



Vol.3
Sept 2024

Innovate With AI: The Future Enterprise

Driving industry success
with artificial intelligence
and MongoDB Atlas



Atlas for Industries

*A look at how AI and
MongoDB are creating
value across industries*

AI and Application Modernization

*Modernizing apps faster for
a better user experience*

Table of Contents

3	<u>Editor's Introduction</u>	68-71	Assessing Business Loan Risks
4	<u>AI and the Developer Data Platform</u>	72-73	Emerging use cases for AI in payments
7-28	<u>Manufacturing and Motion</u>	74-76	Regulatory Compliance and Code Change Assistance
8	AI Investments	77	Other Notable Use Cases
9-15	Inventory Management	77	Contact Information
16-24	Predictive Maintenance		
25-27	Knowledge Management		
28	Other Notable Use Cases		
28	Contact Information		
29-37	<u>Telecommunications and Media</u>	78-93	<u>Insurance</u>
30-31	AI Investments	79-84	Architecting with AI
32	Service Assurance	84-86	Underwriting & Risk Management
33	Fraud Detection and Prevention	86-87	Claim Processing
34-35	Content Discovery	88	Customer Experience
36	Search Generative Experiences (SGE)	89-92	Domain-Driven AI & Solution Library
37	Other Notable Use Cases	93	Other Notable Use Cases
37	Contact Information	93	Contact Information
38-51	<u>Retail</u>	94-110	<u>Healthcare and Life Sciences</u>
39-40	AI Investments	95-96	Potential of Generative AI
41-43	AI-Augmented Search and Vector Search	97-98	Patient Experience and Engagement
43-46	Personalized Marketing & Content Generation	99-100	Enhanced Clinical Decision Making
46-48	Demand Forecasting & Predictive Analytics	101-103	Enhancing Medical Imaging with Generative AI
49-51	Conversational Chatbots	104-105	Revolutionizing Clinical Study Reports
51	Contact Information	106-109	Accelerating Drug Discovery
		110	Other Notable Use Cases
		110	Contact Information
52-77	<u>Financial Services</u>	111-121	<u>AI Leaders in Action</u>
53-54	AI Investments and Challenges	112-116	Ben Ellencweig from McKinsey
55-57	Relationship Management Support with Chat Bots	117-121	Asaf Somekh from Iguazio
58-59	Risk and Fraud Prevention		
60-61	Financial Document Search and Summarization	122-132	<u>AI and Application Modernization Programs</u>
61-63	ESG Analysis	133-225	<u>AI Partnerships</u>
64-67	Transforming Credit Scoring with AI	134-137	Hyperscaler cloud services
		138-143	System Integrators
		144-190	AI With SaaS
		191-225	Component-Based AI
		226-227	<u>Conclusion and Next Steps</u>

Editor's Introduction

I am delighted to present our third edition of the art of possible at the the dynamic intersection of AI, innovation, and industry solutions. This e-book 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. I strongly suggest to browse use cases from other industries - often they are applicable in a changed context as well.

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 and our MongoDB AI Application Program (MAAP). Whether a SaaS end-to-end solution you can implement, or component-based solution you can plug in, there is something here for you.

We invite you to take advantage of our [innovation workshops](#). These ideation sessions are available to you at your location in your environment. You will meet MongoDB industry experts, our partners and discuss the art of the possible based on your requirements and needs. Additionally check out our [MongoDB AI Applications Program](#) on the MongoDB website designed to help organizations rapidly build and deploy modern applications enriched with AI technology at enterprise scale.

Boris Bialek: Vice President and 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 such as OpenAI and Hugging Face opened possibilities for generating and analyzing text data. Today, generative AI (gen AI) applications go beyond text data, creating hyper-personalized experiences.

While implementing AI technology can be risky, complex, and time-consuming, the potential for benefits such as 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 gen 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 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' worth of cryptocurrency daily using machine learning (ML) models 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 gen 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

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 gen AI, this flexibility enables developers to define additional filters on vector-based queries, such as combining metadata, aggregations, and geospatial

LEARN MORE

Real-Time Analytics and Dynamic Pricing

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 data.

With [MongoDB Atlas App Services](#), these industries are able to reinvent pricing strategies to reflect 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 gen AI really is.

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).

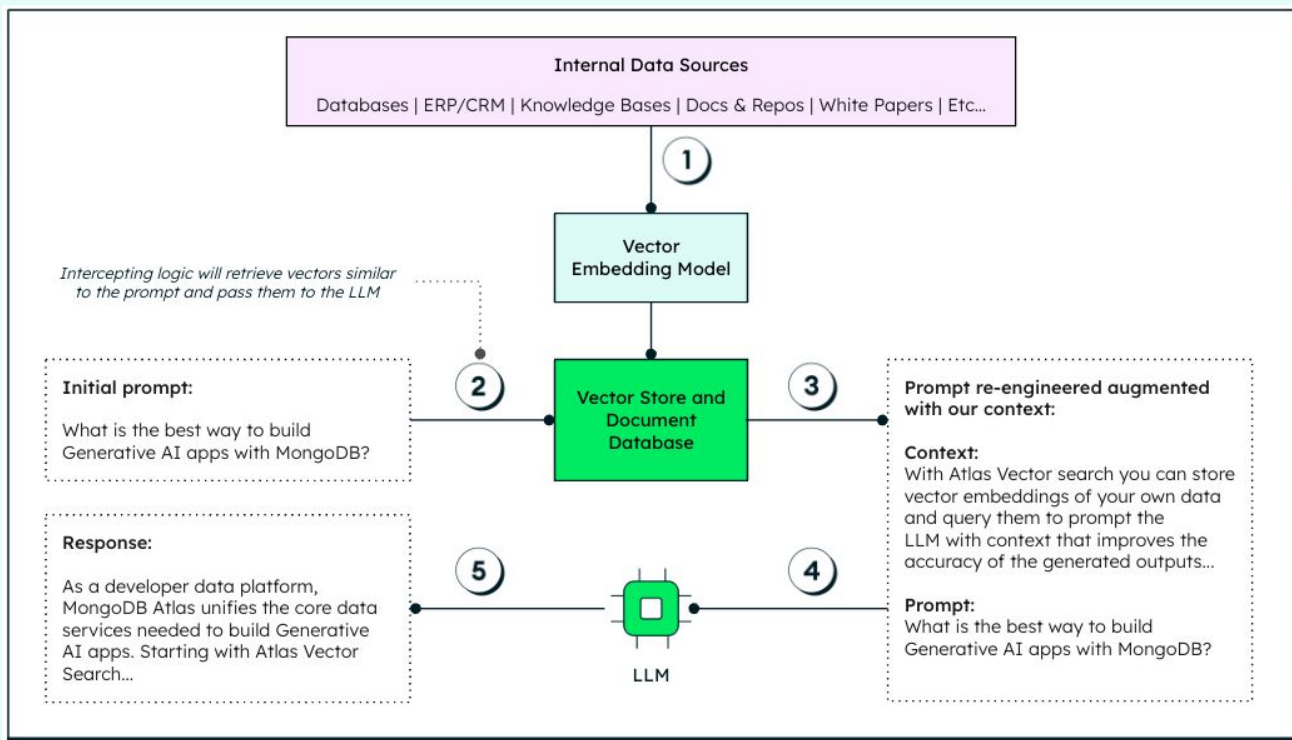


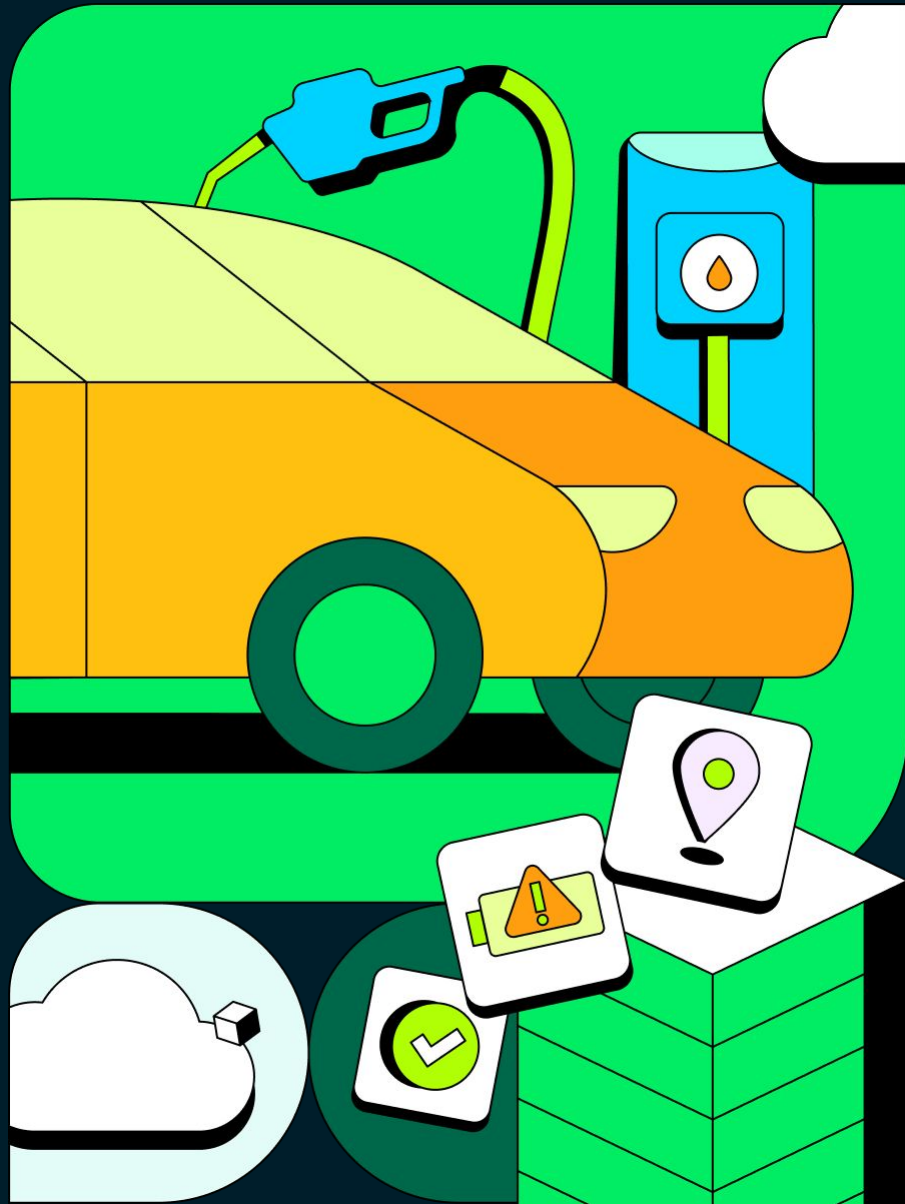
Figure 1: Data is transformed from unstructured internal sources to actionable, impactful insights



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.



AI investments in the manufacturing industry



AI in manufacturing is a game-changer. It has the potential to transform performance across the breadth and depth of manufacturing operations. Companies are in a race to embrace AI, as these technologies are critical enablers of the Fifth Industrial Revolution (also known as Industry 5.0). Artificial intelligence in manufacturing is bringing factories into the future and will ultimately empower the manufacturing market to continue to be the backbone of the global economy.

Industry-wide, manufacturers are facing a range of challenges that make it difficult to speed production while still providing high-value and high-quality products to their customers. All the while, companies need to implement a digital infrastructure that positions them to fully embrace the skills and knowledge of their best assets — people.

The manufacturing industry today relies on automation just as much as people. But the factory of the future, which is a marriage of physical and digital capabilities, requires more: real-time data, connectivity and AI technology at the forefront. In fact, [more than 80%](#) of C-suite executives believe they must leverage AI to achieve their growth objectives. Customer requirements for delivering on-time and on-budget product are of the utmost importance, and efficiency is a goal in everything.

AI's ability to drive impact in this regard is real. According to a study of Capgemini, three use cases stand out in terms of their suitability for kickstarting a manufacturer's AI journey: demand planning, intelligent maintenance and product quality control. These use cases have

an optimal combination of several characteristics, that make them an ideal place to start:

- Clear business value/benefits
- Relative ease of implementation
- Availability of data

Smart manufacturing use cases are revolutionizing many organizations, and a key driver of this is the incorporation of artificial intelligence into manufacturing processes. Many firms have embarked on significant digital transformation journeys in the past two years with the goal of improving efficiency and resilience. However, a concerning gap exists between tech adoption and return on investment. While 89% of organizations have begun digital and AI transformations, only 31% have seen the expected revenue lift, and only 25% have realized the expected cost savings ([McKinsey](#)). In some cases, situations have even worsened.

This highlights the importance of not just implementing new technologies, but implementing them strategically. In other words, simply deploying AI isn't a guaranteed path to success. Manufacturers need to carefully consider how AI can address their specific challenges, and then integrate it into existing processes effectively.

This chapter unpacks how major players in the manufacturing industry are leveraging AI to improve operations, deliver better outcomes for customers, and realize innovation. It delves into three high impact value drivers and AI use cases: Predictive Maintenance, Inventory Management and Knowledge Management.

The path to success



Successful organizations exhibit common traits across five key areas:

- **Identifying high-impact value drivers and AI use cases:** Efforts should be concentrated on domains where artificial intelligence yields maximal utility rather than employing it arbitrarily.
- **Aligning AI strategy with data strategy:** Organizations must establish a strong data foundation with a data strategy that directly supports their AI goals.
- **Continuous data enrichment and accessibility:** High-quality data, readily available and usable across the organization, is essential for the success of AI initiatives.
- **Empowering talent and fostering development:** By equipping their workforce with training and resources, organizations can empower them to leverage AI effectively.
- **Enabling scalable AI adoption:** Building a strong and scalable infrastructure is key to unlocking the full potential of AI by enabling its smooth and ongoing integration across the organization.

Inventory Management and Optimization

Current State and Challenges

Modern manufacturing supply chains are complex systems, interconnected across the globe. 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. While maintaining higher inventory levels allows for suppliers to deal with unexpected fluctuations in demand, they come with higher inventory holding costs that may be passed on to customers. Thus, every player in the supply chain is motivated to strike a

balance between inventory levels to maximize profitability and competitive advantage in the market. Effective inventory management mitigates the risk of 'bullwhip effect', where sudden demands can disrupt the supply chain costs and performance.

Key components of supply chain management include procurement and sourcing, manufacturing and production, distribution, logistics and retail. Technological advancements including IoT and AI (including Gen AI) are being integrated into SCM to improve

transparency, efficiency and adaptability of the supply chain, allowing for real time monitoring, predictive analytics, and enhanced decision-making capabilities.

technology-driven ecosystem that requires collaboration throughout the supply chain

between OEMs, tier1-n suppliers and customers, always aiming for reduced costs, quicker production and response times and heightened customer satisfaction, all of which will result in stronger market position.

