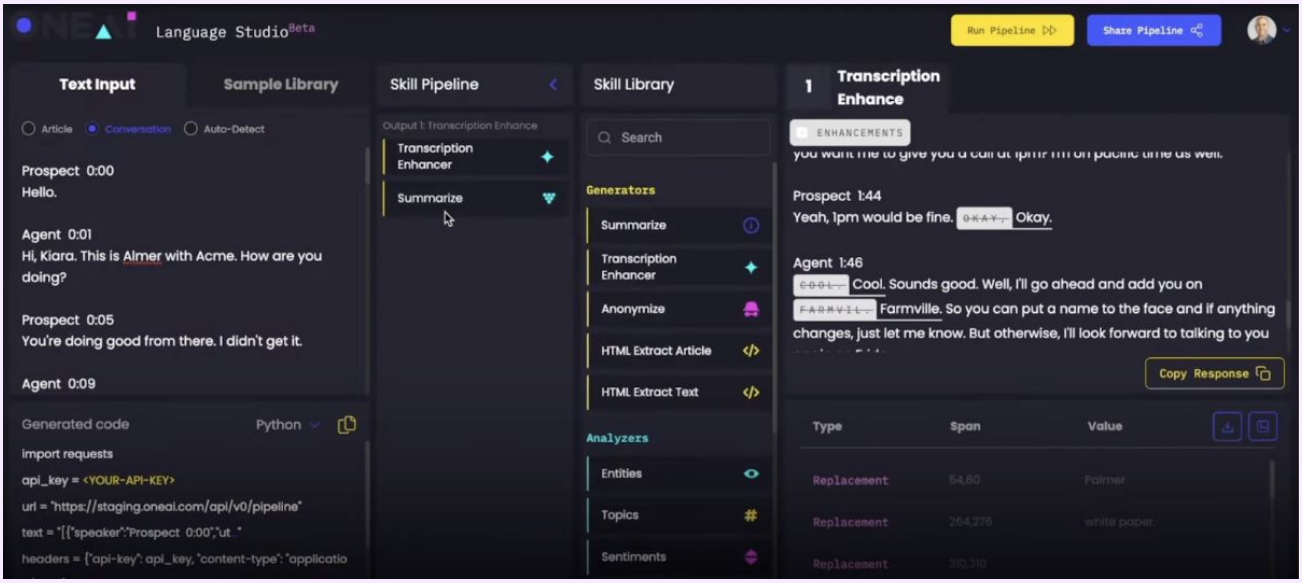


Figure 30: The One AI Language Studio

The One AI APIs let developers analyze, process, and transform language input in their code. No training data or NLP/ML knowledge are required.

"The MongoDB document model really allows us to spread our wings and freely explore new capabilities for the AI, such as new predictions, new insights, and new output data points." Ben adds, "With any other platform, we would have to constantly go back to the underlying infrastructure and maintain it. Now, we can add, expand, and explore new capabilities on a continuous basis."

Amit Ben, CEO One AI

The company also benefits from the regular new releases from MongoDB, such as Atlas Vector Search, which Ben sees as a highly valuable addition to the platform's toolkit. Ben explains: "The ability to have that vectorized language representation in the same database as other representations, which you can then access via a single query interface, solves a core problem for us as an API company."

To learn more, watch the interview with Amit Ben.

MongoDB. | ONE AI

Kovai: Bringing the power of Vector Search to enterprise knowledge bases

Founded in 2011, Kovai is an enterprise software company that offers multiple products in both the enterprise and B2B SaaS arena. Since its founding, the company has grown to nearly 300 employees serving over 2,500 customers.

Document 360

Kovai's key product, Document360, is a knowledge base platform designed for SaaS companies seeking a self-service software documentation solution. It enables efficient management and sharing of critical information.

AI Assistant "Eddy"

Kovai recognized the growing importance of AI and developed an AI assistant named "Eddy". Eddy leverages LLMs (Language Models) and retrieves information from the Document360 knowledge base to provide accurate answers to customer queries.

"Atlas Vector Search is robust, cost-effective, and blazingly fast!"

Said Saravana Kumar, *CEO, Kovai, when speaking about his team's experience*

Choice of MongoDB as Developer Data Platform

- MongoDB Vector Search offers architectural simplicity, making it easier for Kovai to optimize the technical architecture needed to implement their AI assistant, "Eddy." This simplicity likely **streamlines development efforts and reduces complexity** in integrating the search functionality into their system.
- MongoDB Vector Search delivers **faster query response times at scale**, ensuring a **positive user experience** for Kovai's customers interacting with the AI assistant.
- Atlas Vector Search enables Kovai to store both knowledge base articles and their embeddings together in MongoDB collections. This eliminates the need for data syncing between multiple databases, which not only **simplifies operations** but also **reduces potential inaccuracies** in answers provided by the assistant. Operational efficiency is crucial for a **seamless user experience.**

MongoDB | Kovai.co

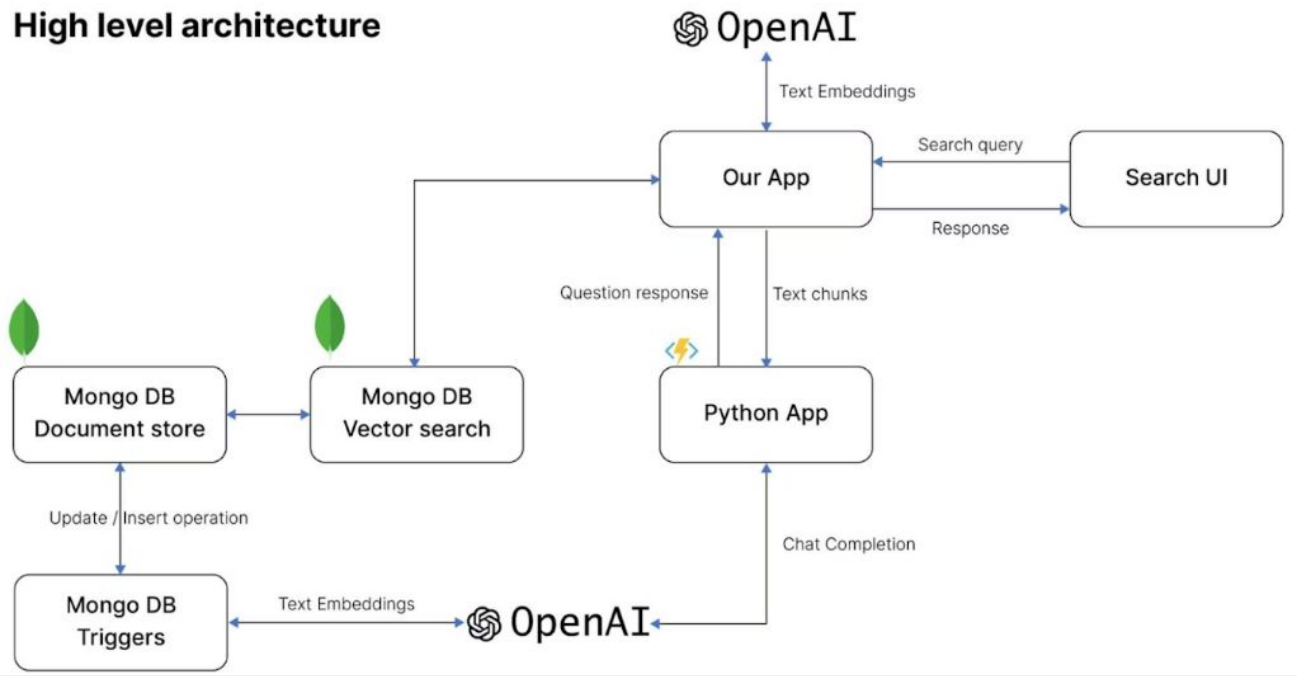


Figure 32: Reference architecture

Faster, Simpler, More Efficient: How Kovai Leverages MongoDB Atlas Vector Search

The release of MongoDB Atlas Vector Search provided a solution with three key advantages for the engineers:

- **Architectural simplicity:** MongoDB Vector Search's architectural simplicity helps Kovai optimize the technical architecture needed to implement Eddy.
- **Operational efficiency:** Atlas Vector Search allows Kovai to store both knowledge base articles and their embeddings together in MongoDB collections, eliminating "data syncing" issues that come with other vendors.
- **Performance:** Kovai gets faster query response from MongoDB Vector Search at scale to ensure a positive user experience.

Specifically, the team has seen the average time taken to return three, five, and 10 chunks between two and four milliseconds, and if the question is a closed loop, the average time reduces to less than two milliseconds.

You can learn more about Kovai's journey into the world of RAG in the full case study.

Robust Intelligence: Securing generative AI, supercharged by your data

Robust Intelligence safeguards organizations from AI's risks. Their end-to-end platform continuously validates models, protecting them with an AI Firewall. This empowers confident AI adoption for any model type, from basic to generative. Trusted by leaders like JPMorgan Chase, Robust Intelligence is your key to unlocking AI's potential.

Recent advancements in generative AI have motivated companies to experiment with potential applications, but a lack of security controls has exposed companies to unmanaged risks. This challenge is exacerbated when sensitive company information is used to enrich pre-trained models, such as connecting vector databases, in order to increase the relevance to the end user.

Robust Intelligence's AI Firewall safeguards large language models (LLMs) in production by validating inputs and outputs in real-time. It addresses operational risks like hallucinations, ethical risks such as model bias and toxic outputs, and security risks like prompt injections and PII extraction. By intercepting harmful inputs and filtering out undesirable AI-generated outcomes, the AI Firewall ensures model integrity and application safety.

"By incorporating MongoDB's Atlas Vector Search into the AI validation process, customers can confidently use their databases to enhance LLM responses knowing that sensitive information will remain secure. The integration provides seamless protection against a comprehensive set of security, ethical, and operational risks."