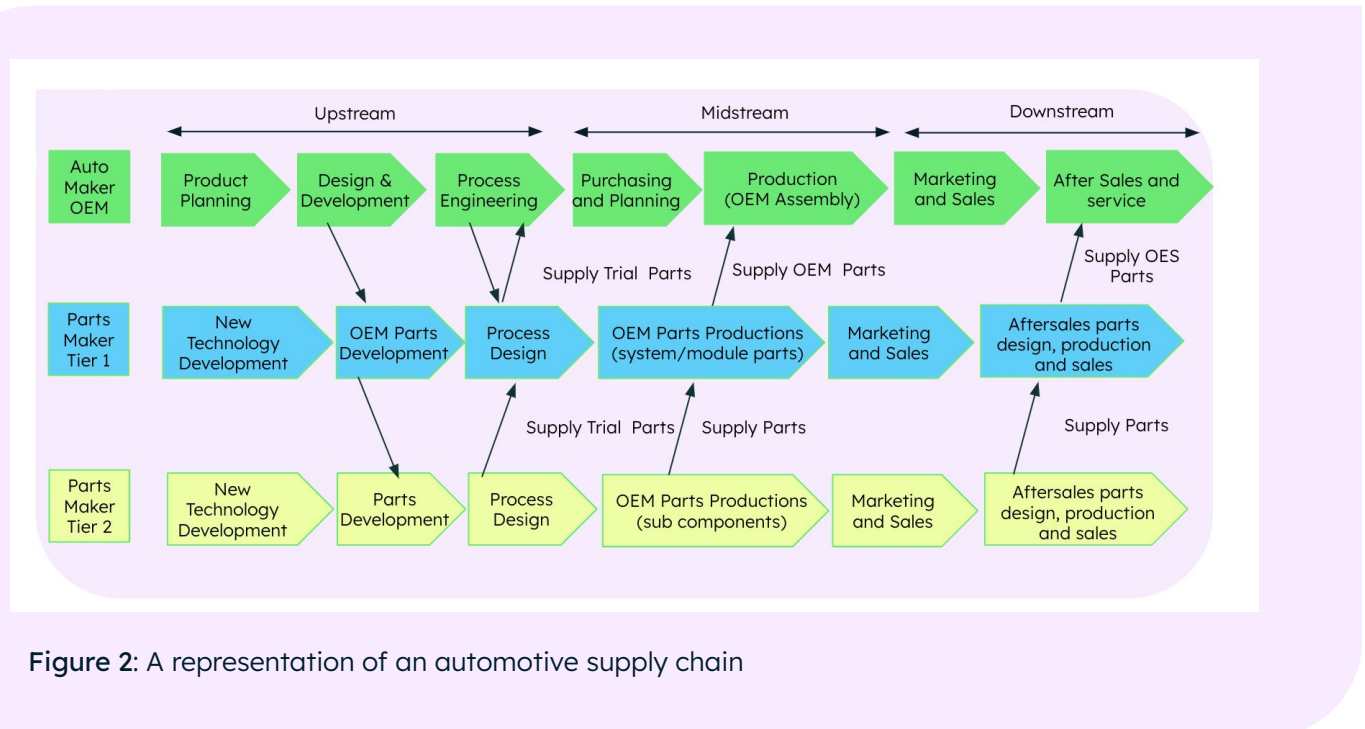


Figure 2: A representation of an automotive supply chain

Inventory management is essential for achieving the goals of efficient supply chains, controlling costs, and delivering to customers with minimal delays. Inventory management is primarily concerned with the planning and controlling an industry's inventory. It includes business processes such as estimating material requirements at various points in the supply chain, determining necessary material's amount, ordering frequency, and safety stock levels. It also includes inventory visibility, inventory forecasting, lead time management, inventory shipping costs, inventory valuation, forecasting future inventory prices, available physical space, quality management, returns and defective goods, and demand forecasting. It plays a very important role in reducing overall costs and rapid response objectives. Effective inventory management requires the

right inventory in the right place at the right time to minimize system costs and meet customer needs.

Usually, companies do supply chain planning at several levels, namely strategic, tactical, and operational. Each level differs in its objectives, planning horizon, and level of detail. Strategic and tactical planning are crucial to successful supply chain management. The so-called Pareto's law could be applicable here as 20% of efforts in strategic and tactical planning brings 80% of the total effect. At the strategic level, the leadership team makes high-level decisions that affect the entire organization. Scenario planning is done at this level. The analysts go through scores of internal and external data including global news, political developments, think tank studies and scientific

literature to pinpoint the most strategic concerns and trends that the organization needs to focus on. The team can then use these outputs to develop a set of draft scenarios for consideration.

This tedious process comes with its own set of challenges. Predicting long-term demand, market trends and economic conditions is challenging because of the long-term horizon. This long planning horizon increases the uncertainty in predicting demand, as market conditions, consumer preferences, and technological advancements can change significantly over time.

At the tactical and operational level, for manufacturers, to manage and optimize inventory levels, the first step is to maintain an accurate and real-time view of inventory levels across multiple plants, warehouses, and suppliers. This is absolutely essential as without having visibility on the current inventory levels, it is impossible to optimize. The second step is to reduce inventory carrying costs while still ensuring that the required parts and materials are available to ship out when needed. Finally, the data from multiple customers needs to be aggregated and analyzed despite being in different formats, each with its own unique lead times and order quantities.

However, efficient inventory management for manufacturers presents complex data challenges too, primarily in forecasting demand accurately and optimizing stock levels. One issue routinely faced is the variability and unpredictability of customer demand patterns, making it difficult to precisely anticipate inventory needs. Managing diverse data streams from sales records, production schedules, supplier information, and market trends poses a considerable data

integration challenge. The spread of data across multiple systems and locations (on-prem systems, cloud regions etc.) can lead to data silos and hinder visibility into overall inventory levels and movements. Finally, when there is sparse historical inventory data available, then traditional ML models may suffer in accuracy.

Generative AI and IoT technologies hold potential to address some of these challenges. Generative AI in particular can assist in scenario planning by generating various potential outcomes based on a wide range of data, allowing the organization to prepare better for an uncertain future.

How AI and MongoDB Help

We will start with scenario planning to generate hypothetical situations which could affect inventory requirements, supply chain performance and overall operations. Effective scenario planning helps companies plan for optimal inventory levels.

After scenario planning, we will look into inventory classification use cases where AI can be used to categorize inventory based on factors such as demand variability, lead times and criticality.

Scenario planning

A scenario planning process has two stages

- Scenario generation
- Strategy generation

In scenario generation, a generative AI application can look at a vast amount of data including internal and external business data, competition data, political news and events and social media news, find the correlation between each piece of unstructured information and then rank these areas of concern in terms of their estimated

significance. It is important to utilize general knowledge that a Large Language Model processes as well as internal company data in a Retrieval Augmented Generation (RAG) model to avoid hallucinations.

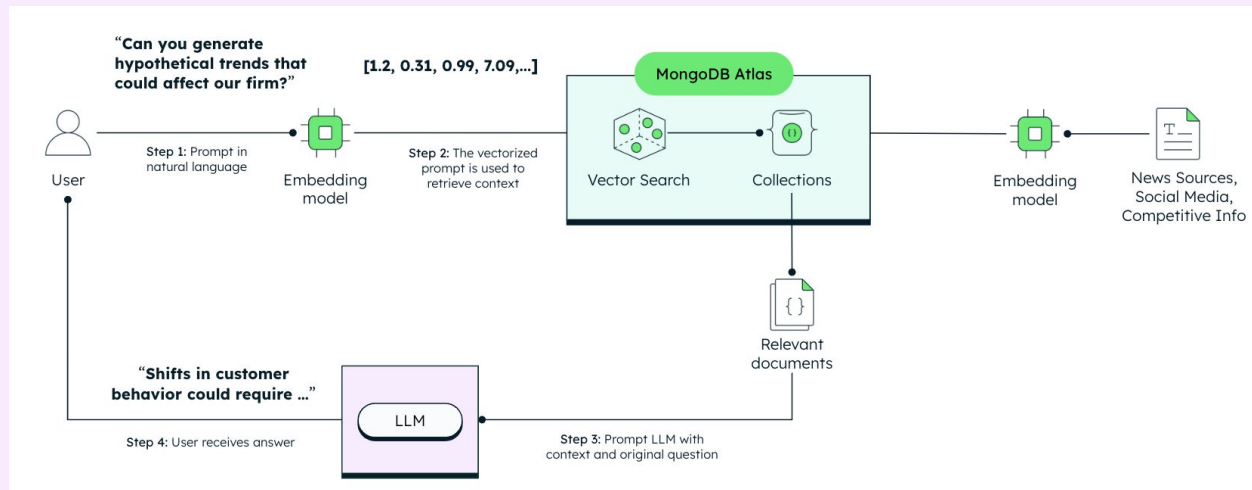


Figure 3: A Scenario Generation RAG App with MongoDB

The figure above shows a scenario generation application data flow. This retrieval-augmented generation (RAG) system consists of four parts. First, an AI data store aggregates and operationalizes structured and unstructured data. In our case, the majority of the data is in unstructured format such as news sources, social media and competitive information typically stored as unstructured PDF files. The PDF files are vectorized and stored in MongoDB Atlas. Atlas Vector Search is then utilized to perform semantic searches and to find meaningful context from the PDF embeddings.

[Atlas Vector Search](#) can be triggered using an AI application, connecting to MongoDB Atlas to retrieve the right context, which is then fed into the large language model to answer questions like "Can you generate hypothetical trends that can affect our firm?" The response might suggest customer behavior shifts or economic

factors due to certain reasons, including but not limited to the political landscape and global supply chain disruptions etc.

MongoDB Atlas streamlines RAG implementation as it handles everything under the hood. App data, metadata, context data, and vectors are all stored in the same place. As the app evolves, the document model is inherently flexible and ideal for storing structured and unstructured data. You can add data to the same collections inside the database as needed, vectorize it, and store the vectors alongside it.

Once the data is stored, vector search capabilities are provided right out of the box, and search operations can be optimized using dedicated search nodes.

With MongoDB Atlas, it's just one query in one technology, one language, and one infrastructure to manage and scale, with no data duplication, ultimately leading to a lower total cost of ownership and a unified developer experience.

The same application can then categorize the trends in terms of their probability of occurrence and impact. Next, a strategy can be created automatically, which provides guidance on what steps to take in response to the trends and scenarios generated. The strategy can also contain information about the financial implications and risks associated with the response, for example any significant initial investment in hiring more AI scientists to develop RAG applications for inventory management.

For scenario planning, generative AI can be used to generate and evaluate strategies. However, it is important to provide enough context to the LLM so that it does not hallucinate. MongoDB Atlas Vector Search is key to creating a RAG application. Additionally, these AI-generated strategies should be viewed as initial concepts for further exploration rather than as final solutions to be adopted without additional analysis. Generative AI should complement human efforts by supporting the identification, evaluation and timely execution of appropriate strategies.

Inventory Optimization

One of the most significant applications of AI in inventory management is in demand forecasting. AI algorithms can be used to analyze complex datasets to predict future demand of 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, market trends, manufacturers can determine the most efficient stock levels. AI systems can also place orders automatically based on predicted demand and targeted stock levels. This automation not just saves time but also reduces human error. Finally, AI utility can be extended to supplier selection and relationship management. By analyzing supplier performance data, AI based software can assist in choosing the right suppliers who will meet the company's quality, delivery and cost requirements.

MongoDB Atlas provides a flexible, scalable, and highly available developer data platform for managing inventory data. The document data model can handle complex inventory structures and hierarchies, making it easy to manage inventory across multiple plants and suppliers.

At the warehouse, the inventory can be scanned using a mobile device. This data can be persisted in a MongoDB collection. Once data is in Atlas, it can serve as the central repository for all inventory-related data which includes stock, supplier, and customer information, bill of materials and production line data. This repository becomes the source of data for the inventory management AI applications. This approach removes data silos and improves visibility into overall inventory levels and movements.

However, the challenge of poor or sparse data at the source systems may still remain. To solve this, manufacturers can take advantage of generative AI and Atlas Vector Search to implement a Retrieval Augmented Generation (RAG) architecture to generate synthetic data whenever needed. They can take multimodal content such as product descriptions and

specifications, customer feedback and reviews and inventory notes, vectorize them and store vector embeddings alongside the operational data in MongoDB Atlas. This allows them to supercharge their inventory optimization using

RAG. They 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.

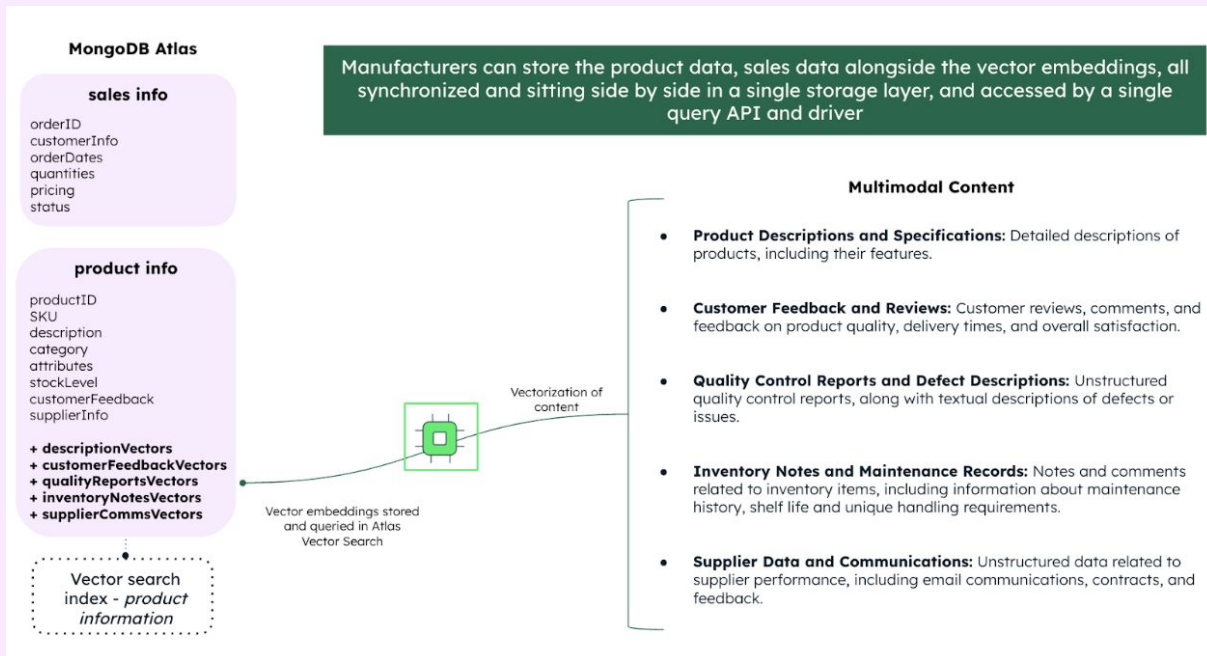


Figure 4: Enhancing Inventory Data with Vector Embeddings

The figure below shows a reference architecture of generative AI+AI enabled demand forecasting with MongoDB Atlas. The accurate demand forecasting results will help in stocking up on right inventory levels. For new products, the historical sales data is not available. Generative AI models can create realistic and diverse synthetic data by learning patterns from existing datasets of similar products. This synthetic data can mimic the sales trends and seasonality that new products might experience. Atlas Vector search can find similar products attributes and feed that context into the generative AI model. By finding semantics in similar products, Atlas vector search can help refine the synthetic data generation, ensuring that it closely reflects potential market conditions and

customer behaviors. This approach not only fills the gap of missing historical data for new products but also provides a robust foundation for demand forecasting enabling manufacturers to optimize their inventory levels.

Solution demo

Discover how to building an event-driven inventory management system.

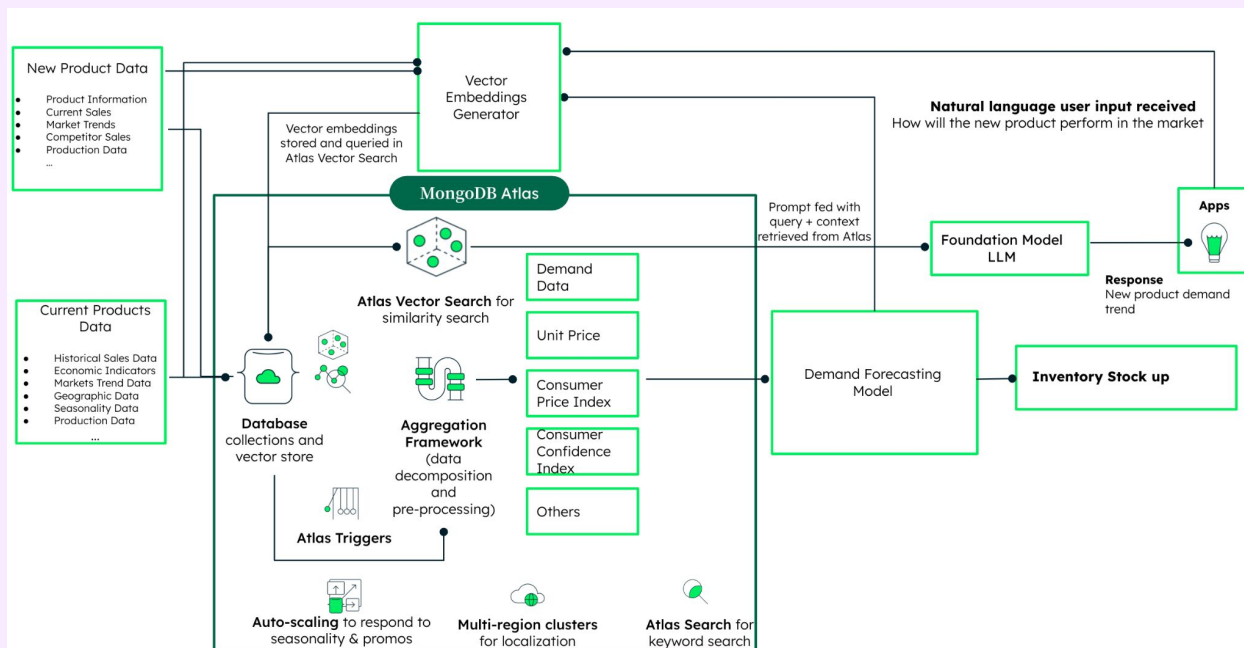


Figure 5: Gen AI enabled demand forecasting with MongoDB Atlas



Ceto is revolutionizing maritime operations with MongoDB time series

Ceto is on a mission to bring the maritime industry into the digital age—and to transform maritime operations into a model of efficiency and sustainability.

To make its mission a reality, Ceto partnered with MongoDB, leveraging its robust data handling capabilities to integrate AI with real-time data collected from thousands of sensors across its customers' fleets. This allows Ceto to predict and preempt potential failures, streamline operations, and manage risks proactively. This shift not only enhances safety and reliability but also propels maritime logistics into a new era of technological advancement, making Ceto a transformative force in global commerce.

MongoDB's architecture provided Ceto with several key features that are crucial for their operations. Scalability was essential for managing the increasing data volumes generated by their expanding fleet. [Time Series](#) Collections offered advanced data compression capabilities, crucial for managing the large volumes of data generated daily.

“MongoDB's Time Series collections have revolutionized how we manage and utilize data from our fleet. The ability to process and analyze data in real-time has significantly enhanced our predictive maintenance capabilities.”

[Learn more](#)

Ben Harrison
CTO, Ceto

Predictive Maintenance

A well-defined maintenance strategy can be a game-changer for any organization, driving significant revenue and cost savings. Here's how it works:

First, identify the equipment that is most crucial for your operations. Downtime for this equipment can lead to bottlenecks, halting production.

Second, equip these critical assets with sensors to enable condition monitoring. This allows you to monitor the health of the equipment in real time, identifying potential issues before they escalate into catastrophic failures.

Third, based on the prediction, the system can generate work orders, schedule maintenance activities, and even provide guidance to maintenance personnel. This ensures that maintenance is performed only when necessary, optimizing resource allocation.

This series of activities delivers tangible benefits. Costs are reduced through saved labor hours and extended machine lifespan. Additionally, revenue increases as your machines operate at optimal performance levels.

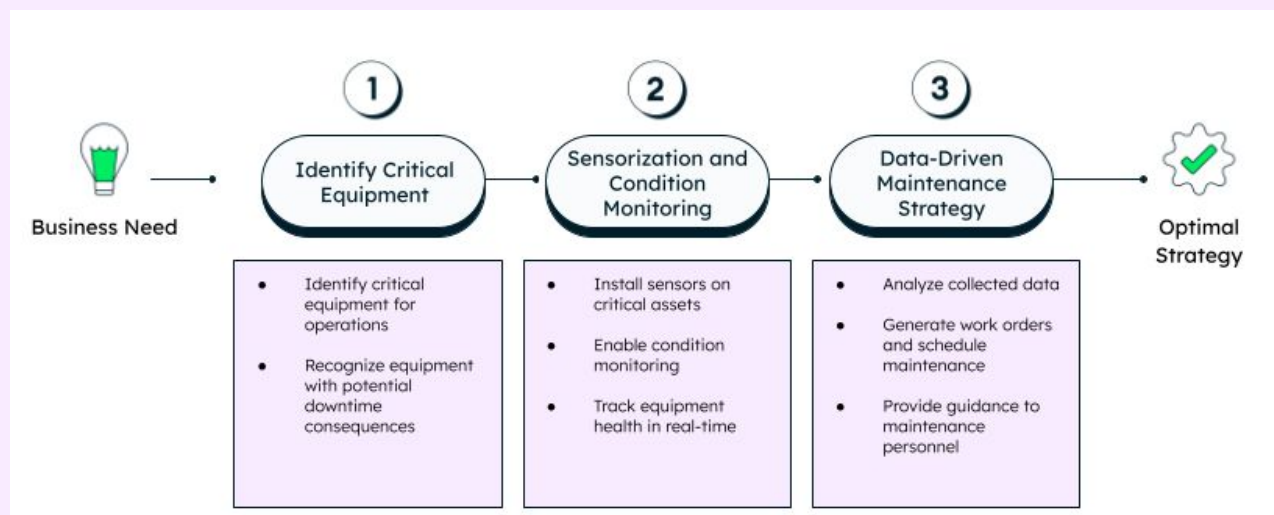


Figure 6: Steps required for an optimal maintenance strategy

Optimal Maintenance Strategy

An optimal maintenance strategy isn't a one-size-fits-all solution. It's about choosing the right blend of approaches based on your specific equipment and operational needs.

In today's processes, we see a spectrum of maintenance approaches. These methods range from highly complex and expensive at one end to simpler, more affordable options on the other:

- **Reactive maintenance:** This is the most basic approach in which maintenance is performed after a machine fails. While simple, it can lead to unexpected downtime and higher repair costs.
- **Preventive maintenance:** This is a proactive approach that involves scheduling maintenance tasks based on predetermined time intervals or usage metrics. This helps prevent breakdowns but can be inefficient as machine conditions can vary. Thresholds for these tasks may need to be adjusted due to factors like aging equipment, changes in processes, or different materials being used.
- **Condition-based maintenance (CBM):** This approach continuously monitors the health of the machine through sensors and data analysis. Maintenance is then triggered based on the actual condition of the equipment rather than a set schedule. This is more efficient than preventive maintenance as it avoids unnecessary maintenance. Threshold-based alerting systems are often used with CBM.
- **Predictive maintenance:** This is the most advanced approach, using data analytics to predict potential equipment failures before they occur,

which allows for proactive maintenance and minimizes downtime. Predictive maintenance requires significant upfront investment in sensors and data analysis tools.

Predictive maintenance uses data analysis to identify problems in machines before they fail. This allows organizations to schedule maintenance at the optimal time, maximizing machine reliability and efficiency.

Here's how predictive maintenance can benefit manufacturing operations, [according to Deloitte](#):

3-5% Reduction in new equipment costs

5-20% increase in labor productivity

15-20% reduction in facility downtime

10-30% reduction in inventory levels

5-20% reduction in carrying costs

Predictive maintenance is constantly evolving. We've moved beyond basic threshold-based monitoring to advanced techniques like machine learning (ML) models. These models can not only predict failures but also diagnose the root cause, allowing for targeted repairs.

The latest trend in predictive maintenance is automated strategy creation. This involves using AI to not only predict equipment breakdowns but also to generate repair plans, ensuring the right fixes are made at the right time.

Automated strategy creation requires substantial investment in R&D, along with deep industry knowledge, access to relevant data, and practical operational experience. The question is, can generative AI help?

Current State and Challenges

The answer is yes, generative AI can help. But there are challenges at each stage of implementation that organizations must consider. Each stage involves a key question and associated challenges, highlighting the steps and issues faced in predictive maintenance and machinery upkeep.

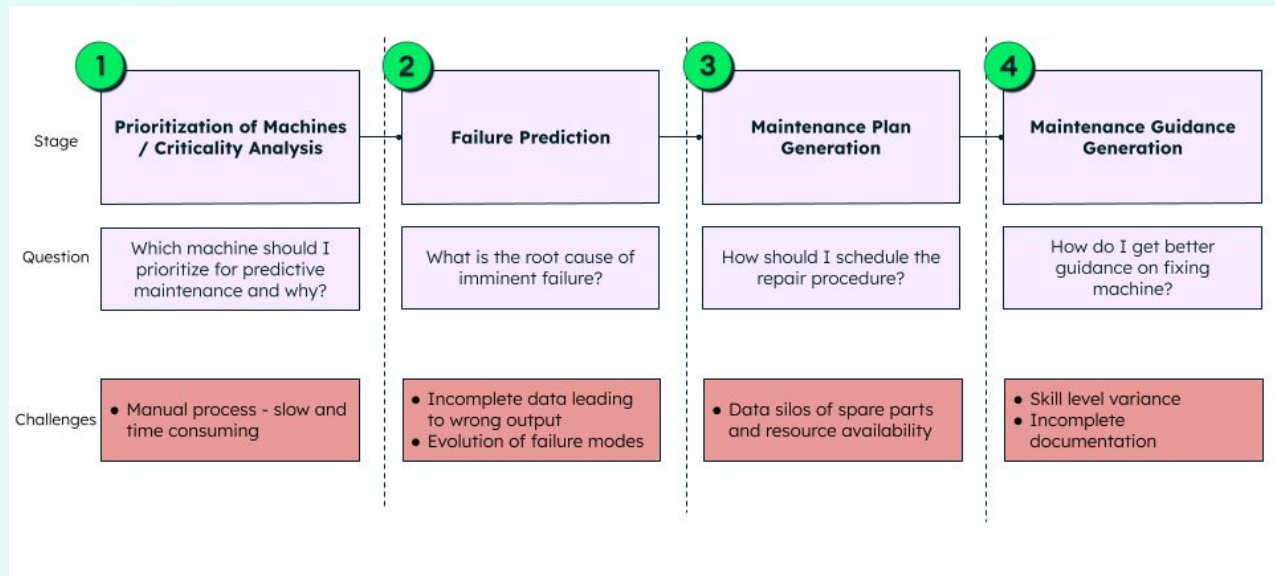


Figure 7: Different challenges seen at each stage of the predictive maintenance deployment

Now, let's envision a factory containing both automated and manual machines. Let's call it Gear Transmission Systems Ltd, whose primary output is gearboxes.

Within this factory, we have an array of equipment: cutting machines, milling machines, measurement devices, and more. As its general manager, you are tasked with managing the budget allocated for maintenance processes and improving strategies. One pressing question you must address is: which machines should take priority for the predictive maintenance projects, and why? This involves consulting with the

maintenance managers and leaders and conducting quantitative analyses, a rather manual process.

Once you've identified the machines, the next step is to install sensors and train the machine learning model. However, two major challenges arise. First, you lack sufficient "run to failure" data to effectively train the model. Secondly, machine health deteriorates over time, leading to evolving failure modes with the age of the machine.

Assuming you manage to overcome these hurdles, the next phase involves maintenance

scheduling and execution. You're faced with a myriad of data silos, including inventory data and resource availability data, which need to be integrated to formulate a comprehensive repair plan. Furthermore, it's essential to ensure that operators are adept at addressing minor machine issues to reduce reliance on external experts. While complex issues may still require OEM or SI support, internal troubleshooting capabilities are invaluable. Therefore, developing easy-to-follow documentation tailored to the skill levels of our staff is important.

As the Figure below shows, different data is required for solving above listed challenges:

- **Prioritization of machines/criticality analysis:** At this stage, we require both structured and unstructured data. We need previous machine failure data as well as expert analysis/opinion on which machines to prioritize for predictive maintenance and why.
- **Failure prediction:** This stage involves structured data such as sensor data and maintenance logs to identify the root cause of imminent failure.
- **Maintenance plan and guidance generation:** In both of these stages, we deal with both structured and unstructured data. The objective is to combine this data to generate an optimal repair plan and operator guidance.