Yaron Singer, CEO and co-founder at Robus Intelligence

MongoDB. | ROBUST INTELLIGENCE

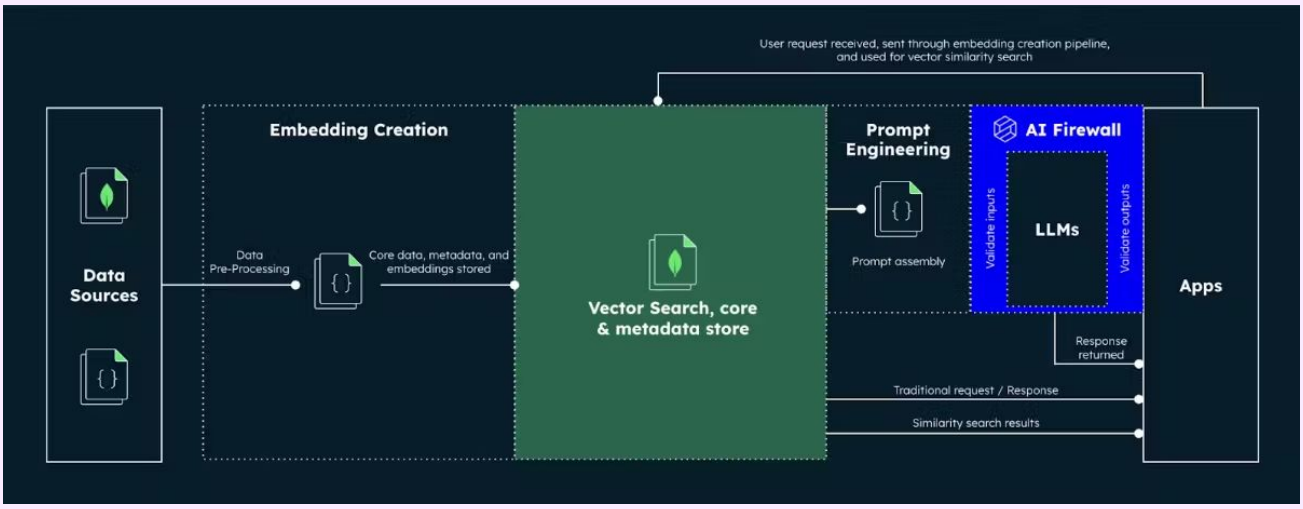


Figure 34: High level architecture

Unlocking Personalized Customer Experiences with Algomo's Conversational AI

Customers can confidently connect MongoDB Atlas Vector Search to any commercial or open-source LLM for secure retrieval-augmented generation with the AI Firewall integration. Atlas Vector Search serves as the memory and fact database for AI Firewall, ensuring the AI model provides enriched responses without hallucinating.

Additionally, it serves as the memory and database to store historical data points. This is important in the context of identifying more advanced security attacks, such as data poisoning and model extraction, which often manifest across a cluster of data points as opposed to a single data point.

Component-Based AI for Development Teams

Solutions built out of building blocks can be seamlessly integrated into existing systems without disrupting other functions or data.

Fireworks AI and MongoDB: The Fastest AI Apps with the Best Models, Powered By Your Data

Fireworks AI and MongoDB are now partnering to make innovating with generative AI faster, more efficient, and more secure. Fireworks AI was founded in late 2022 by industry veterans from Meta's PyTorch team, where they focused on performance optimization, improving the developer experience, and running AI apps at scale. It's this expertise that Fireworks AI brings to its production AI platform, curating and optimizing the industry's leading open models. Benchmarking by the company shows gen AI models running on Fireworks AI deliver up to 4x faster inference speeds than alternative platforms, with up to 8x higher throughput and scale.

Models are one part of the application stack. But for developers to unlock the power of gen AI, they also need to bring enterprise data to those models. That's why Fireworks AI has partnered with MongoDB, addressing one of the toughest challenges to adopting AI. With MongoDB Atlas, developers can securely unify operational data, unstructured data, and vector embeddings to safely build consistent, correct, and differentiated AI applications and experiences. Fireworks AI and MongoDB provide a solution for developers who want to leverage highly curated and optimized open-source models, and combine these with their organization's own proprietary data — and to do it all with unparalleled speed and security.

Lightning-fast models from Fireworks AI: Enabling speed, efficiency, and value

With its lightning-fast inference platform, Fireworks AI curates, optimizes, and deploys 40+ different AI models, resulting in significant cost savings, reduced latency, and improved throughput. Their platform delivers this via:

- **Off-the-shelf models, optimized models, and add-ons**: Fireworks AI provides a collection of top-quality text, embedding, and image foundation models. Developers can leverage these models or fine-tune and deploy their own, pairing them with their own proprietary data using MongoDB Atlas.
- **Fine-tuning capabilities**: To further improve model accuracy and speed, Fireworks AI also offers a fine-tuning service using its CLI to ingest JSON-formatted objects from databases such as MongoDB Atlas.
- **Simple interfaces and APIs for development and production**: The Fireworks AI playground allows developers to interact with models right in a browser. It can also be accessed programmatically via a convenient REST API. This is OpenAI API-compatible and thus interoperates with the broader LLM ecosystem.
- **Cookbook**: A simple and easy-to-use cookbook provides a comprehensive set of ready-to-use recipes that can be adapted for various use cases, including fine-tuning, generation, and evaluation.

MongoDB | Fireworks AI

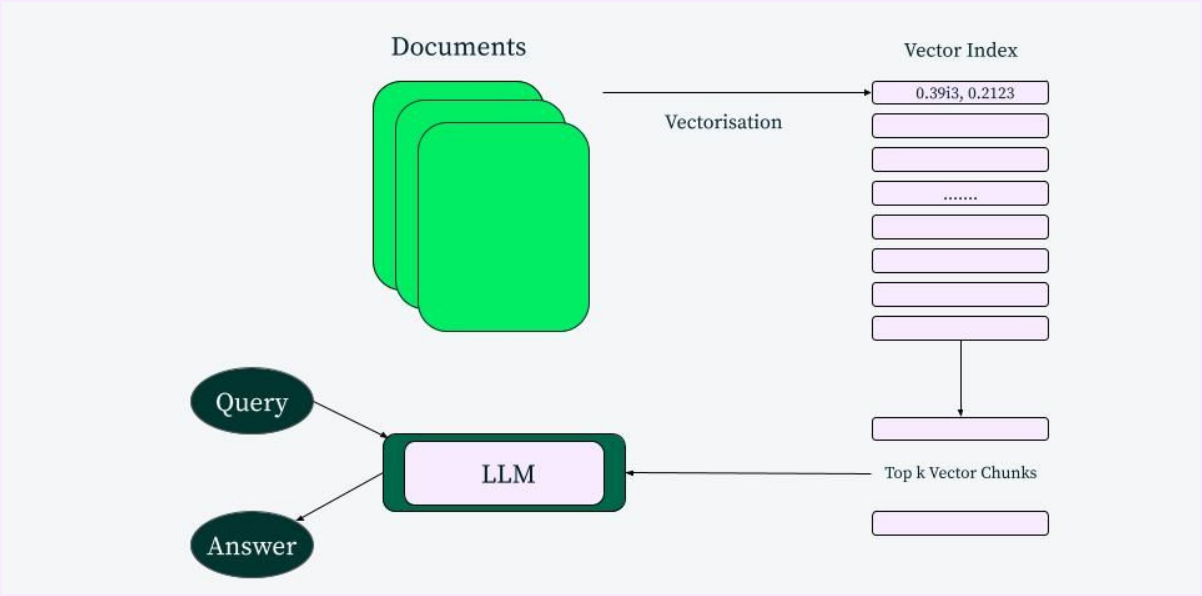


Figure 20: Bringing your data to LLMs

Getting started: The Fireworks tutorial showcases how to bring your own data to LLMs with retrieval-augmented generation (RAG) and MongoDB Atlas

With Fireworks AI and MongoDB Atlas, apps run in isolated environments ensuring uptime and privacy, protected by sophisticated security controls that meet the toughest regulatory standards:

- As one of the top open-source model API providers, Fireworks AI serves 66 billion tokens per day (and growing).
- With Atlas, you run your apps on a proven platform that serves tens of thousands of customers, from high-growth startups to the largest enterprises and governments.

Together, the Fireworks AI and MongoDB joint solution enables:

- **Retrieval-augmented generation (RAG) or Q&A from a vast pool of documents**: Ingest a large number of documents to produce summaries and structured data that can then power conversational AI.
- **Classification through semantic/similarity search**: Classify and analyze concepts and emotions from sales calls, video conferences, and more to provide better intelligence and strategies. Or, organize and classify a product catalog using product images and text.
- **Images to structured data extraction**: Extract meaning from images to produce structured data that can be processed and searched in a range of vision apps — from stock photos, to fashion, to object detection, to medical diagnostics.
- **Alert intelligence**: Process large amounts of data in real-time to automatically detect and alert on instances of fraud, cybersecurity threats, and more.

Getting started with Fireworks AI and MongoDB Atlas: review the Optimizing RAG with MongoDB Atlas and Fireworks AI tutorial, which shows you how to build a movie recommendation app.

MongoDB. | Fireworks AI

How GoBots AI for E-commerce Increases Retailer Sales Conversion by 40%

Major retail brands have long been using various forms of AI, for example statistical analysis and machine learning models, to better serve their customers. But with its high barriers to entry, one key channel has been slower to embrace the technology. By connecting large and small brands with customers, e-commerce marketplaces such as Amazon, Mercado Libre, and Shopify are among the fastest growing retail routes to market. Since 2016, GoBots has been working to extend the benefits of AI to any retailer on any marketplace. It uses AI, analytics, and MongoDB Atlas to make e-commerce easier, more convenient, and smarter for brands serving Latin America.

GoBots increases engagement and conversion rates for over 600 clients across Latin America, including Adidas, Bosch, Canon, Chevrolet, Dell, Electrolux, Hering, HP, Nike, and Samsung.