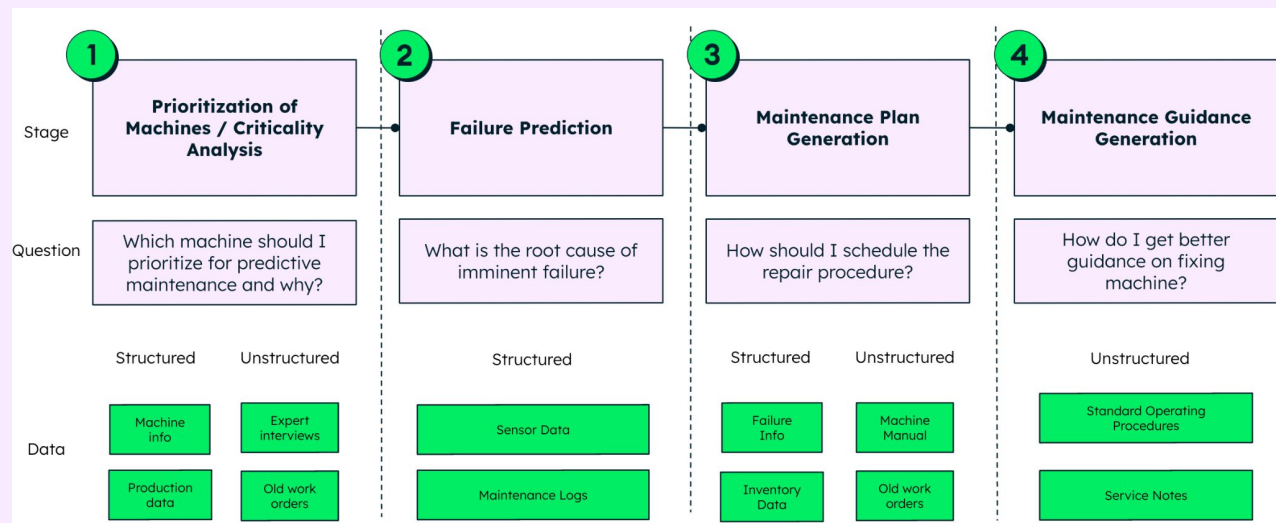


Figure 8: Different data requirements at each stage

How AI and MongoDB Help

MongoDB Atlas is the only multi-cloud developer data platform designed to accelerate and simplify how developers work with data. Using MongoDB Atlas, developers can power end-to-end value chain optimization with AI/ML, advanced analytics, and real-time data processing for innovative mobile, edge, and IoT applications.

Stage 1: Machine prioritization

Current machine prioritization for predictive maintenance relies heavily on manual analysis. Factory personnel gather historical and current machine data on utilization losses due to breakdowns. This data is then reviewed alongside the experience of maintenance managers and leaders. Based on this combined analysis, a roadmap for the predictive maintenance project is recommended, highlighting which machines should be prioritized.

However, this approach has limitations. A reliance on manual analysis can be time-consuming and may not always capture the full picture of the maintenance project due to the limited use of quantitative data sources. Additionally, inconsistencies in interpretation can lead to an overdependence on institutional knowledge, which in turn can result in false analyses that impact the project's return on investment (ROI).

But with the arrival of generative AI, things have changed. A generative AI-based machine prioritization tool can be created to reduce the time manufacturing experts spend on manual analysis, and to decrease the risk of poor investments. To leverage AI, experts need a data store capable of storing and operationalizing both structured and unstructured data. Having such a data store will allow them to perform semantic searches

and to provide the right context to the large language model, ensuring it generates responses based on factory data without hallucinating. Such a system can result in positive business outcomes. Here's how the system can look with MongoDB Atlas as the AI data store:

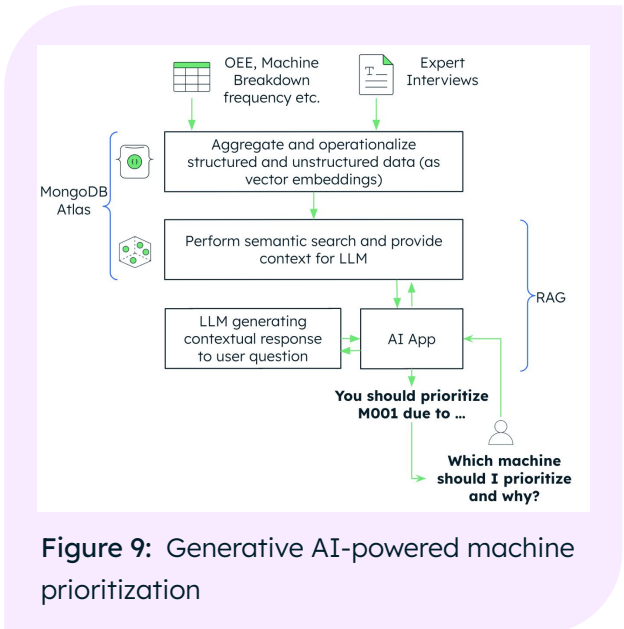


Figure 9: Generative AI-powered machine prioritization

This retrieval-augmented generation (RAG) system consists of four parts. First, an AI data store aggregates and operationalizes structured and unstructured data. In the Figure above, machine breakdown history and operational parameters are represented as structured data, while expert interviews are stored as unstructured PDF files. The PDF files are vectorized and stored in MongoDB Atlas. Atlas Vector Search is then utilized to perform semantic searches and to find meaningful context from the PDF embeddings.

[Atlas Vector Search](#) can be triggered using an AI application, connecting to MongoDB Atlas to retrieve the right context, which is then fed into the large language model to answer questions like "Which machine should I prioritize and why?" The response might suggest prioritizing Machine M001 or M002 due to certain reasons, including but not limited to the criticality of the machine, high maintenance cost, etc.

Stage 2: Failure prediction

Now that we've discussed prioritizing equipment, let's move on to failure prediction. MongoDB Atlas provides all the necessary building blocks or tools to implement failure prediction. By providing a unified view of

operational data, real-time processing capabilities, integrated monitoring and alerting, and seamless compatibility with machine learning tools, MongoDB Atlas enables organizations to optimize machine performance and minimize downtime.

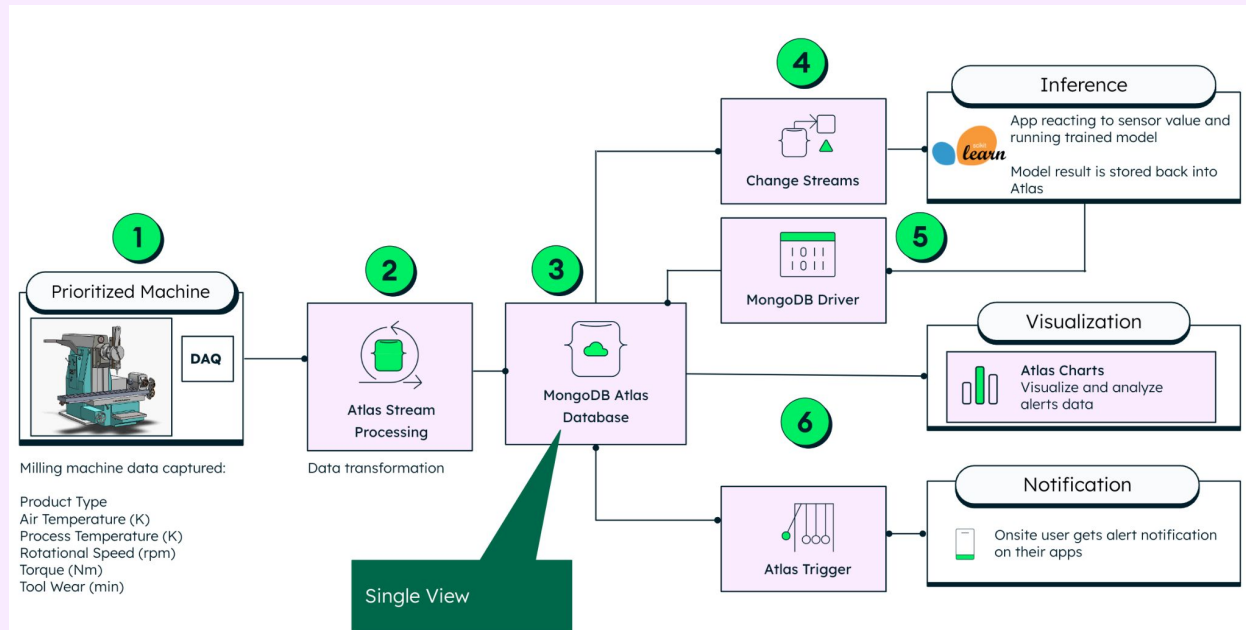


Figure 10: Smart milling machine uses real-time data to predict failures

As seen in the Figure above, we have our prioritized machine, which is a milling machine with attached sensors that collect data such as air temperature, rotational speed, torque, and tool wear. This data will be processed through Atlas Stream Processing, enabling the processing of streams of complex data using the same data model and Query API used in Atlas databases. [Atlas Stream Processing](#) enables developers to build aggregation pipelines to continuously operate on streaming data without the delays inherent to batch processing. Results can be continuously published to MongoDB Atlas or to a Kafka

topic. This allows data transformation and enrichment before it even lands in the database.

Once the data is in MongoDB, another application can react to sensor values and run a trained model designed to predict failures. The model results can be stored back into Atlas (between steps 4 and 5/Inference in the Figure above). These results can then be visualized using [Atlas Charts](#). Finally, [Atlas Triggers](#) and Functions can be used to push notifications to on-site users. This establishes an end-to-end system for failure prediction.

Stage 3: Repair plan generators

Having identified the nature of the equipment failures, the implementation of a comprehensive repair strategy becomes paramount. First, we have to generate a maintenance work order. This order should include repair instructions, spare parts needed,

schedule, and resource availability information. In this case, both structured and unstructured data are involved. The repair instructions will come from the machine manual. For this process, MongoDB Atlas acts as the operational data layer, seamlessly integrating structured and unstructured data.

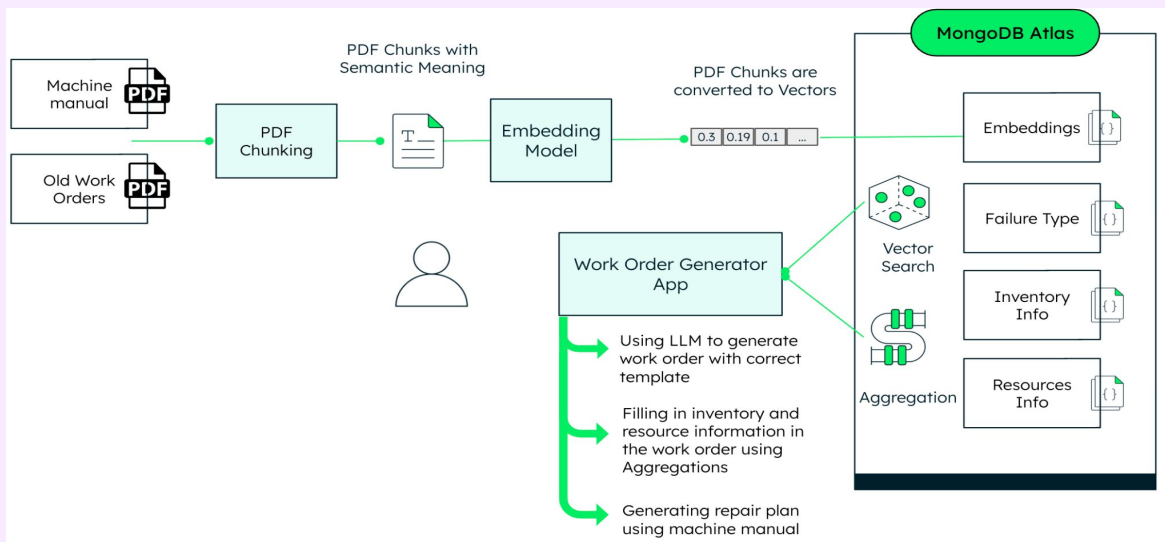


Figure 11: MongoDB Atlas as the operational data layer for structured and unstructured data

The Figure above shows the process of work order generation using generative AI. First, we must extract chunks of information from a milling machine's operating and repair manual, as well as from old work orders stored as PDF files, and convert them into vectors. These embeddings are then stored in MongoDB Atlas. MongoDB's versatility allows for the storage of both structured and unstructured data within the same database. Leveraging Atlas Vector Search and Aggregation pipelines, we can integrate this data to feed into a large language model (LLM) powering a work order generator application. The LLM analyzes the data to generate the appropriate work order and template, drawing from past examples. It populates inventory and resource details using aggregation techniques and structured data. Finally, it generates a repair plan similar to the

old work orders. What sets this approach apart is the ability to use the same MongoDB database to store structured data such as failure types, spare parts inventory, and resource information. By employing the aggregation framework to extract relevant information from structured data and vector search to glean insights from vectors, the LLM within the work order generator application gains contextual understanding.

This application seamlessly utilizes the LLM to generate work orders with the correct template, filling in inventory and resource details through aggregations, and ultimately creating repair plans based on machine manuals. This application can run inside a central maintenance management system.

Stage 4: Maintenance guidance generation

So we come to the last step: How can we use gen AI to enhance the operator or technician guidance to maintain the machine?

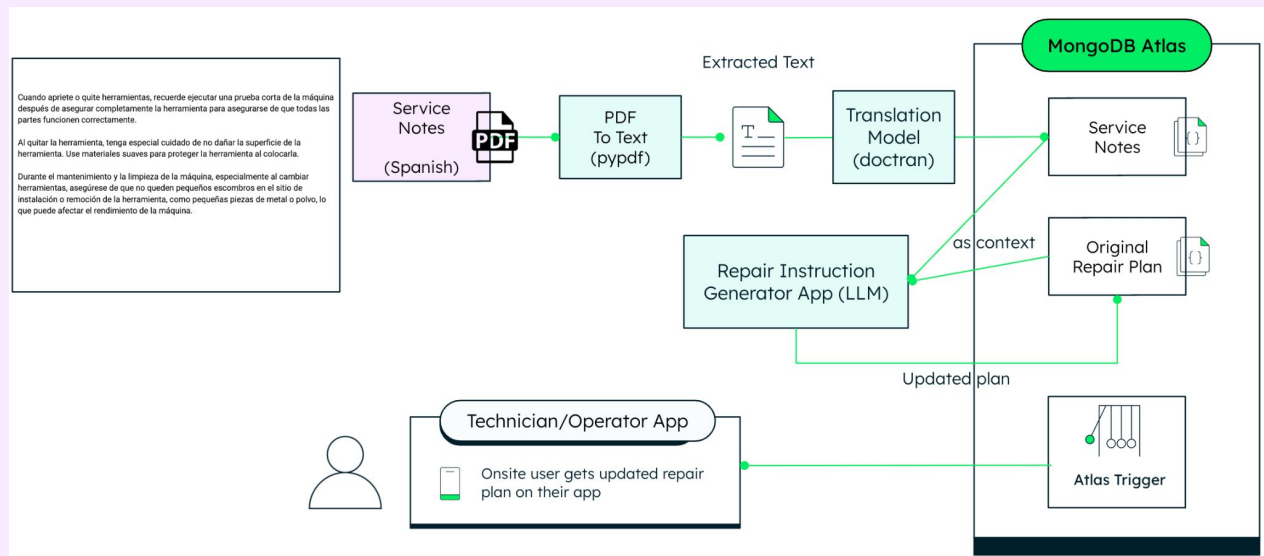


Figure 12: Using the RAG Approach for Operator Work Instructions

Let's walk through an example scenario here. The repair plan was generated in the last step. Now, the computerized maintenance management system (CMMS) has found some service notes uploaded to the platform by another technician, but they're written in another language, let's say Spanish. We can use the RAG architecture again to intelligently merge these service notes with the repair instructions generated in the previous step.

We first need to extract text from the PDF, translate it into English since our other data is in English, and then provide the service notes as well as the repair plan to the LLM as context. So, we have the original plan from the previous steps, and we combine it integrally using the LLM with the service notes obtained in this step. Note that we're not performing vector search here. Once the plan is updated,

then we can publish notifications down to the technician's application via Atlas Triggers and Functions.

In summary, we are essentially integrating AI and gen AI apps to implement an end-to-end predictive maintenance strategy (shown in the Figure on the next page).

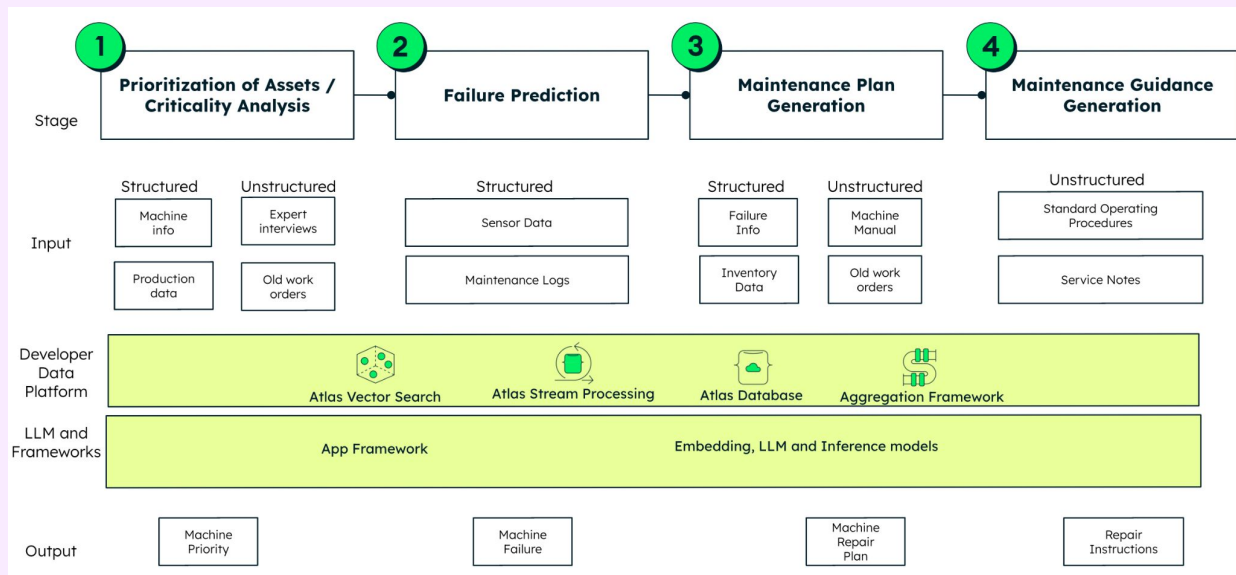


Figure 13: Model chaining with unified data store

Our input consisted of a combination of structured and unstructured data. We leveraged the various services offered by the MongoDB Atlas developer data platform, including Atlas Vector Search, Atlas Stream Processing, and, of course, the MongoDB database and aggregation framework. These features enabled us to provide the right context to the LLM and the appropriate data to the AI model.

Ultimately, we obtain the desired output at each stage, ranging from machine prioritization, failure type identification, and repair plan formulation, to instruction generation.

Solution Demo

[Find out](#) how AI is being used in renewable energy by leveraging MongoDB Atlas Vector Search to drive efficiency through real-time, audio diagnostics.

Knowledge Management

The Unique Challenges of Preserving Knowledge

Preserving and maintaining knowledge in manufacturing is just as much a challenge as its accessibility. Due to aging population worldwide, as experienced workers retire, valuable tribal knowledge is lost. Transferring their expertise to the in-experienced workforce is difficult. Siloed data resulting from mergers and acquisitions or legacy systems makes it even harder to consolidate knowledge for decision making.

How Generative AI and Atlas Vector Search Help

Manufacturers can capture and index the valuable knowledge left by experienced

workers, including both textual and unstructured information. Creating semantic vectors from these documents, manuals, and notes simplifies this process. The task of locating and transferring knowledge from data silos gets eliminated by indexing and generating vectors from a wide range of data sources, encompassing both structured documents and unstructured data such as handwritten notes. This enables users to perform cross-system searches using natural language queries facilitating seamless access to information across different platforms.

Preserving Expertise through MongoDB Atlas

Documents and Manuals: Operation manuals, maintenance guides, process documents

Email Archives: Archived email communications related to specific projects, issues, or processes

Project Reports: Reports from completed projects, including success stories, lessons learned, and best practices

Standard Operating Procedures (SOPs): Any existing SOPs that experienced workers have followed and contributed to

Incident Reports and Case Studies: Records of past incidents, near-misses, and case studies

Meeting Notes and Minutes: Notes from meetings, particularly those involving experienced workers

Interviews and Personal Conversations: Transcripts or recordings of interviews with retiring workers, capturing their insights and experiences

Photos and Videos: Visual documentation of machinery, equipment, and processes

External Resources: Information from external sources, such as industry best practices

Technical Specifications and Blueprints: Detailed technical specifications, blueprints, and engineering drawings

Supplier and Vendor Documentation: Information provided by suppliers and vendors, including manuals and documentation

Regulatory and Compliance Documents: Documents related to industry regulations, safety guidelines, and compliance standards.

Customer Feedback and Quality Reports: Feedback from customers, quality assurance reports, and data on product defects or improvements

Employee Training Materials: Training materials used for onboarding and skill development

Embedding
Creation

MongoDB Atlas



Vectors, Core &
Metadata Store

Example LLM Prompts

- How to calibrate the de-palletizing robot?
- How to fix the shot peening machine?
- Give me summary of technical manual of my collaborative robot

Figure 14: Preserving knowledge in MongoDB Atlas

Knowledge collection from shop floor

The time being wasted for consolidating data from different systems to take decisions on a daily basis can be heavily reduced by using gen AI. A shift leader, as an example, spends a lot of time to collect data from different sources like MES, SCADA, or from handwritten notes of the night shift workers to get an overview about the condition of the equipment after the last shift. Traditionally it takes a lot of time to collect all that data from various sources and locations for getting a holistic overview to understand current safety, maintenance, inventory and quality needs.

A shifter leader may struggle to make timely decisions due to the fragmented nature of the data, leading to delays in addressing production issues and optimizing workflows.

With a knowledge management application on the shopfloor all the input from the production equipment as well as from the workers can be collected. The application takes all the structured and unstructured text input and categorizes it into one of many categories (defects, breakdowns, alarms, etc.).

Having all that data collected, contextualized, and indexed allows a chatbot application to get an immediate overview on the status of the shopfloor by prompting: Provide me a list of machines with problems in the last shift, followed by prompting on how to solve that problem. The RAG application can therefore use all the preserved information from the experienced workers, stored as vectors in MongoDB Atlas.

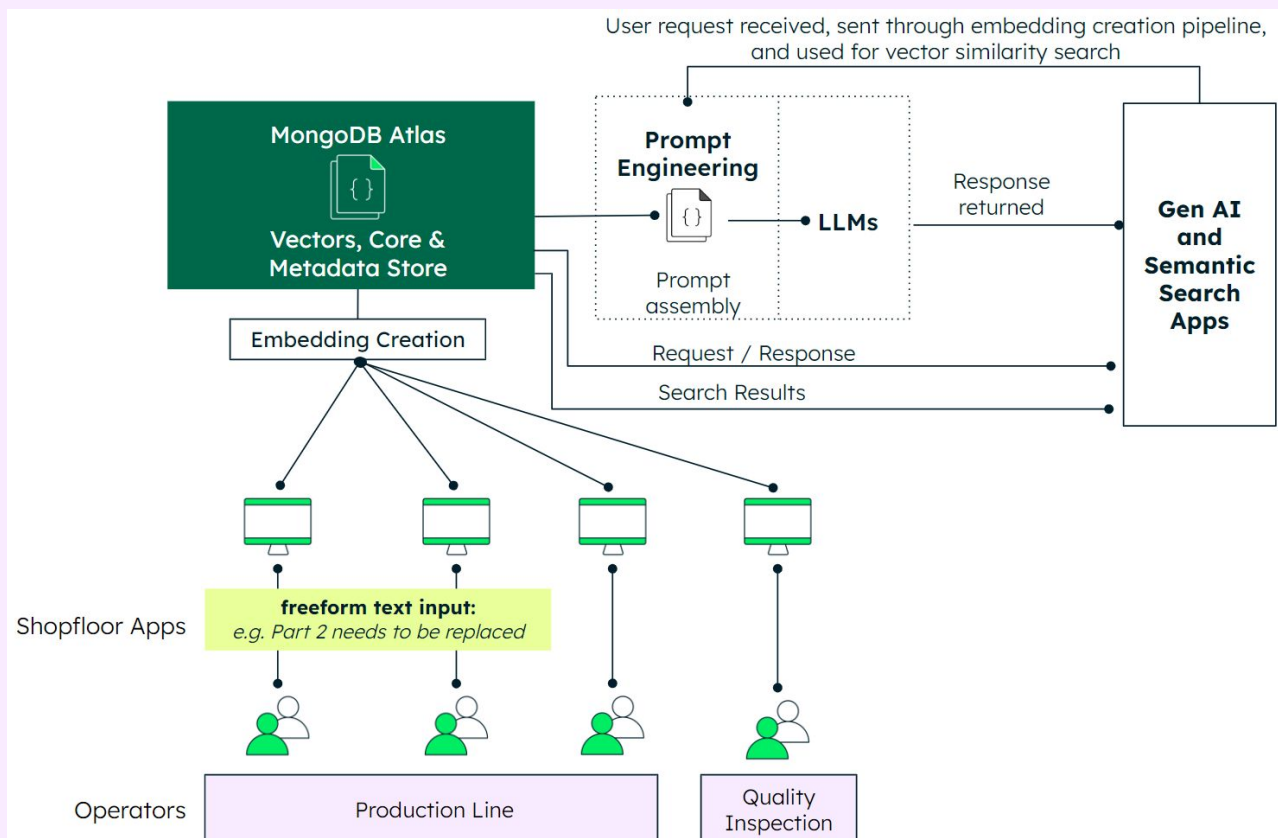


Figure 15: Knowledge management architecture



Eni makes terabytes of subsurface unstructured data actionable with MongoDB Atlas

Based in Italy, **Eni** is a leading integrated energy company with more than 30,000 employees across 69 countries. Its operations vary from exploring and drilling for natural gas and oil to cogenerating electricity, renewables, biorefining, and chemical production.

Eni partnered with [MongoDB Consulting](#) for training and to support the migration of workloads into [MongoDB Atlas](#). Eni wanted to move to a managed service with a seamless user experience and easy-to-use interface for developers.



Improving staff productivity at Enel using Amazon Bedrock

Enel is a leading integrated electric utility with a presence across 32 countries and an 82-GW generation capacity

Enel identified the opportunity to use generative AI to boost IT service desk efficiency by extending automation to nontrivial tasks through basic troubleshooting, providing resolution steps and ticket routing without human involvement.

The solution is designed around a retrieval-augmented generation architecture using Amazon Bedrock.

With MongoDB Atlas, Eni users can quickly find data spanning multiple years and geographies to identify trends and analyze models. [MongoDB Atlas Search](#) also assists by filtering out irrelevant documents. The team also integrated AI and machine learning models with the platform to make it even easier to identify patterns.

“MongoDB Atlas isn’t just a database, it’s a complete set of products and services. It’s cloud agnostic and combines rich functionality with the flexibility we needed to make it our own.”

[Learn more](#)

Sabato Severino

*Senior AI Solution Architect for
Geoscience at Eni*

The solution uses Amazon Titan, a family of models exclusive to Amazon Bedrock. Specifically, it uses the Amazon Titan Text Embeddings model to generate embeddings (vectors capturing semantics of text) from Enel’s knowledge base, which consists of a series of runbooks containing incidents classes, preconditions, root causes, resolutions steps, and operations information related to the applications. Embeddings are computed and persisted in a vector database instance using [MongoDB Atlas Vector Search](#), which supports similarity search.

[Learn more](#)

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.

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

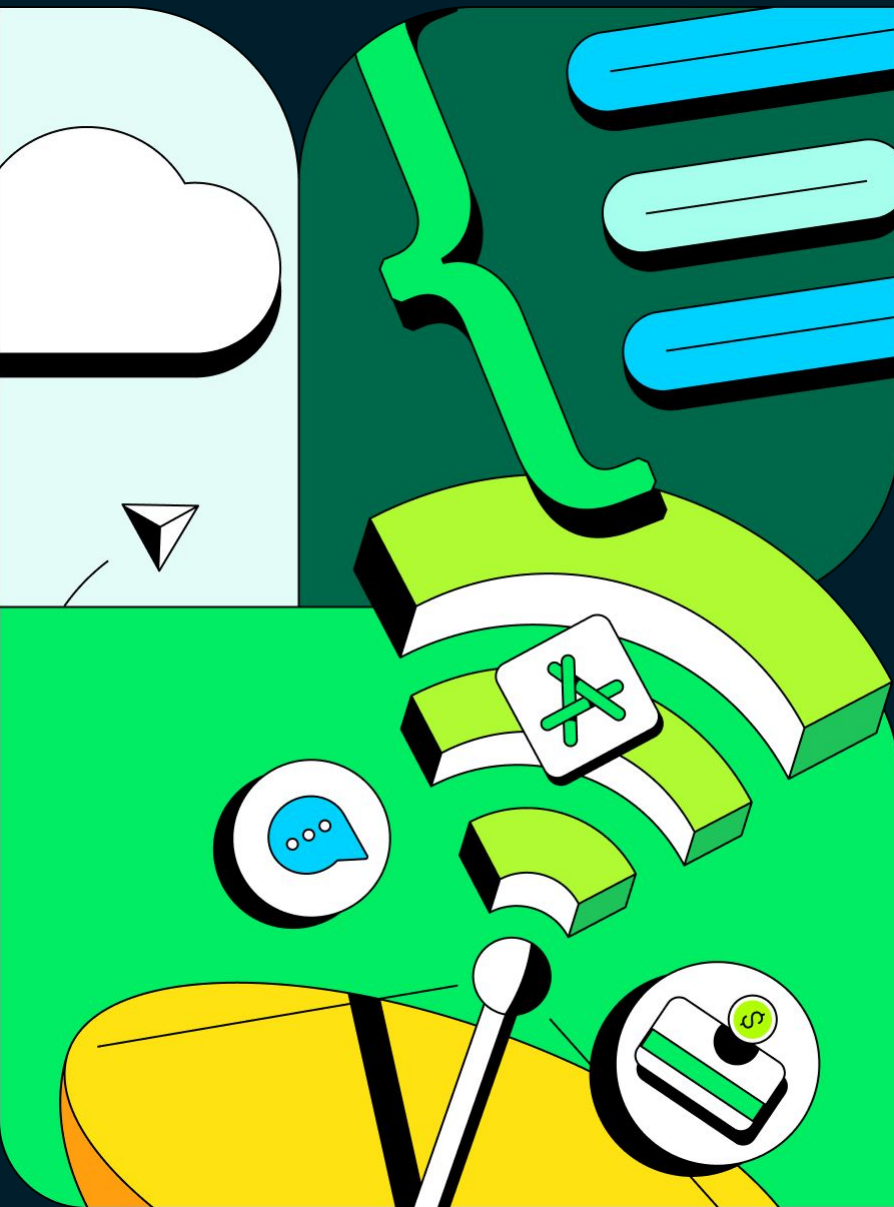
Manufacturing & Motion
Industry Solutions Principal
raphael.schor@mongodb.com



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.