The solution makes the benefits of AI available to any retailer, whether large or small. With the GoBots natural language understanding (NLU) model, retailers automate customer interactions such as answering questions and resolving issues through intelligent assistants. At the same time, they leverage data analytics to offer personalized customer experiences.

By using GoBots AI for ecommerce with MongoDB Atlas, customers have grown sales conversions by 40% and reduced time to customer response by 72%.

With the power of MongoDB's developer data platform and flexibility of MongoDB's document model, GoBots builds higher-performing AI-powered applications faster:

- MongoDB Atlas provides a **single data platform** that serves multiple operational and AI use cases. This includes user data and product catalogs as well as a store for AI model inferences, outputs of multiple AI models for experimentation and evaluation purposes, a data source for fine-tuning models, and for vector search.
- GoBots is evaluating the use of Atlas Triggers for invoking AI model API calls in an event-driven manner as the underlying data changes.
- The flexibility provided by MongoDB's document model allows the development team to continually **enrich historical questions** with outputs generated by different models and compare the results. This means that they are not blocked behind complex schema changes that would otherwise slow down the pace of harnessing new data in their models for training and inference.
- The question-answer pairs output by the company's NLU models and LLMs are complex data structures with many nested entities and arrays. Being able to persist these directly to the database without first having to transform them into a tabular structure improves developer productivity and reduces application latency.

MongoDB | GoBots

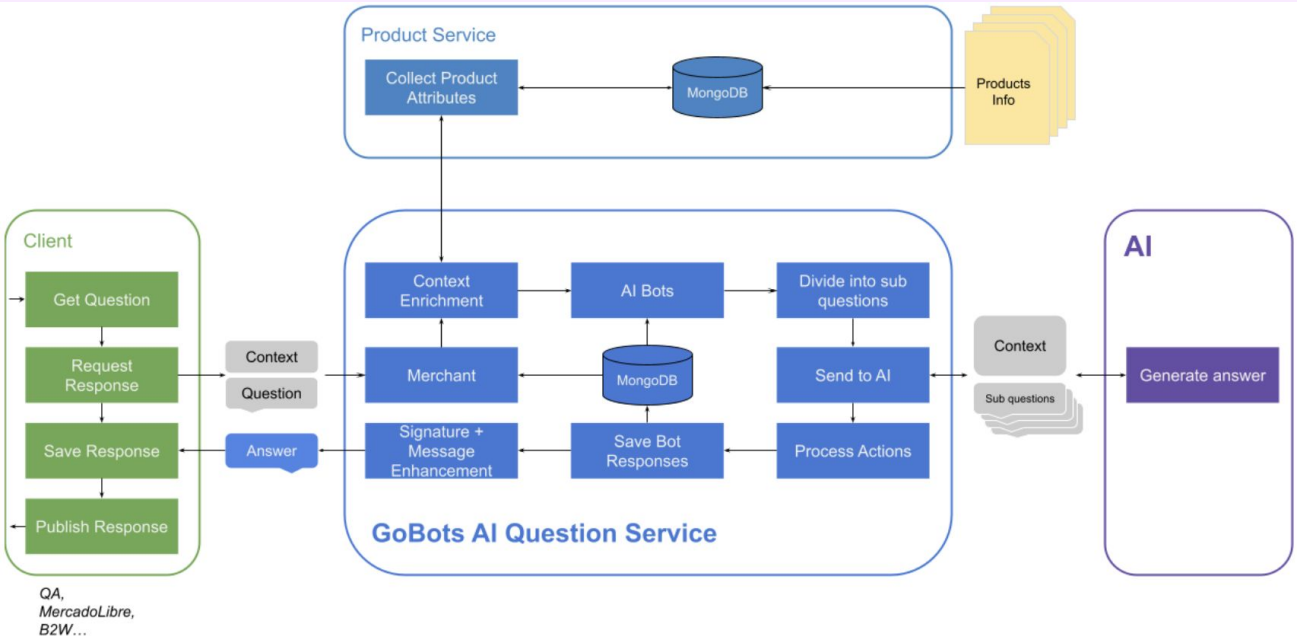


Figure 21: GoBots question processing architecture

GoBots' custom NLU models are built using the Rasa framework with a neural network trained on over 150 million question-answer examples and more than 50 bots — specialists in different segments — to understand more specific questions.

Models are fine tuned with data from the retailer's own product catalog and website corpus. The model runtime is powered by a PyTorch microservice on Google Cloud. The larger GoBots platform is built with Kotlin and orchestrated by Kubernetes, providing the company with cloud freedom as its business expands and evolves.

The GoBots AI assistants kick into action as soon as a customer asks a question on the marketplace site, with the questions stored in MongoDB Atlas. GoBots' natural language models are programmatically called via a REST API to perform tasks like named entity recognition (NER), user intent detection, and question-answer generation with all inferences also stored in MongoDB. If the models are able to generate an answer with high confidence, the GoBots service will respond directly to the customer in real time. In case of a low confidence response, the models flag the question to a customer service representative who receives a pre-generated suggested response.

With all question-answer pairs from the different models written to the MongoDB Atlas database, the data is used to further tune the natural language models while also guiding model evaluations. The company has also recently started using Atlas Vector Search to identify and retrieve semantically similar answers to past questions. The search results power a co-pilot-like experience for customer service representatives and provide in-context training to its fleet of LLMs.

Story Tools Studio Brings Gen AI To Gaming With Myth Maker AI

Story Tools Studio harnesses cutting-edge generative AI (gen AI) technologies to craft immersive, personalized, and infinite storytelling experiences. Their flagship game Myth Maker AI leverages MUSE (Modular User Story Engine), an internally developed AI-powered, expert-guided story generator that blends a growing collection of advanced AI technology with creative artistry to weave real time narratives.

MUSE (Modular User Story Engine) combines professionally crafted stories with user-empowered experiences. Players make intentional choices that guide the story with AI adapting to each decision in real time, providing a unique and personalized journey. MUSE separates the story from game mechanics, allowing the development of multiple game types. Its use of AI creates more agile teams with fewer dependencies.

"By selecting MongoDB, we were able to create a prototype of our game in just 48 hours. It is only with MongoDB that we can release new features to production multiple times per day. We couldn't achieve any of this with a relational database."

Roy Altman, Founder and CEO at Story Tools Studio

AI, transactions, and analytics with MongoDB

The engineering team has used MongoDB Atlas from the very start of the company. MongoDB **stores all of the data used in the platform**: — user data, scripts, characters, worlds, coins, and prompts are all richly structured objects stored natively in MongoDB. The games are built in React and Javascript.

Beyond gameplay, the company's developers are now exploring MongoDB's ACID transactional integrity to support in-game monetization, alongside in-app intelligence to further improve the gaming experience through player analytics.

By running MongoDB in Atlas, Story Tools Studio's engineering team:

- Is **free to focus** on AI-driven gaming experiences, and not on the grind of managing a database
- Was able to **scale seamlessly and automatically** as the team graduated from their closed beta into public beta
- **Manage player demands** with dozens of new players every day – every 24 hours they are organically adding dozens of new players with tens of gigabytes of new data streaming into the platform

MongoDB | STORY TOOLS Studio

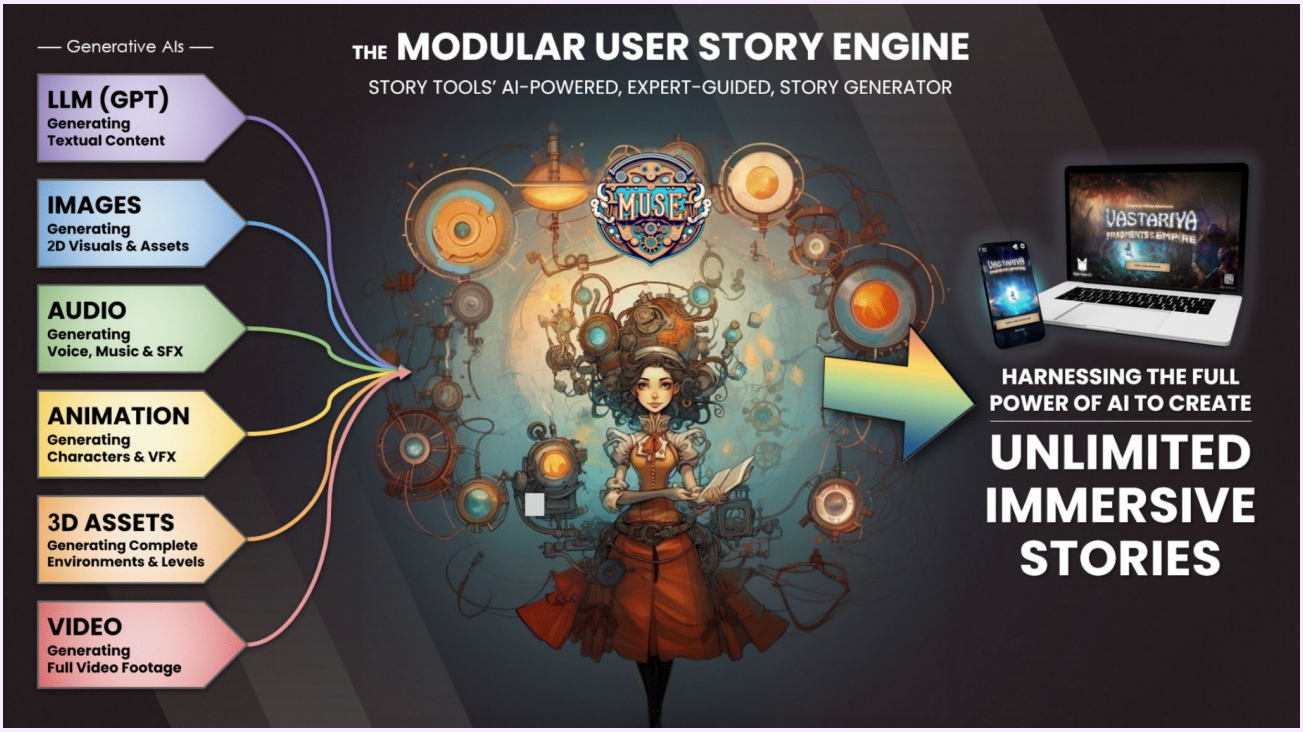


Figure 23: Story Tools modular user experience engine

MUSE orchestrates multimodel genAI to create real time, unlimited stories

When a player starts a game in Myth Maker AI, they are presented with the option to choose their starting hero character. Under the covers, MUSE calls the GPT4 API, which takes the player's selection and writes a fully customized adventure premise. From that initial personalized script, MUSE programmatically calls specialized AI models to collaboratively generate an immersive, multimodal gaming experience using images, animation, audio, and soon, video and 3D.