AI investments in the telco & media industry



Generative AI is set to revolutionize the media and telecommunications sectors, significantly impacting content creation, distribution, service assurance, anti-fraud measures, and overall industry dynamics.

Content Creation and Distribution

AI will dramatically enhance both content creation and distribution. Search generative experiences (SGE) and AI-driven chatbots are expected to provide faster, more intuitive access to information, shifting audience flows and financial dynamics within the industry. As referral traffic from traditional social media platforms like Facebook and X (formerly Twitter) declines, media companies face the challenge of maintaining their audience and revenue streams.

In response to these challenges, media organizations are focusing on building direct relationships with their audiences and investing in intellectual property protection. This strategy aims to reduce dependence on major tech platforms but risks alienating younger, tech-savvy audiences who are already comfortable with algorithmically generated news.

Opportunities and Strategies

Despite the challenges, the shift towards AI offers numerous opportunities. News organizations are exploring innovative ways to integrate AI, emphasizing the importance of balancing technological benefits with potential risks.

These are some of the key findings from an [industry survey](#) of over 300 digital leaders across 50 countries:

- **Referral Traffic:** 63% are worried about declining traffic from social media, with significant drops reported from platforms like Facebook and X/Twitter. Traffic to news sites from Facebook fell 48% in 2023, with traffic from X/Twitter declining by 27%.
- **Platform Strategies:** Publishers are prioritizing platforms like WhatsApp and Instagram while maintaining interest in video networks such as TikTok and YouTube.
- **Content Focus:** There is a strong emphasis on increasing video, newsletters, and podcasts to engage audiences, while concerns about news fatigue and selective avoidance persist.

Revenue and AI Integration

Subscription and membership models are seen as vital revenue streams, with publishers also exploring lucrative AI platform licensing deals. AI is primarily viewed as a tool for backend automation, improved recommendations, and commercial uses.

Today's leading companies in the telecommunications and media industry are leveraging AI to enhance operations, safeguard their businesses, deliver relevant content and services to customers, and drive innovation.

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. 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 has been revolutionized by AI through several key capabilities: ML can be the powerful foundation for predictive maintenance, analyzing patterns and predicting network failures, 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. AI can also have a huge impact on 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. It stores 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 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 considered, with reaction time being a critical factor in preventing the worst things from happening, 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 identified, for instance, suspicious user activities. In addition to Atlas Vector Search, the MongoDB Query API supports stream processing, simplifying data ingestion from various sources.

One of the customers utilizing MongoDB for its anti-fraud strategy is [AT&T](#). The company selected MongoDB for its ability to ingest and store rapidly changing data. Moreover, MongoDB Atlas meets AT&T's key requirements for performance, availability, and security. The fraud prevention platform now operates over 50 different AI models, utilizes more than 1,200 features, stores 30TB of data, and processes up to 20 million events daily. Events are processed in under 200 milliseconds, enabling real-time fraud detection and prevention. As a result, fraud—particularly iPhone-related fraud, which was previously a \$1 billion problem—has been reduced by over 80%.

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 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.

Content Suggestions and Personalization

By utilizing user data, behavior analytics, and the multi-dimensional vectorization of media items, the platform suggests content that aligns with individual preferences and past interactions. This not only enhances user engagement but also increases the likelihood of converting free users into paying subscribers. Vectors are embedded directly in MongoDB documents. This has several advantages. For instance, there are no complexities of a polyglot persistence architecture, and no need to extract, transform, and load (ETL) data between different database systems, which simplifies the data architecture and reduces overhead. Furthermore, MongoDB's built-in scalability and resilience become particularly advantageous when handling vector search operations. Organizations can scale their operations vertically or horizontally, and they can even choose to scale search nodes independently from operational database nodes, thus adapting to the specific load scenario.

Content Summarization and Reformatting

In an age of information overload, the solution provides concise summaries and adapts content formats based on user preferences and device specifications. This tailored approach addresses the diverse consumption habits of users across different platforms.

Keyword Extraction

Essential information is drawn from content through advanced keyword extraction, enabling users to grasp key news dimensions quickly and enhancing the searchability of content within the platform. Keywords are fundamental to how content is indexed and found in search engines, and they significantly influence the SEO (search engine optimization) performance of digital content. With the help of the underlying [LLM](#), the solution extracts keywords automatically and with high sophistication.

Automatic Creation of Insights and Summaries

Our system can automatically generate comprehensive insights and summaries from multiple articles. This feature is particularly valuable for users interested in deep dives into specific topics or events, providing them with a rich, contextual experience. This capability leverages the power of one or more large language models (LLMs) to generate natural language output, enhancing the richness and accessibility of information derived from across multiple source articles.

Solution demo

[Discover](#) how to leverage MongoDB and Vector Search effectively to create more engaging, personalized user experiences.

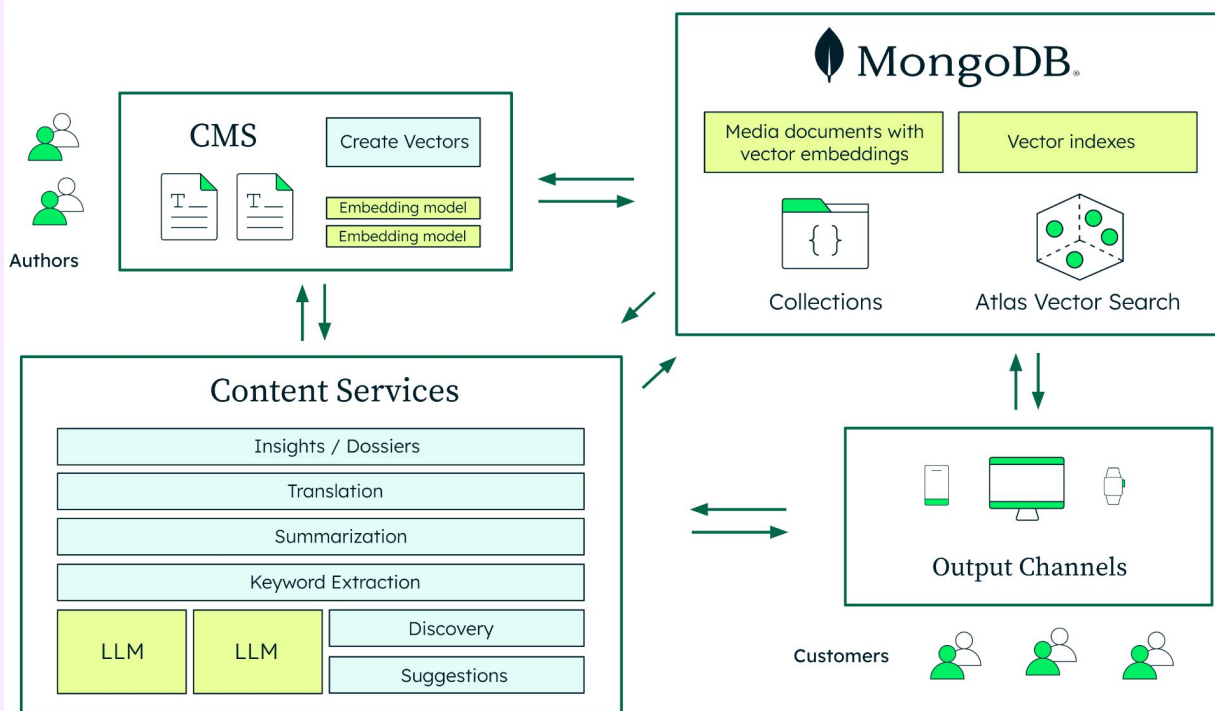


Figure 16: Reference architecture for to creating more engaging, personalized user experiences with MongoDB and Vector Search

Search Generative Experiences (SGE)

Search generative experiences (SGE) represent a transformative shift in how media and content industries interact with information retrieval and user engagement. SGE utilizes artificial intelligence to generate search results that are not just links to existing content, but comprehensive, synthesized responses tailored to the user's query. By providing immediate, concise answers, SGE helps media platforms keep users engaged longer, reducing bounce rates and increasing the likelihood of deeper interaction with the content.

With MongoDB and Vector Search being placed at the center of an SGE solution (as shown below in the reference architecture), multiple new revenue streams can be established.

Intelligent Research Tools

Provide easy access for researchers to the publisher's news archive, returning not just

the most relevant articles for a topic of interest, but also combining them into a concise summary, elevating the value for the customer.

Smart Conversational Interfaces

Content retrieval products that are accessible from smartwatches, AR and VR glasses, lapel pins, and other wearable devices, or from the car, allow for the consumption of data in more day-to-day situations.

Gamified Learning Experiences

Develop educational content with interactive quizzes and activities that adapt to the user's knowledge level and learning pace using Atlas Vector Search and SGE to provide personalized feedback.

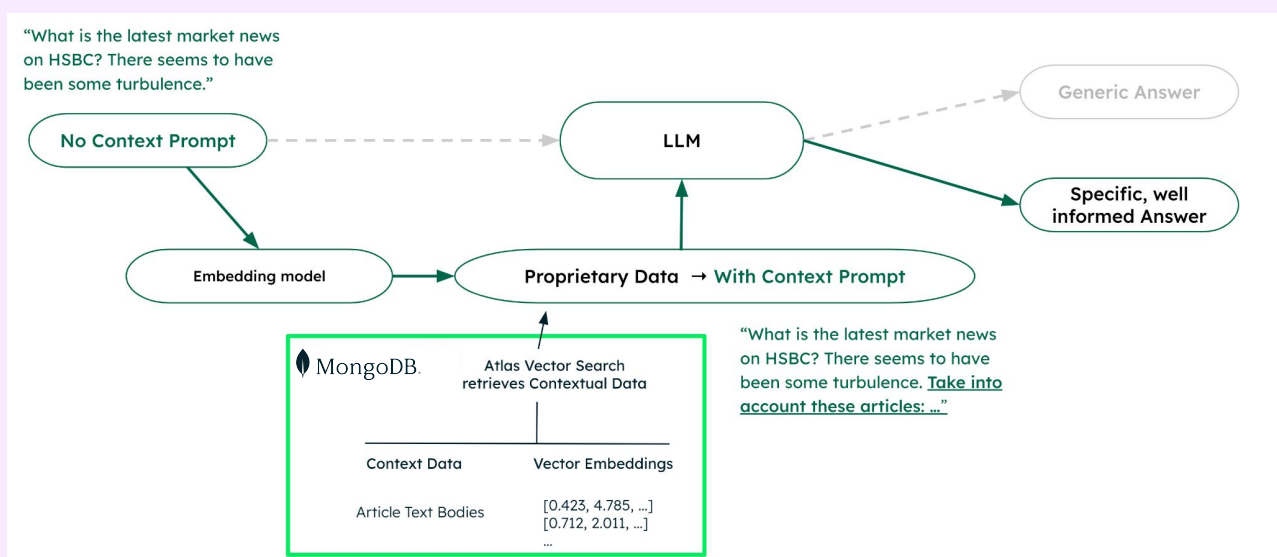


Figure 17: Reference architecture for creating Search Generative Experiences

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 an ML algorithm.

This method facilitates the adoption of dynamic pricing models instead of sticking to a standard price list, thereby increasing overall revenue, and enlarging the paying customer base.

Backend Automation

Extract relevant keywords from news articles using one or several LLMs, thus supporting search engine optimization (SEO), as well as content indexing and categorization.

Assist in copyediting, by identifying and correcting grammatical errors, spelling mistakes, and punctuation issues. Suggest improvements in sentence structure, style, and tone, ensuring high-quality content production.



Vodafone's New Developer Speed and Dexterity—Powered by MongoDB

Vodafone transformed its software development with a "telco-as-a-service" (TaaS) model and MongoDB Atlas, enabling faster, more secure app development. This shift to serverless computing cut costs, improved customer experience, and boosted developer productivity, allowing teams to quickly deploy code and innovate.

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

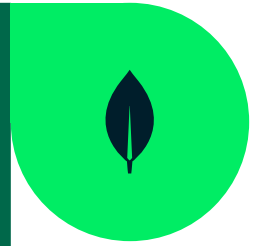
Retail



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.

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

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



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

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

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

- **Content Creation:** Generative AI can produce personalized marketing content, such as product descriptions, advertisements, and social media posts, tailored to specific audiences.
- **Hyper Personalization:** Beyond recommendations, generative AI can

create personalized shopping experiences by dynamically generating web and mobile interface elements based on user behavior.

- **Conversational Chat:** AI-generated virtual shopping assistants can provide more natural and engaging interactions with customers, improving the overall shopping experience.

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

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

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

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

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

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

The most creative AI use cases for retailers is to understand customer needs and choices that change continually with season, trends, and socio economic shifts. By analyzing customer data and behavior, generative AI can also create personalized product recommendations, customized marketing materials, and unique shopping experiences that are tailored to individual preferences.

AI plays a critical role in decision making at retailer enterprises—product decisions such as design, pricing, demand forecasting, and distribution strategies require complex understanding of a vast amount of information from across the organization.

To ensure that the right products in the right quantities are in the right place at the right time, back office teams leveraged machine learning arithmetic algorithms for years.