For story generation and text to voice, this is run in Azure's OpenAI service. Visual assets are created via Leonardo AI, and the team are constantly experimenting with new models to create richer modalities. Currently, the team is working on generating enhanced 3D assets and video from text prompts. With the pace of AI advancement, the creativity of the team, and the input from game testers, Story Tools Studio has the flexibility to continuously deploy new features with MongoDB's dynamic and flexible document data model. This enables Story Tools Studio to build a truly innovative, artistic platform, opening up a whole new world of experiences for both creators and audiences alike.

Accelerating App Development With the Codeium AI Toolkit

Of the many use cases set to be transformed by generative AI (gen AI), the bleeding edge of this revolution is underway with software development. Developers are using gen AI to improve productivity by writing higher-quality code faster. Tasks include autocompleting code, writing docs, generating tests, and answering natural language queries across a code base. How does this translate to adoption? A recent survey showed 44% of new code being committed was written by an AI code assistant.

Codeium is one of the leaders in the fast-growing AI code assistant space. Its AI toolkit is used by hundreds of thousands of developers for more than 70 languages across more than 40 IDEs including Visual Studio Code, the JetBrains suite, Eclipse, and Jupyter Notebooks. The company describes its toolkit as "the modern coding superpower," reflected by its recent $65 million Series B funding round and five-star reviews across extension marketplaces. Codeium was developed by a team of researchers and engineers to build on the industry-wide momentum around large language models, specifically for code. They realized that their specialized generative models, when deployed on their world-class optimized deep learning serving software, could provide users with top-quality AI-based products at the lowest possible costs.

Training models on MongoDB

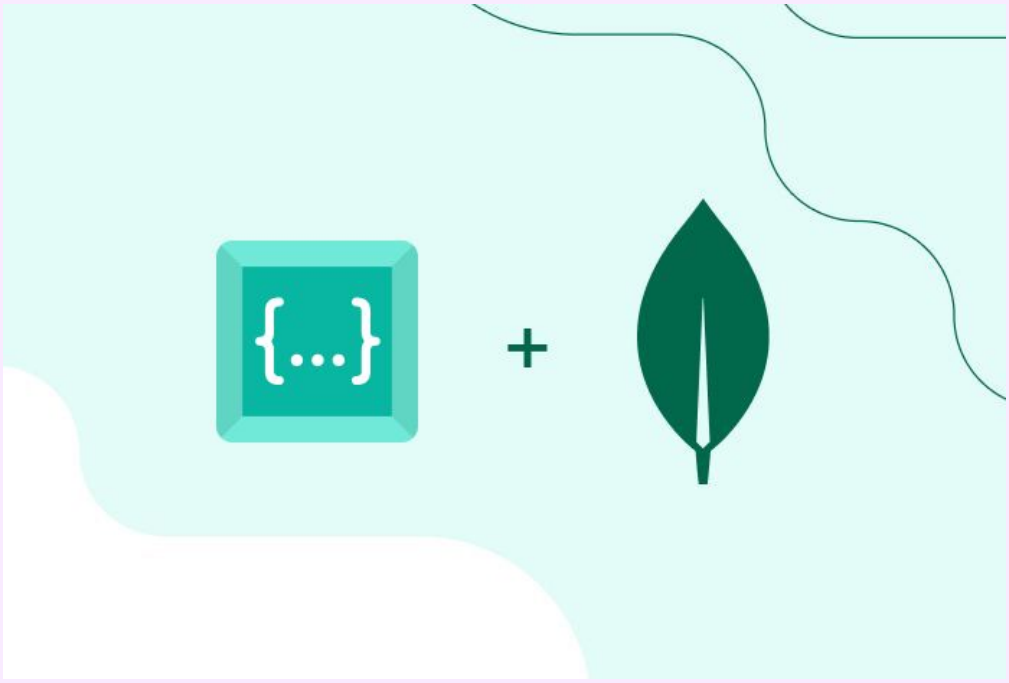
Codeium has recently trained its models on MongoDB code, libraries, and documentation. Now developers building apps with MongoDB can install the Codeium extension on the IDE of their choice and enjoy **rapid code completion** and **codebase-aware chat and search**.

Developers can stay in the flow while they build, coding at the speed of thought, knowing that Codeium has ingested MongoDB best practices and documentation.

"MongoDB is wildly popular across the developer community. This is because Atlas integrates the fully managed database services that provide a unified developer experience across transactional, analytical, and generative AI apps."

Anshul Ramachandran, Head of Enterprise & Partnerships, Codeium

MongoDB | codeium



Getting Started with MongoDB and Codeium

MongoDB APIs are incredibly powerful, but due to the breadth and richness of the APIs, it is possible for developers to be spending more time than necessary looking through API documentation or using the APIs inefficiently for the task at hand. An AI assistant, if trained properly, can effectively assist the developer in retrieval and usage quality of these APIs. Unlike other AI code assistants, we at Codeium build our LLMs from scratch and own the underlying data layer. This means we accelerate and optimize the developer experience in unique and novel ways unmatched by others.

In its announcement blog post and YouTube video, the Codeium team shows how to build an app in VSCode with MongoDB serving as the data layer. Developers can ask questions on how to read and write to the database, get code competition suggestions, explore specific functions and syntax, handle errors, and more. This was all done at no cost using the MongoDB Atlas free tier and Codeium 100% free.

You can get started today by registering for MongoDB Atlas and then downloading the Codeium extension.

MongoDB. | codeium

Nomic AI: Cost-effective, Open Source Embeddings at Scale

Nomic Embed v1.5 is a truly open source text embedder for the big data era. Out of the box, this model supports a 8192 token context length, resizable embedding dimensions, and binary quantization, all while outperforming similar models such as OpenAI's Ada-002 and text-embedding-3-small on both short and long context tasks.

Truly Open Source

Nomic Embed provides open-source model weights and training code under the Apache-2 license, with curated training data available on the Nomic website. This ensures full reproducibility and auditability.

High Throughput

Nomic Embed provides high-quality, compact embeddings, ideal for high-throughput, data-heavy workflows. On an AWS Sagemaker single GPU ml.g5.xlarge instance, it returns an embedding roughly every 0.01 seconds.

Long Context

Nomic Embed supports a 8192 token context length making it well-suited for real-world applications with large PDFs and text documents.

Cost-effective Storage

Nomic Embed offers flexible embedding sizes via Matryoshka representation learning. Users can choose to store 64, 128, 256, or 512 embedding dimensions from the full 768. Smaller sizes reduce performance loss and storage costs linearly.

Unleashing Nomic Embeddings with MongoDB Atlas

- **Seamless Integration**: MongoDB Atlas integrates Nomic embeddings effortlessly, storing both embeddings and metadata in MongoDB collections, either together or separately. Both MongoDB Atlas and Nomic Embed are available on AWS Marketplace for identical VPC deployments.
- **Powerful Analytics Capabilities**: MongoDB Vector Search combines Nomic embeddings for fast semantic search, enabling the fusion of vector search and traditional database queries on metadata. It's a flexible analytics tool for data insights, user recommendations, and more.
- **Streaming and Triggers**: Mongo Stream Processing is a perfect fit for Nomic Embed's high throughput capabilities. Incoming data streams are robustly processed and can be combined with MongoDB Triggers to generate embeddings for immediate downstream use. Given Nomic Embed's lightweight nature and offline capabilities (via private or local deployments from open source), embeddings can be produced and ingested into MongoDB at extremely rapid transfer rates.

MongoDB | NOMIC

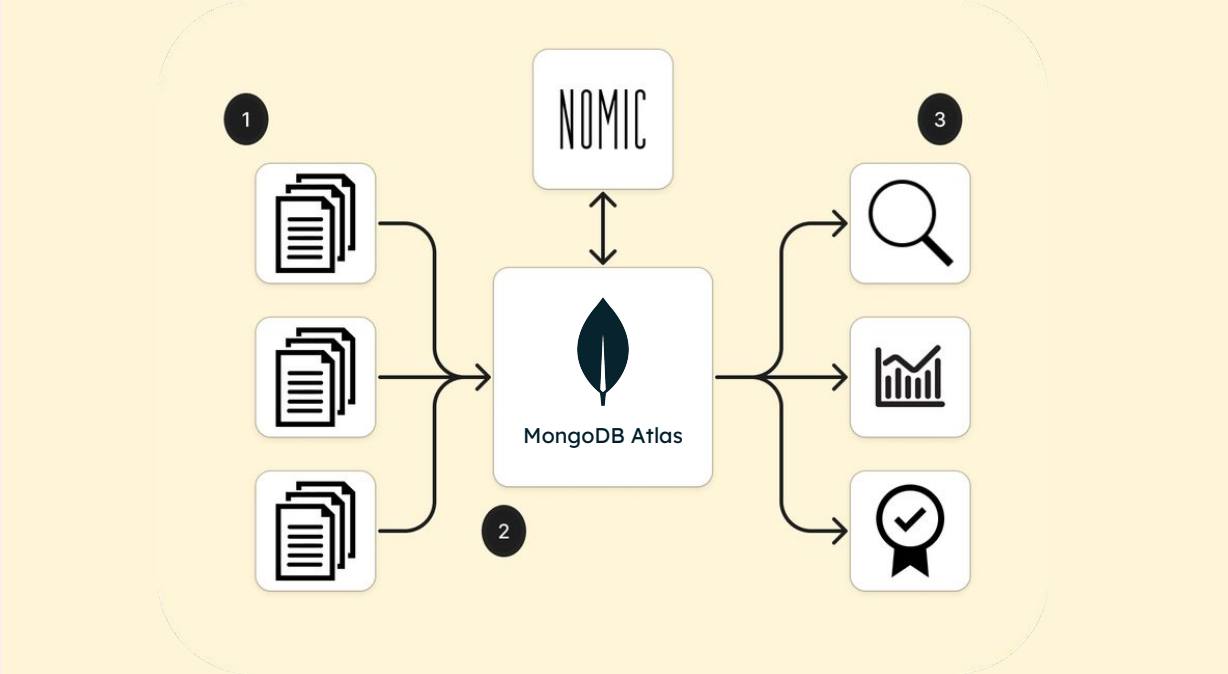


Figure 24: Use-case: PDF Search

PDF Search is a use-case that combines the capabilities of Nomic Embed and MongoDB Atlas for an accessible, high-throughput solution. Nomic Embed simplifies the process of embedding PDF previews directly into your application, while MongoDB Atlas provides a powerful and scalable NoSQL database to store and index your PDFs for efficient searching. This combination allows you to quickly build a user-friendly search experience for your PDFs without worrying about complex infrastructure management.

- Large PDFs can be chunked and ingested into MongoDB Atlas via stream processing, while
- Nomic Embed can quickly produce long-context embeddings from the processed text.
- MongoDB Vector Search integrates semantic search on Nomic embeddings with traditional database queries for multi-faceted downstream analysis.

Together AI: Advancing the Frontier of AI With Open Source Embeddings, Inference, and MongoDB Atlas

Founded in San Francisco in 2022, Together AI is on a mission to create the fastest cloud platform for building and running generative AI (gen AI). The company has so far raised over $120 million, counting Nvidia, Kleiner Perkins, Lux, and NEA as investors.

Together Embeddings Endpoint

Together AI introduces Together Embeddings. This service provides access to eight leading open-source embedding models at a significantly lower cost (up to 12x cheaper) compared to proprietary options.

Customized AI Outputs

With the retrieval-augmented generation (RAG) pattern, developers can feed gen AI models with their own up-to-date, domain-specific data, resulting in more reliable and customized outputs, reducing the risk of hallucinations.

"We prioritized integrating with MongoDB because of its relevance and importance in the AI stack."

Vipul Ved Prakash, Founder and CEO at Together AI

Build Better AI Apps Faster

Atlas Vector Search is used to store and index embeddings and then perform semantic search to retrieve relevant data examples for natural language queries against a sample Airbnb listing dataset. With this RAG pattern, the gen AI model can recommend properties that **meet the user's criteria while adhering to factual information.**

- **Reduced Complexity and Cost**: MongoDB Atlas helps developers reduce complexity and cost. It brings together live application data synchronized right alongside vector embeddings in a single platform, which simplifies the development process.
- **Faster Time-to-Market**: With MongoDB Atlas, developers can bring cutting-edge apps to market faster. This is crucial for staying competitive in the rapidly evolving field of AI.

MongoDB | together.ai

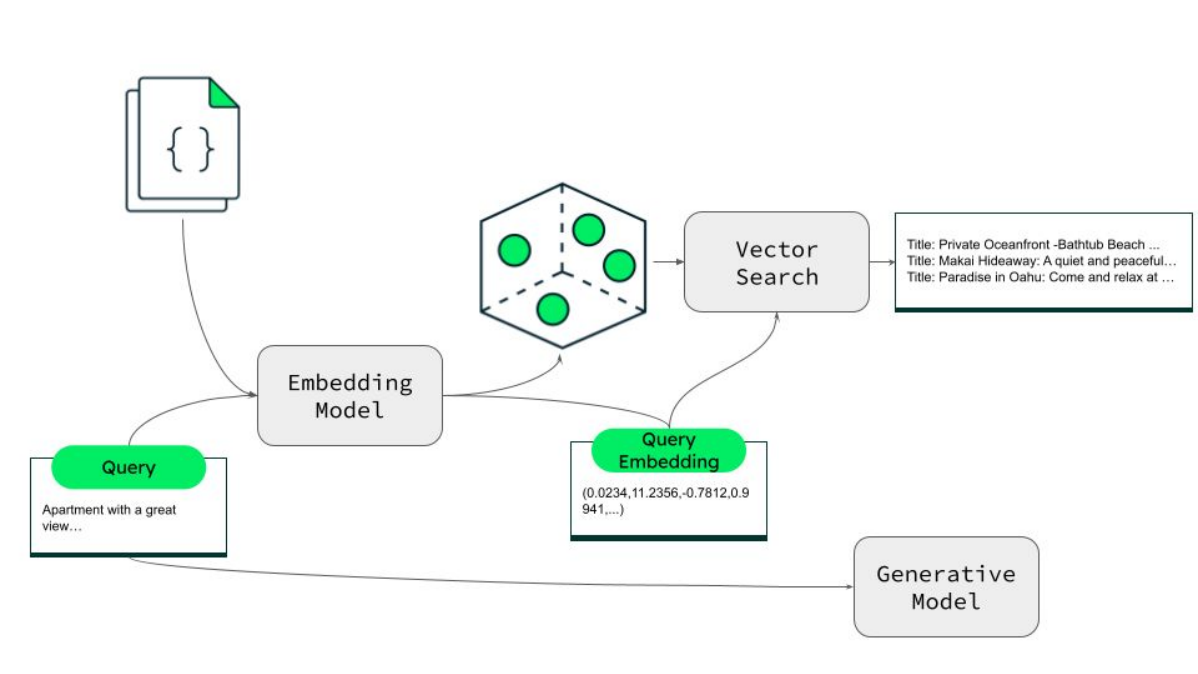


Figure 25: Together AI framework

Unlock the Power of Open-Source Embeddings for 12x Less

The Together Embeddings endpoint offers access to eight leading open-source embedding models at up to 12x cheaper price than proprietary alternatives. The list of the models includes top models from the MTEB leaderboard (Massive Text Embedding Benchmark), such as UAE-Large-v1 and BGE models, and state-of-the-art long context retrieval models. Together Embeddings also offers integrations to MongoDB Atlas, LangChain, and LlamaIndex for RAG.

To demonstrate this integration, the engineering team at Together AI created a tutorial for developers exploring how to build a RAG application with MongoDB Atlas. This tutorial shows how to use Together Embeddings and Together Inference to generate embeddings and language responses.

Putting Jina AI's Breakthrough Open Source Embedding Model To Work

Jina AI has swiftly risen as a leader in multimodal AI, focusing on prompt engineering and embedding models. With its commitment to open-source and open research, Jina AI is bridging the gap between advanced AI theory and the real world AI-powered applications being built by developers and data scientists. Over 400,000 users are registered to use the Jina AI platform.

Jina AI's work in embedding models has caught significant industry interest. As many developers now know, embeddings are essential to generative AI (gen AI). Embedding models are sophisticated algorithms that transform and embed data of any structure into multi-dimensional numerical encodings called vectors. These vectors give data semantic meaning by capturing its patterns and relationships. This means we can analyze and search for unstructured data in the same way we've always been able to with structured business data. Considering that over 80% of the data we create every day is unstructured, we start to appreciate how transformational embeddings — when combined with a powerful solution such as MongoDB Atlas Vector Search — are for gen AI.

> "Our Embedding API is natively integrated with key technologies within the gen AI developer stack including MongoDB Atlas, LangChain, LlamaIndex, Dify, and Haystack. MongoDB Atlas unifies application data and vector embeddings in a single platform, keeping both fully synced. Atlas Triggers keeps embeddings fresh by calling our Embeddings API whenever data is inserted or updated in the database. This integrated approach makes developers more productive as they build new, cutting-edge AI-powered apps for the business."
>
> **Dr. Han Xiao**, Founder and CEO at Jina AI

MongoDB | Jina

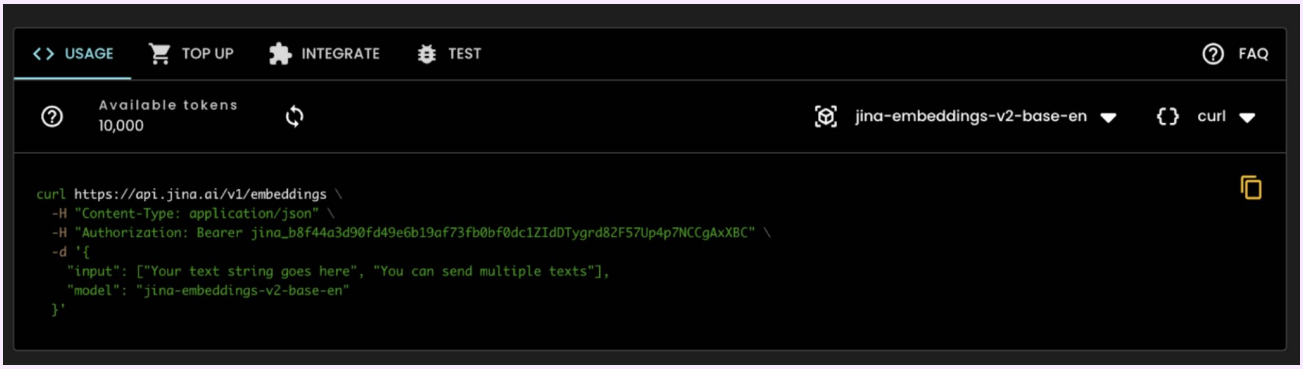


Figure 26: Jina AI's world-class embedding models improve search and RAG systems.