As technology has advanced and the barrier for entry is lowered for adopting AI, retailers are moving towards data-driven decision making where AI is leveraged in real time. Generative AI is used to consolidate information and provide dramatic insights that could be immediately utilized across the enterprise.

AI-Augmented Search and Vector Search

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

One of the key customer interactions in an e-commerce experience is search. Through the implementation of full-text search engines, customers can more easily find items that match their search, and retailers are given the opportunity to rank those results in a way that will give the customer the best option. Traditionally, decisions on how to rank search results in a personalized way were made by segmentation of customers through data acquisition from various operational systems, moving it all into a data warehouse, and subsequently running classical AI with various machine learning algorithms on such data. Typically, this would run in a batch mode (every 24, 48, or even 72 hours or a few days), and the next time a customer logs in, they will have a personalized experience. It does not, however, capture the customer's true desire in real time.

Modern retailers augment search ranking with data from real-time responses and/or

analytics from AI algorithms. Also, it's now possible to incorporate factors such as the current shopping cart or basket and customer clickstream and trending purchases across shoppers.

The first step in truly understanding the customer is to build a customer operational data store that combines data from disparate systems and silos in the organization: support, e-commerce transactions, in-store interactions, wish lists, reviews, and more. MongoDB's flexible document model enables bringing data of different types and formats in one document to get a clear view of the customer in one place. As the retailer captures more data points about the customer, they can easily add fields without the need for downtime due to schema changes.

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 workloads (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. Retailers can build dynamic ranking by using the MongoDB aggregation framework for advanced analytical queries or triggering an AI model in real time to give an answer that can be embedded into the search ranking.

The benefit of an all-in-one platform is huge here, as instead of having to update your search indexing to incorporate your AI augmentation, MongoDB has Lucene Search built in. This whole flow can be completed in one data platform automatically—as your data is being augmented with AI results, the search indexing will sync to match.

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

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

READ MORE

AI-Enhanced Search in E-commerce With MongoDB

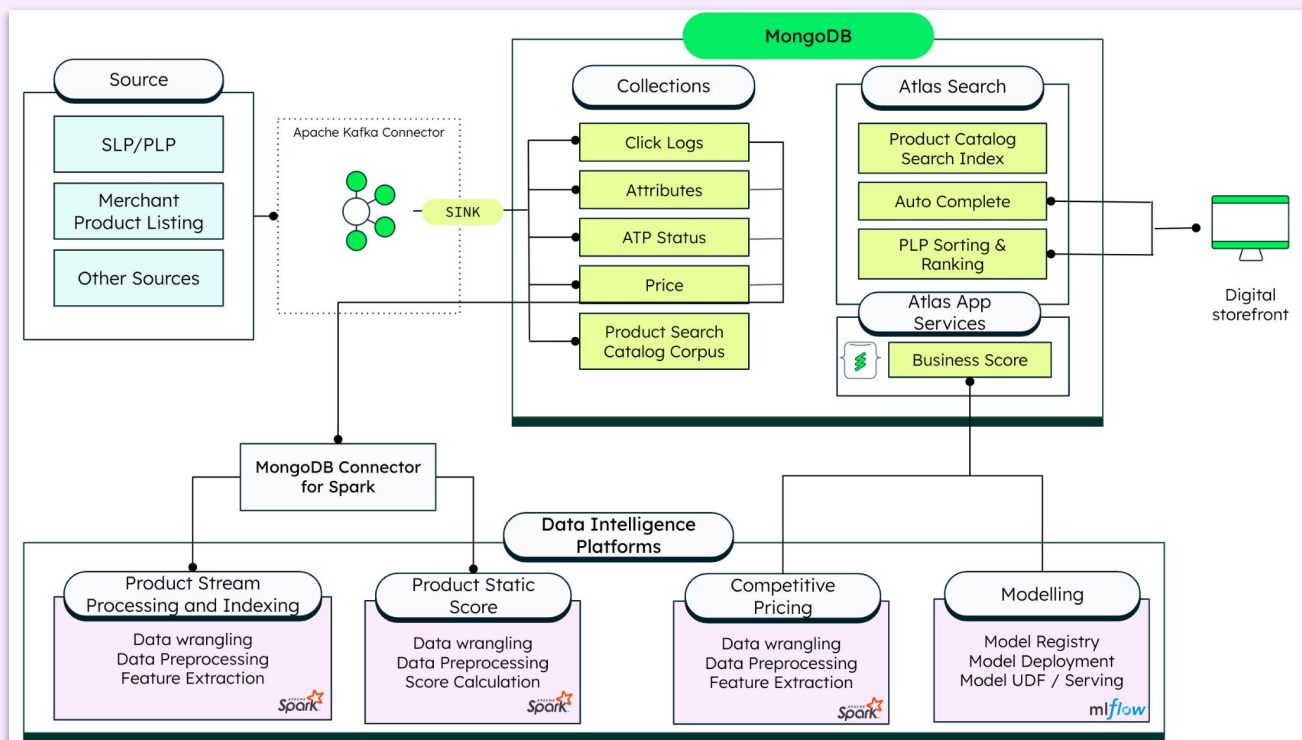


Figure 18: Architecture of an AI-enhanced search engine explaining the different MongoDB Atlas components and Data Intelligent Platforms and workflows used for data cleaning and preparation, product scoring, dynamic pricing, and vector search



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

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

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

The solution was MongoDB Atlas, a scalable, high-performing developer data platform that integrates easily with the AI ecosystem. With it the new Item Replacement Tool is being piloted first in the Middle East. By providing personalized recommendations against live inventory, Delivery Hero expects to see an increase in its monthly gross merchandise value.

“We can implement things that would take six months on other platforms in just a few days.”

[Learn more](#)

Andrii Hrachov

Principal Software Engineer at Delivery Hero

Personalized Marketing & Content Generation

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

Customer buying patterns, constantly updating product catalog, and inventory availability are critical components of retail operations. Along with this there is also the task of ensuring that the product literature is in the right tone of voice to reflect the brand in multiple languages. The product images need to be relevant to the audience in the locale. Traditionally, this required a huge amount of labor in copywriting and editing, photography of different models, and generation of visuals and graphics.

The retailer must also understand in real time what the impact of campaigns is so they can quickly redirect their marketing spend and strategy to reflect what is working. In an industry where marketing and branding are critical business activities, retailers need as much insight on their customers as possible so they can reach them with the right message at the right time.

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

Gen AI has also rapidly increased retailers' ability to personalize the interactions with their customers. Retrieval-augmented generation using Large Language Models (LLMs) is capable of creating individualized marketing material, newsletters, social posts, and email outreach that is unique to each customer in seconds. Visuals, graphics, and even photo-realistic images can be generated using AI to leverage the vast amount of data the retailer already has. This reduces manual work and accelerates time to market.

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

The key to creating content that is personalized to the customer and the brand is leveraging the vast amount of data that retailers have in-house to provide an LLM with context.

In MongoDB, the Apache Spark Connector allows for model training of LLMs so that prompts such as “create a personalized newsletter for each customer suggesting an item based on what is on offer and their previous purchases” can use data, images, and 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. Training the model with your internal data provides an invaluable resource to retailers in reaching their audience easily

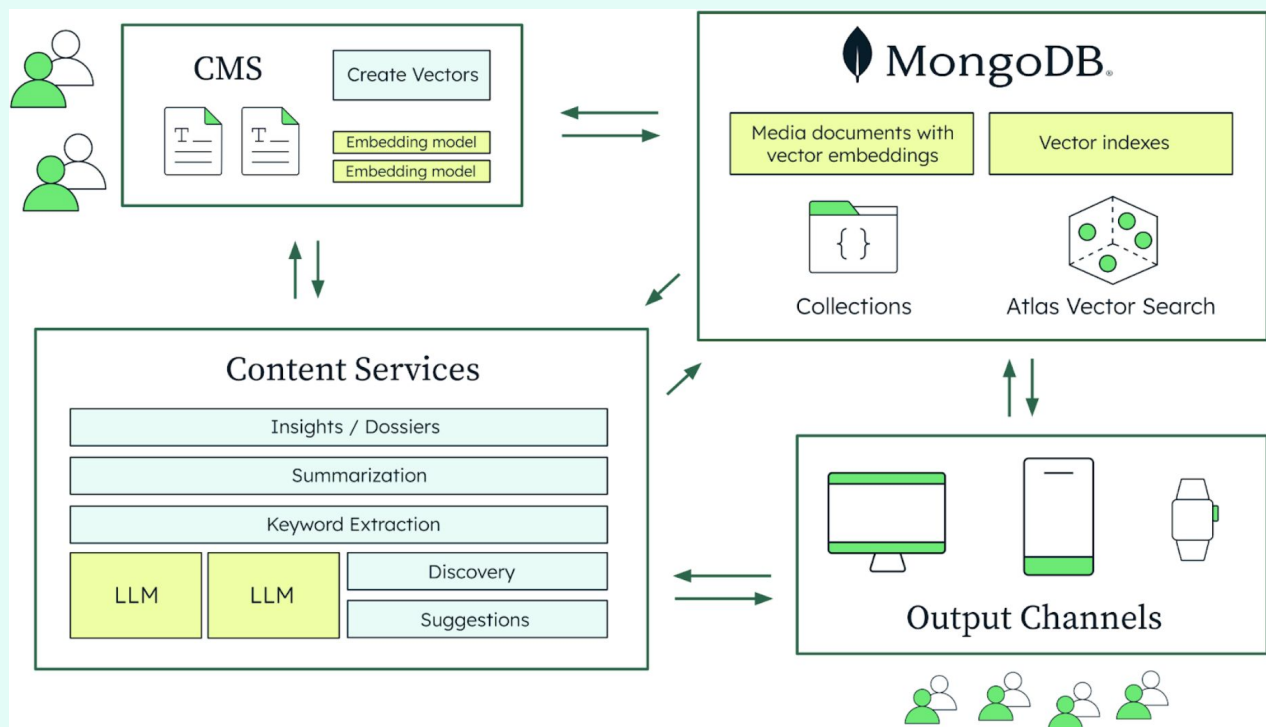


Figure 19: AI-powered personalization architecture

The above shows a reference architecture highlighting where MongoDB can be leveraged to achieve AI-powered personalization. By leveraging user data and the multi-dimensional vectorization of media content, MongoDB Atlas can be applied to multiple AI use cases. This allows for the utilization of media channels to more effectively improve end-user experiences.

By doing so, media organizations can suggest content that aligns more closely with individual preferences and past interactions. This not only enhances user engagement but also increases the likelihood of converting free users into paying subscribers.

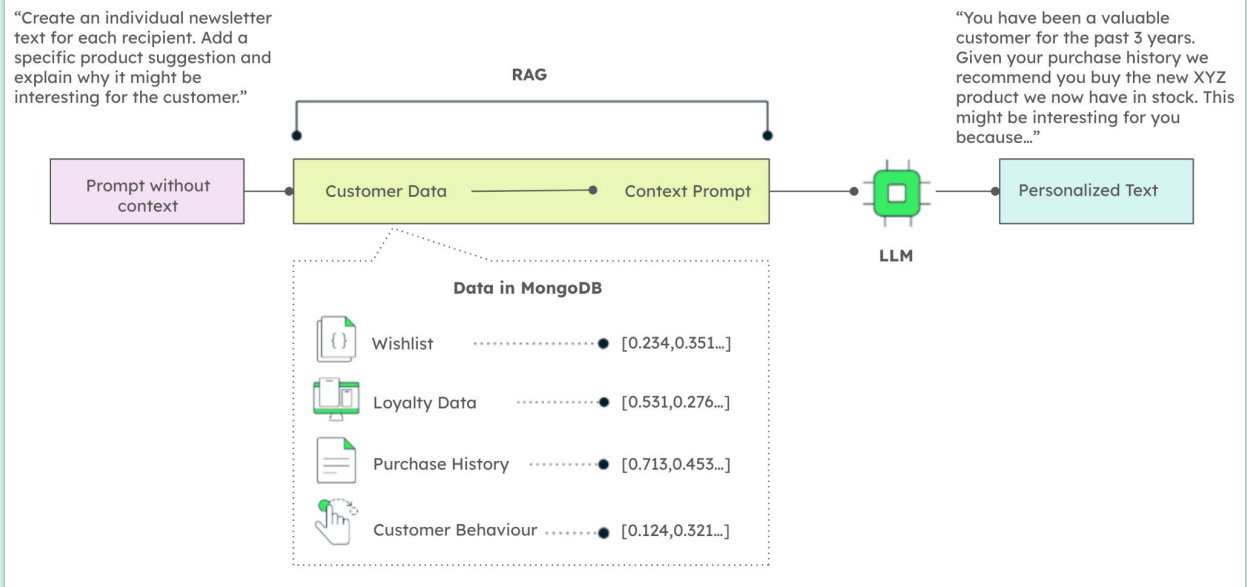


Figure 20: Example of the data flow for an AI-generated personalized newsletter. The prompt is entered by a user on the left hand side and context is added via the vectorized data in MongoDB—wishlist, loyalty data, purchase history, and customer behavior. Using RAG, the LLM can produce a personalized newsletter per customer in seconds, allowing the retailer to create vast amounts of customized content

Demand Forecasting & Predictive Analytics

Accurate demand planning using AI in retail optimizes inventory levels, reducing costs and stockouts while enhancing customer satisfaction through better availability of products. It also enables data-driven decisions, leading to improved sales forecasts and efficient supply chain management. Retailers either develop homegrown applications for demand prediction using traditional machine learning models or buy specialized products designed to provide these insights across the segments for demand prediction and forecasting. The homegrown systems require significant infrastructure for data and machine learning implementation and dedicated technical expertise to develop, manage, and maintain them. More often than not, these systems require constant care to ensure optimal performance and provide value to the businesses.

Subsequently, feature engineering is there to extract seasonality, promotions, impact, and general economic indicators. A retrieval augmented generation model can be incorporated to improve demand forecasting predictions and reduce the possibility of hallucinations. The same datasets could be utilized from historical data to train and fine-tune the model for improved accuracy.