Jina AI's embedding models

Jina AI's jina-embeddings-v2 is the first open-source 8K text embedding model. Its 8K token length provides deeper context comprehension, significantly enhancing accuracy and relevance for tasks like retrieval-augmented generation (RAG) and semantic search. Jina AI's embeddings offer enhanced data indexing and search capabilities, along with bilingual support. The embedding models are focused on singular languages and language pairs, ensuring state-of-the-art performance on language-specific benchmarks. Currently, Jina Embeddings v2 includes bilingual German-English and Chinese-English models, with other bilingual models in the works.

Jina AI's embedding models excel in classification, reranking, retrieval, and summarization, making them suitable for diverse applications, especially those that are cross-lingual. Recent examples from multinational enterprise customers include the automation of sales sequences, skills matching in HR applications, and payment reconciliation with fraud detection.

In our published Jina Embeddings v2 and MongoDB Atlas article we show developers how to get started in bringing vector embeddings into their apps. The article covers:

1. Creating a MongoDB Atlas instance and loading it with your data. (The article uses a sample Airbnb reviews data set.)
2. Creating embeddings for the data set using the Jina Embeddings API.
3. Storing and indexing the embeddings with Atlas Vector Search.
4. Implementing semantic search using the embeddings.

SuperDuperDB: Bringing AI to your database

SuperDuperDB is an open-source Python package providing tools for developers to apply AI and machine learning on top of their existing data stores. Developers and data scientists continue to use their preferred tools, avoiding both data migration and duplication to specialized data stores. They also have the freedom to run SuperDuperDB anywhere, avoiding lock-in to any one AI ecosystem.

With SuperDuperDB developers can:

- Deploy their chosen AI models to automatically compute outputs (inference) in their database in a single environment with simple Python commands.
- Train models on their data simply by querying without additional ingestion and pre-processing.
- Integrate AI APIs (such as OpenAI) to work together with other models on their data effortlessly.
- Search data with vector search, including model management and serving.

"We integrate MongoDB as one of the key backend databases for our platform, the PyMongo driver for the app connectivity and Atlas Vector Search for storing and querying vector embeddings". "It therefore made sense for us to partner more closely with the company through MongoDB Ventures. We get direct access to the MongoDB engineering team to help optimize our product, along with visibility within MongoDB's vast ecosystem of developers."

Duncan Blythe, co-Founder SuperDuperDB

MongoDB | SuperDuperDB

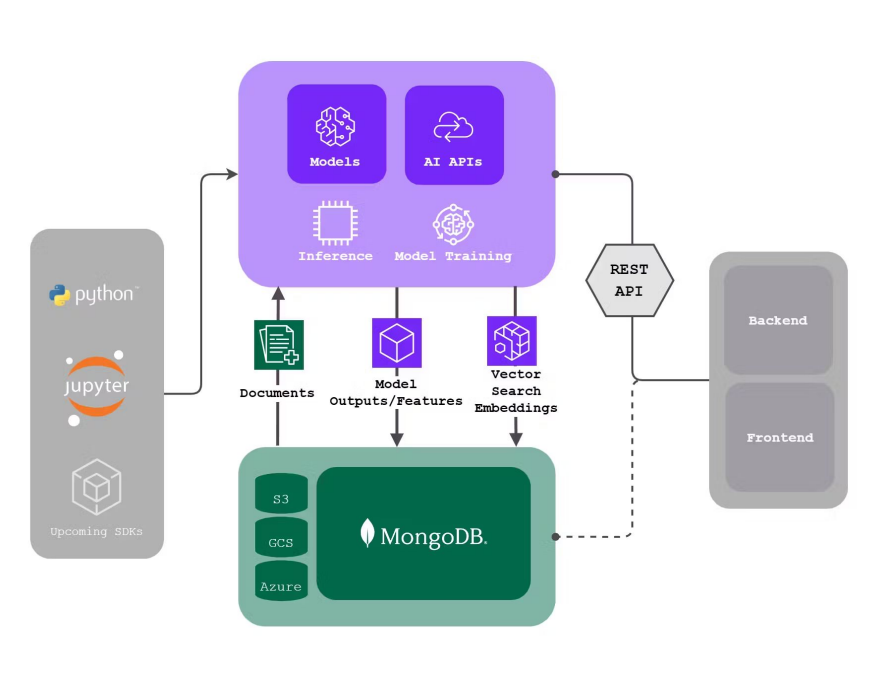


Figure 27: Reference architecture

Build Next-Generation AI Apps on Your Existing Database

SuperDuperDB provides an array of sample use cases and notebooks that developers can use to get started including vector search with MongoDB, multimodal search, retrieval augmented generation (RAG), transfer learning, and many more.

The team has also built an AI chatbot app that allows users to ask questions about technical documentation. The app is built on top of MongoDB and OpenAI with FastAPI and React (FARM stack) + SuperDuperDB. It showcases how easily developers can build next-generation AI applications on top of their existing data stores with SuperDuperDB. You can try the app and read more about how it is built at SuperDuperDB's documentation.

4149.AI: Maximizing team productivity with a hypertasking AI-powered teammate

4149.AI boosts team productivity with a dedicated AI teammate. In a successful private beta, nearly 1,000 teams leveraged this agent to streamline goal tracking and tasks. It analyzes team communication, identifies roadblocks, and takes action in Slack discussions, meetings, calls, reports, emails, and task trackers.

AI-powered team

4149.AI provides teams with their own AI-powered teammate that helps track goals and priorities.

No-code customization

There is a no-code way for people to customize and expand the functionality of their AI teammate.

Participation in tasks

The AI agent participates in various tasks such as joining slack threads, meetings, transcribing calls, generating summaries, responding to emails, and updating issue trackers.

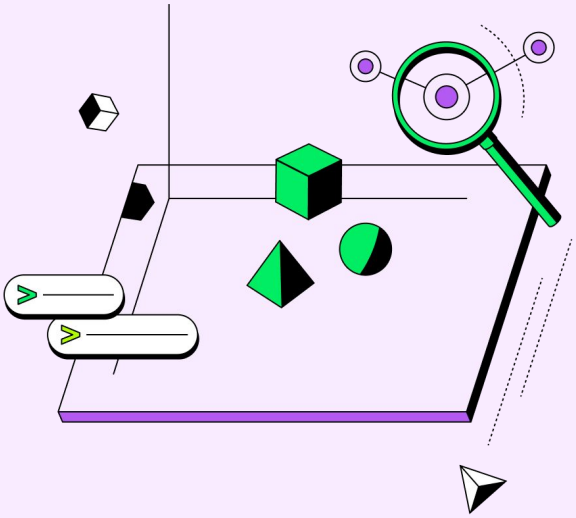
Ambitious Growth Strategy

4149.AI outlines an aggressive roadmap for its products, leveraging the power of chain-of-thought reasoning and multimodal capabilities in advanced language models.

The Power of Unified Data

- The ability to **store summaries and chat history alongside vector embeddings** in the same database accelerates developer velocity and the release of new features.
- The hybrid search capability of MongoDB Atlas allows pre-filtering data with keyword-based Atlas Search before semantically searching vectors, which **helps retrieve relevant information faster.**
- Being part of MongoDB's AI Innovators program provides 4149.AI with access to technical support and free Atlas credits, helping them quickly experiment using the native AI capabilities available in the MongoDB developer data platform.

MongoDB. | 4149 BETA



4149.AI helps teams get more work done by providing them with their very own AI-powered teammate. During the company's private beta program, the autonomous AI agent has been used by close to 1,000 teams to help them track goals and priorities. It does this by building an understanding of team dynamics and unblocking key tasks. It participates in slack threads, joins meetings, transcribes calls, generates summaries from reports and whitepapers, responds to emails, updates issue trackers, and more.

4149.AI uses a custom-built AI-agent framework leveraging a combination of embedding models and LLMs from OpenAI and AI21 Labs, with text generation and entity extraction managed by Langchain. The models process project documentation and team interactions, persisting summaries and associated vector embeddings into Atlas Vector Search. There is even a no-code way for people to customize and expand the functionality of their AI teammate. Over time, the accumulated context generated for each team means more and more tasks can be offloaded to their AI-powered co-worker.

The engineers at 4149.AI evaluated multiple vector stores before deciding on Atlas Vector Search. The ability to store summaries and chat history alongside vector embeddings in the same database accelerates developer velocity and the release of new features. It also simplifies the technology stack by eliminating unnecessary data movement.

Looking forward 4149.AI has an aggressive roadmap for its products as it starts to more fully exploit the chain-of-thought and multimodal capabilities provided by the most advanced language models. This will enable the AI co-worker to handle more creative tasks requiring deep reasoning such as conducting market research, monitoring the competitive landscape, and helping identify new candidates for job vacancies. The goal for these AI teammates is for them to eventually be able to take the initiative in what to do next rather than rely on someone to manually assign them a task.

Zelta.AI: Prioritizing product roadmaps with data-driven customer analytics

In the rapidly evolving digital economy, Zelta.AI stands as a beacon for product managers navigating the sea of customer feedback. Born out of the need to synthesize diverse feedback into coherent development plans, Zelta.AI is revolutionizing the way businesses prioritize their product roadmaps

Generative AI for Customer Insights

Zelta uses generative AI to communicate insights on top of customer pain points found in companies' most valuable asset: qualitative sources of customer feedback such as call transcripts and tickets.

Integration with Multiple Platforms

Zelta.AI has the capability to pull data directly from multiple platforms like Gong, Zoom, Fireflies, Zendesk, Jira, Intercom, among others.

Processing Unstructured Data

Zelta leverages Language Models (LLMs) to process unstructured data and returns actionable insights for product teams.

Real-Time Product Feedback Trends

Zelta.AI offers real-time product feedback trend reporting, enabling faster decisions for product teams, enhancing its value.

Choice of MongoDB as Developer Data Platform

- MongoDB provides Zelta with the **flexibility** to constantly experiment with new features. They can add fields and evolve the data model as needed without any of the expensive schema migration pains imposed by relational databases.
- Zelta makes heavy use of the MongoDB aggregation pipeline for application-driven intelligence. Without having to ETL data out of MongoDB, they can analyze data in place to provide customers with **real-time dashboards and reporting of trends in product feedback.**
- Looking forward, as Zelta plans on creating its **own custom models**, MongoDB will prove invaluable as a source of labeled data for supervised model training.

MongoDB | Zelta

Figure 31: Zelta leverages LLMs to process unstructured data and returns actionable insights for product teams

The company's engineering team uses a combination of fine-tuned OpenAI GPT-4, Cohere, and Anthropic models to extract, classify, and encode source data into trends and sentiment around specific topics and features. MongoDB Atlas is used as the data storage layer for source metadata and model outputs.

"The flexibility MongoDB provides us has been unbelievable. My development team can constantly experiment with new features, just adding fields and evolving the data model as needed without any of the expensive schema migration pains imposed by relational databases."

Mick Cunningham, CTO and Co-Founder at Zelta AI

"We also make heavy use of the MongoDB aggregation pipeline for application-driven intelligence. Without having to ETL data out of MongoDB, we can analyze data in place to provide customers with real-time dashboards and reporting of trends in product feedback. This helps them make product decisions faster, making our service more valuable to them."

Mick Cunningham, CTO and Co-Founder at Zelta AI

Crewmate: Helping brands connect with their communities

Crewmate is a no-code builder for embedded AI-powered communities. The company's builder provides customizable communities for brands to deploy directly onto their websites. Crewmate is already used today across companies in consumer packaged goods (CPG), B2B SaaS, gaming, Web3, and more.

Customizable Communities

Crewmate creates AI-powered communities for SaaS firms, boosting sales, retention, and engagement. Users can interact, share insights, and discuss your product.

Real-Time Data Pipelines

Crewmate implements event-driven pipelines to ensure that community content remains fresh and up-to-date.

Context-Aware Semantic Search

Crewmate's platform features context-aware search powered by Atlas Vector Search, delivering relevant content to users on brand community pages.

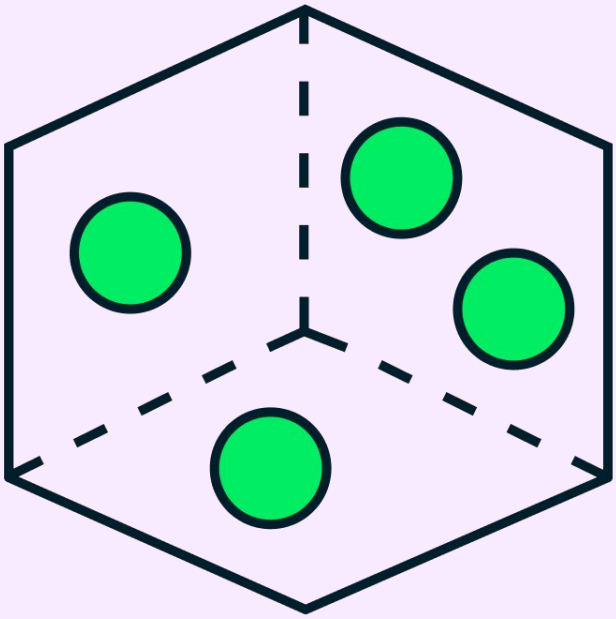
Insightful Analytics for Brands

Crewmate enables brands to extract valuable insights from user engagement data using its powerful analytics capabilities.

Choice of MongoDB as Developer Data Platform

- MongoDB Atlas provides integrations with the fast-evolving AI ecosystem. Crewmate leverages this capability to **easily integrate with other AI models**, such as OpenAI's ada-002 and potentially other models like Llama in the future.
- Crewmate utilizes MongoDB's Query API to process, aggregate, and analyze user engagement data. This allows brands to **track community outreach efforts and conversions directly from the app data stored in MongoDB**, without the need to extract, transform, and load (ETL) it into a separate data warehouse or data lake.
- Crewmate utilizes Atlas Vector Search, a feature provided by MongoDB Atlas, to power **context-aware semantic search**. This enables users visiting a brand's website to automatically access relevant content such as social media posts, forum discussions, job postings, and special offers.

MongoDB | crewmate



Personalized Community Content with Atlas Vector Search

Using context-aware semantic search powered by Atlas Vector Search, users hitting and browsing the community pages on a brand's website are automatically served relevant content. This includes posts from social media feeds, forum discussions, job postings, special offers, and more.

"I've used MongoDB in past projects and knew that its flexible document schema would allow me to store data of any structure. This is particularly important when ingesting many different types of data from my clients' websites,"

Raj Thaker, CTO and Co-Founder of Crewmate.

Thaker goes on to say, "The introduction of Atlas Vector Search and the Building Generative AI Applications tutorial gave me a fast, ready-made blueprint that brings together a database for source data, vector search for AI-powered semantic search, and reactive, real-time data pipelines to keep everything updated, all in a single platform with a single copy of the data and a unified developer API. This keeps my engineering team productive and my tech stack streamlined. Atlas also provides integrations with the fast-evolving AI ecosystem. So while today I'm using OpenAI models, I have the flexibility to easily integrate with other models, such as Llama, in the future."

MongoDB. | crewmate

Video personalization at scale with Potion and MongoDB

Potion enables salespeople to personalize prospecting videos at scale. Already over 7,500 sales professionals at companies including SAP, AppsFlyer, CaptivateIQ, and Opensense are using SendPotion to increase response rates, book more meetings, and build customer trust.

Effortless Video Creation

Sales representatives simply record a video template and select the elements they want to personalize. These elements typically include details like the recipient's name, company, and desired call-to-action.

Bulk Transformation

Imagine turning a single video template into over 1,000 unique video messages, each tailored to an individual contact. Potion achieves this by efficiently reanimating videos in bulk, saving time and effort for sales teams.

Efficient Outreach

With Potion, you can engage, convert, and leave a lasting impact on your prospects. It eliminates the need for manual video recording, streamlining your communication efforts.

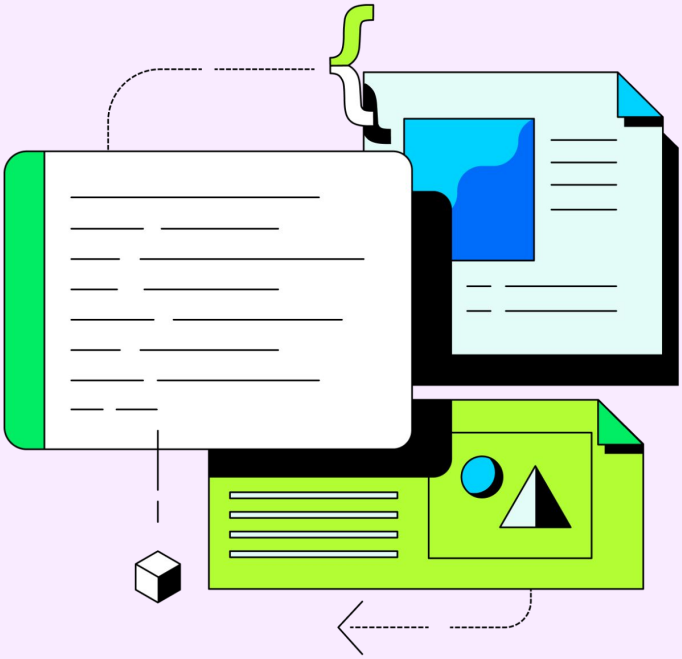
AI Models and Technologies

Potion's custom generative AI models are built using PyTorch and TensorFlow. Their vision model is trained on thousands of faces, allowing them to synthesize videos without individualized AI training. Audio models are tuned on-demand for each voice.

"We use the MongoDB database to store metadata for all the videos, including the source content for personalization, such as the contact list and calls to action. For every new contact entry created in MongoDB, a video is generated for it using our AI models, and a link to that video is stored back in the database. MongoDB also powers all of our application analytics and intelligence. With the insights we generate from MongoDB, we can see how users interact with the service, capturing feedback loops, response rates, video watchtimes, and more. This data is used to continuously train and tune our models in Sagemaker."

Kanad Bahalkar, co-Founder & CEO at Potion

MongoDB | Potion



Scaling Potion with MongoDB Atlas

On selecting MongoDB Kanad says, “I had prior experience of MongoDB and knew how easy and fast it was to get started for both modeling and querying the data. Atlas provides the best-managed database experience out there, meaning we can safely offload running the database to MongoDB. This ease-of-use, speed, and efficiency are all critical as we build and scale the business."

To further enrich the SendPotion service, Kanad is planning to use more of the developer features within MongoDB Atlas. This includes Atlas Vector Search to power AI-driven semantic search and RAG for users who are exploring recommendations across video libraries. The engineering team is also planning on using Atlas Triggers to enable event-driven processing of new video content.

Potion is a member of the MongoDB AI Innovators program. Asked about the value of the program, Kanad responds, “Access to free credits helped support rapid build and experimentation on top of MongoDB, coupled with access to technical guidance and support."

MongoDB. | Potion

Artificial Nerds: The power of custom voice bots without the complexity of fine-tuning

Artificial Nerds, founded in 2017, is a software company that unlocks the potential of AI for businesses through a suite of intelligent virtual assistants. Their custom voice bots streamline customer interactions, allowing teams to focus on building meaningful relationships.

Human-Like Conversations

Artificial Nerds' AI bots are designed to create fluid and personalized conversations with customers. Unlike traditional bots with scripted responses, Artificial Nerds' innovative tools ensure a more natural and user-centric experience.

No-Code Builder

Their platform enables easy adjustments to chatbots without coding. Using cloud tech and templates, businesses can act quickly, cutting development time.

Voicebot Integration

Artificial Nerds offers voicebots that automate phone calls for businesses. These voicebots understand user language and provide solutions, making business interactions more efficient and effective.

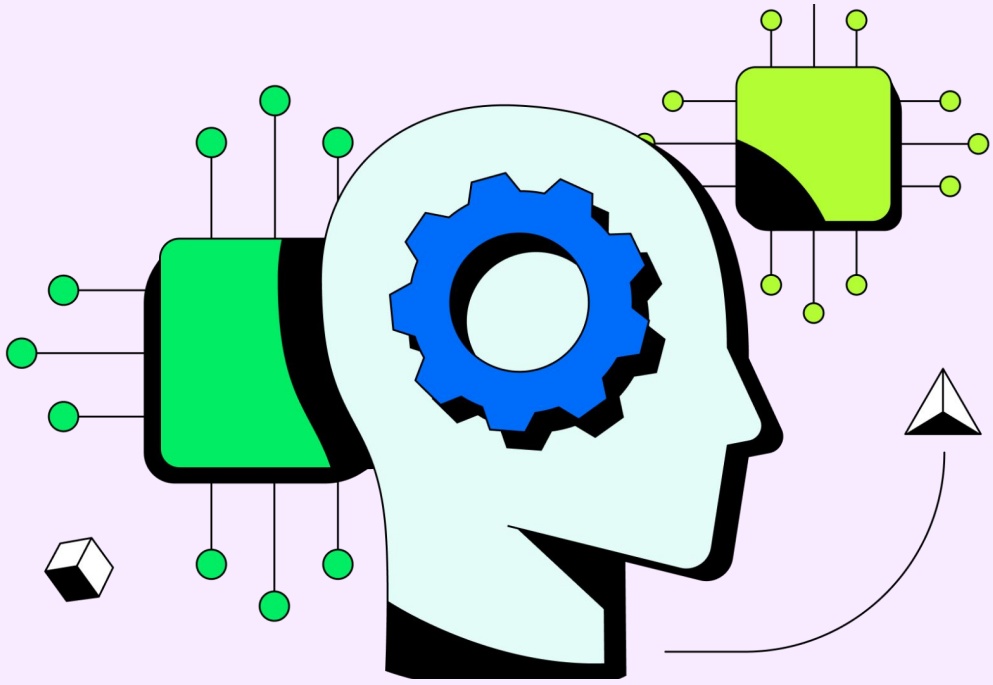
Contact Center Enhancement

Artificial Nerds streamlines tasks, boosting efficiency and freeing up teams for strategic work. Their advanced bots offer visibility, control, and real-time adjustments, all without developer intervention.

Enabling Flexible Data Storage for AI-Powered Products

- Artificial Nerds chose MongoDB for its **flexible schema**, which allows them to store richly structured conversation history, messages, and user data. This flexibility is crucial for a company focused on AI-powered products, as it enables them to adapt and evolve their data structures as needed to support their evolving suite of products and services.
- By **eliminating the need for a separate search engine and ETL**, MongoDB Atlas reduces the complexity of development and management. This allows developers to focus on building their application without worrying about maintaining separate data stores.

MongoDB | Artificial Nerds



AI-Fueled Search and Innovation: Artificial Nerds Speeds Up with MongoDB Atlas

By adopting Atlas Search, the company streamlined its search capabilities, integrating a powerful full-text index directly onto its database collections. This eliminated the need for separate search engines and ETL mechanisms, reducing cognitive overhead. Similarly, the release of Atlas Vector Search further enhanced efficiency by replacing a standalone vector database with MongoDB Atlas, resulting in improved developer productivity and a 4x reduction in latency for a better customer experience.

Artificial Nerds is growing fast, with revenues expanding 8% every month. The company continues to push the boundaries of customer service by experimenting with new models including the Llama 2 LLM and multilingual sentence transformers hosted in Hugging Face. Being part of the MongoDB AI Innovators program helps Artificial Nerds stay abreast of all of the latest MongoDB product enhancements and provides the company with free Atlas credits to build new features.

MongoDB. | Artificial Nerds

Algomo: Conversational support, powered by generative AI

Algomo uses generative AI to help companies offer their best service to both their customers and employees across more than 100 languages. The company's name is a portmanteau of the words Algorithm (originating from Arabic) and Homo, (human in Latin). It reflects the two core design principles underlying Algomo's products:

AI Agents with Human-Level Reasoning

Algomo provides AI agents with human-like understanding and decision-making capabilities, enhancing customer service by efficiently managing tasks and seamlessly transitioning complex issues to support teams.

Efficient Help Desk

Algomo's Helpdesk integrates teams, channels, and data into a single workspace, simplifying support operations. Furthermore, Algomo's AI operates in Co-Pilot Mode, offering suggestions to enhance the efficiency of customer service teams.

Personalized Interactions

Algomo's AI chatbot delivers personalized interactions, tailoring content to individual customers, posing clarifying questions, and capable of communicating in over 100+ languages.

Omnichannel Support

Algomo's Messenger offers customization options to align with any brand and enables the reception of messages from multiple channels, such as email, WhatsApp, and social media.

Alamo Optimizes Support with MongoDB Atlas

- Alamo chose MongoDB due to its **flexible document data model**, allowing them to store customer data alongside conversation history and messages, ensuring long-term memory for context and continuity in support interactions.
- MongoDB Atlas as a fully managed cloud service relieves Alamo's team from operational heavy lifting, enabling them to **focus on building conversational experiences** rather than managing infrastructure.
- Alamo's engineers are considering Atlas Vector Search as a replacement for their current standalone vector database. This move not only **reduces costs but also simplifies their codebase** by eliminating the need to synchronize data across two separate systems.

MongoDB | Algomo

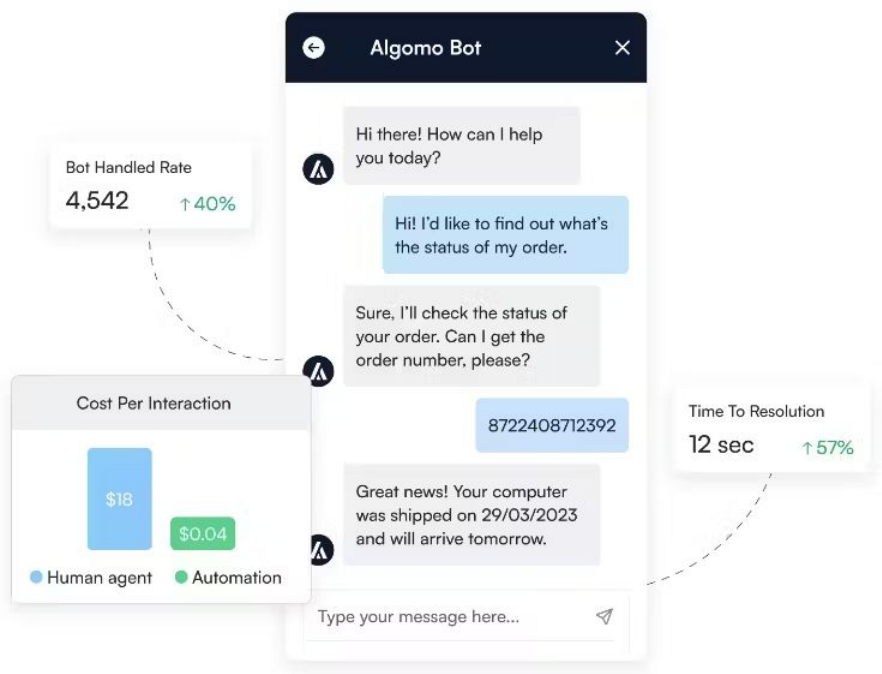


Figure 33: Algomo Bot

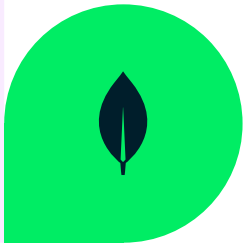
Unlocking Personalized Customer Experiences with Algomo's Conversational AI

With Algomo, customers can get a ChatGPT-powered bot up on their site in less than 3 minutes. More than just a bot, Algomo also provides a complete conversational platform. This includes Question-Answering text generators and autonomous agents that triage and orchestrate support processes, escalating to human support staff for live chat as needed. It works across any communication channel from web and Google Chat to Intercom, Slack, WhatsApp, and more.

Customers can instantly turn their support articles, past conversations, slack channels, Notion pages, Google Docs, and content on their public website into personalized answers. Algomo vectorizes customer content, using that alongside OpenAI's ChatGPT. The company uses RAG (Retrieval Augmented Generation) prompting to inject relevant context to LLM prompts and Chain-Of-Thought prompting to increase answer accuracy. A fine-tuned implementation of BERT is also used to classify user intent and retrieve custom FAQs.

Conclusion

Across industries, AI has captured the imaginations of executives and consumers alike. Whether you're a customer of a bank, insurance company, telecommunications enterprise, or retail conglomerate, AI has and will transform and enhance the way you do business with corporations. For the industries 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 industries' 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.

Next Steps with MongoDB

MongoDB's unique blend of speed, flexibility, and robust security offers a compelling proposition for organizations building AI-enriched applications. Our ability to provide a scalable, resilient, and efficient data management solution, deployment flexibility and support for multi-cloud strategies positions MongoDB as a leader for intelligent applications.

The conversation about leveraging MongoDB within industries doesn't end here. We invite you to delve deeper into MongoDB's capabilities and offerings to discover how you can build the future of AI applications.

Contact us today and **click here for MongoDB Industry Solutions**.

Innovation workshops

Sign up for our MongoDB Atlas for Industries program to take advantage of innovation workshops and more.

AI resources

Get full access to our resources to Build AI-powered Apps including articles, reports, case studies and more.