Such efforts lead to the following business benefits:

- Precision in demand forecasting
- Optimized product / supply planning
- Accuracy in inventory management
- Enhanced customer satisfaction

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

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

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

MongoDB for Predictive Analytics

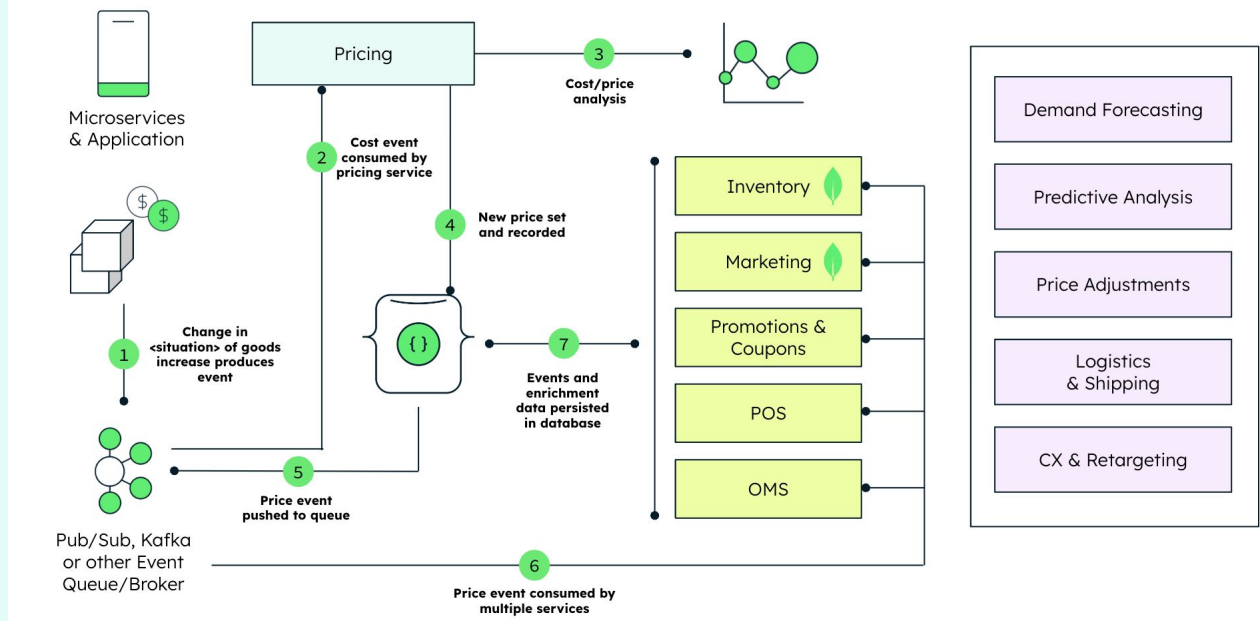


Figure 21: An illustration of a price-change scenario where fuel costs have risen, which leads to a rise in shipping costs and, in turn, pricing

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

Conversational Chatbots

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

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

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

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

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

To retrieve relevant information based on user queries, MongoDB Atlas Vector Search is utilized. In this example OpenAI is being used in tandem with Vector Search to generate answers to the customer questions. Using data from private data sources and enhanced by the LLMs the data is augmented, given context, and then returned to the user. The Azure OpenAI embeddings API are used to convert MongoDB documentation and user queries into vector embeddings to help find the most relevant content for queries using Atlas Vector Search.

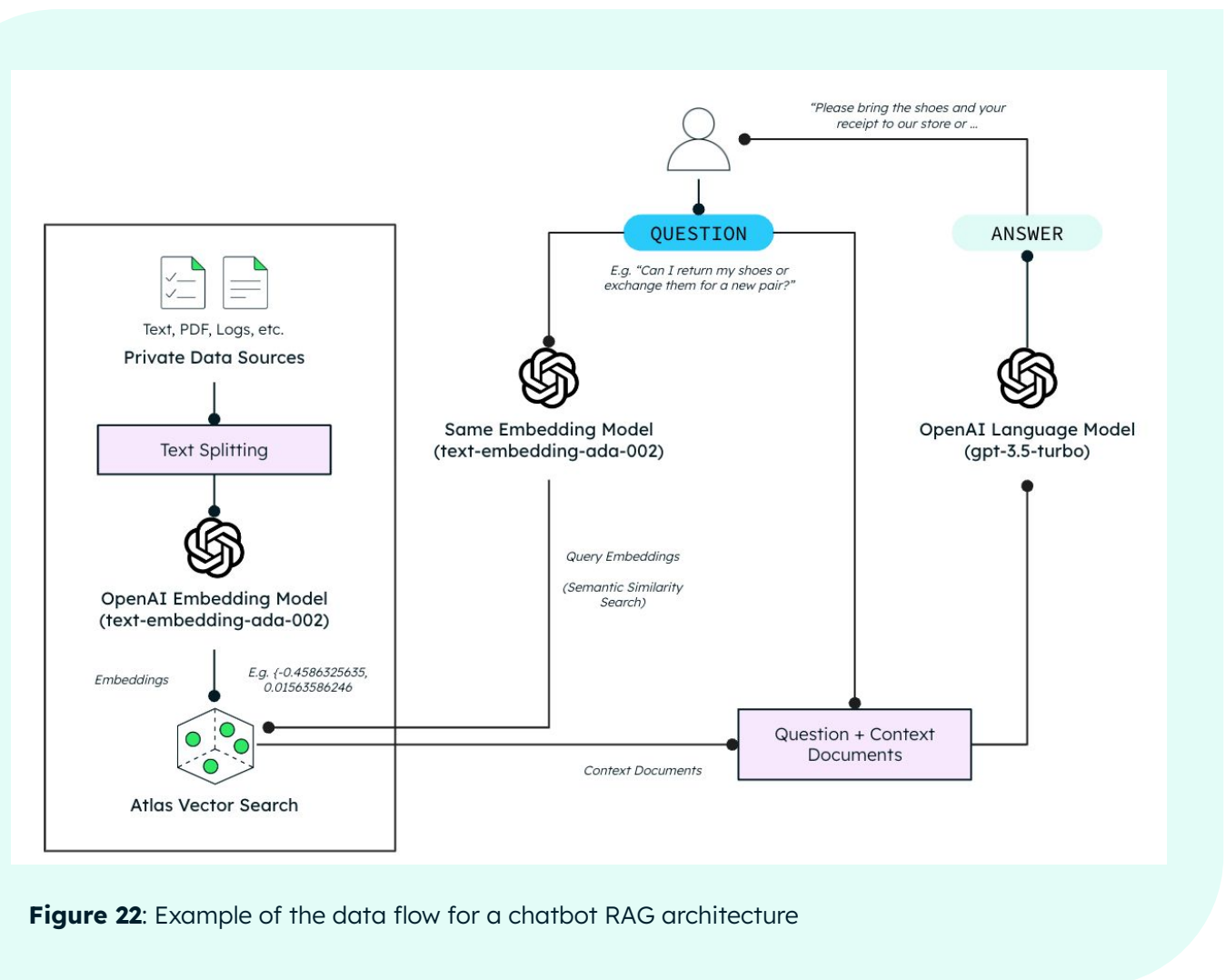


Figure 22: Example of the data flow for a chatbot RAG architecture



L'Oréal Improves App Performance and Velocity with MongoDB Atlas

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

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

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

[Learn more](#)

Moutia Khatiri

CTO, Tech Accelerator, L'Oréal

Artificial intelligence is revolutionizing the way retailers enhance their competitive edge by providing deeper insights into customer behavior and optimizing profit margins through smart decision-making processes. By incorporating both traditional and generative AI, retailers can harness the

benefits of enhanced and semantic search capabilities, create targeted marketing content based on current market trends, effectively utilize predictive analytics for demand forecasting, employ conversational chatbots, and significantly elevate the overall customer experience.

Contact Information



Genevieve Broadhead

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



Prashant Juttukonda

Retail Industry
Solutions Principal
prashant.juttukonda@mongodb.com



Rodrigo Leal

Retail Industry
Solutions Principal
rodrigo.leal@mongodb.com

FOR MORE INFORMATION AND RESOURCES

Visit MongoDB for Retail

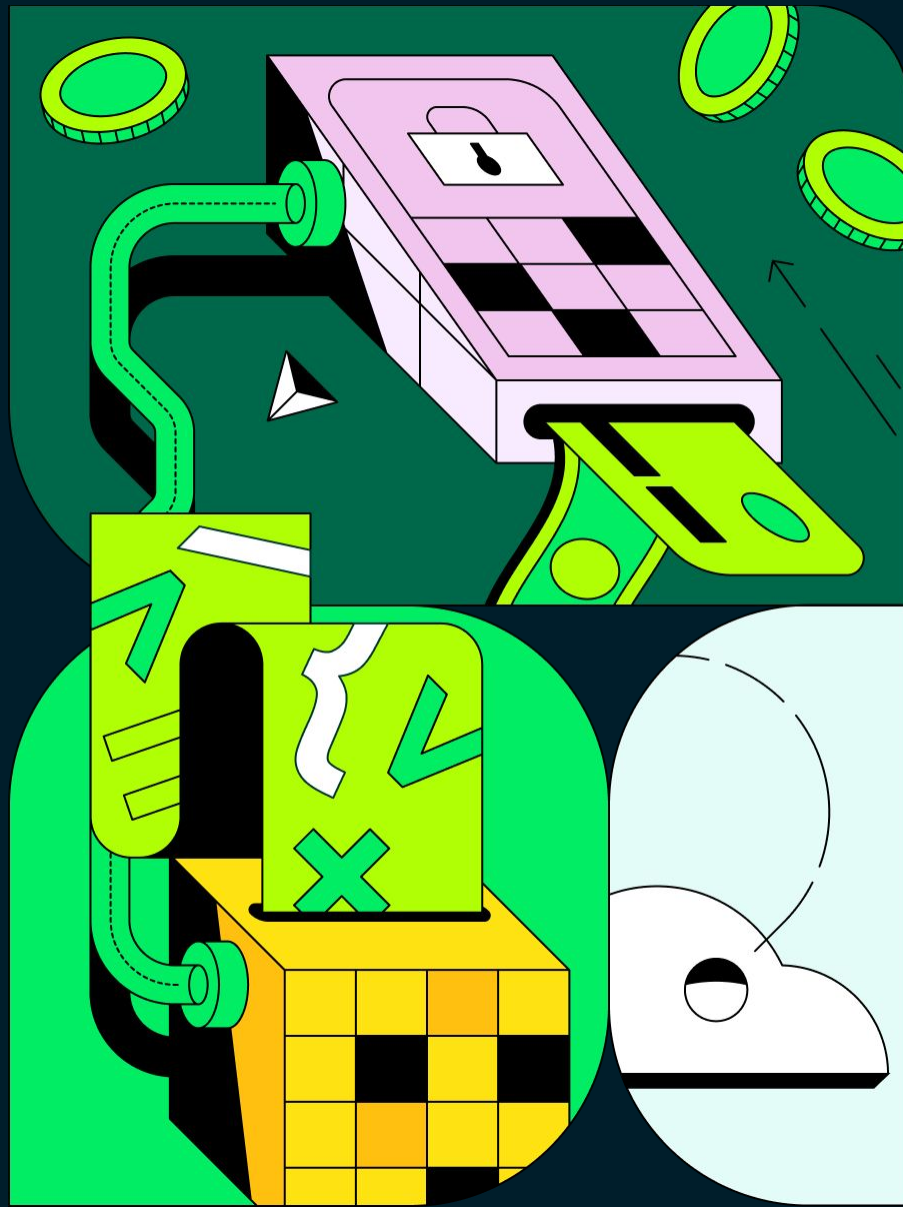




Atlas for Industries

Financial Services

Analyst firms, Celent and McKinsey, both identify financial services as one of the industries most likely to be transformed AI. While the initial benefits are likely to be process-related as firms slowly ramp up adoption, it won't be long before customer interactions are also completely re-invented by AI-driven experiences.



AI investments in the financial services industry



Already an integral part of the industry in areas such as risk, fraud, and compliance for example, the use of AI has been commonplace for years and continues to deepen, while there is a long (and growing) list of other use cases for which banks are also making use of AI.

AI technologies are used today to address a wide range of different workflows and customer-facing services from process automation and optimization in the middle and back office, to areas such as real-time risk and service personalization in the front office. However, it is to be expected that the benefits will be first in internal process optimizations and will only slowly start to be seen in direct customer engagement and interaction.

The conversation around generative AI became more nuanced through the latter part of 2023. This is understandable given the complexities of applying large language models (LLMs) to sensitive customer data, as well as broader regulatory concerns over the explainability of LLM outputs. That said, there are many areas in which gen AI is already being used to support advisors and relationship managers, with further innovation expected. According to Celent's [Harnessing the Benefits of AI in Payments](#) report, 58% of banks are evaluating or testing generative AI in some capacity (as shown below) while a further 23% have projects using this in their roadmap.

Interest in leveraging AI for banking use cases is expected to increase in the next few years.

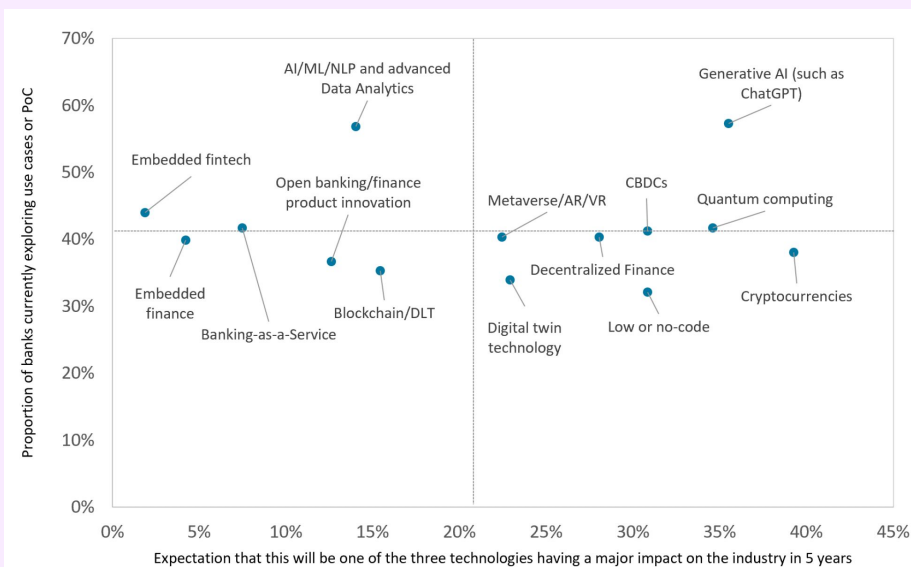


Figure 23: Bank tech exploration vs. expected impact

Addressing the challenges of AI in finance



While the industry has always had to deal with persistent issues like risk management and governance, adopting generative AI and machine learning introduces new challenges that AI specialists have always dealt with, like inherent biases and ethical concerns. One challenge that stands out for generative AI is hallucination—the generation of content that is not accurate, factual, or reflective of the real world.

Generative AI models, especially in natural language processing, might generate text that is coherent and contextually appropriate but lacks factual accuracy. This poses challenges, for example:

- **Misleading financial planning advice:** In financial advisory services, hallucinated information may result in misleading advice, leading to unexpected risks or missed opportunities.
- **Incorrect risk assessments for lending:** Inaccurate risk profiles may lead to poor risk assessments for loan applicants, which can cause financial institutions to approve a loan with a higher risk of default than the firm would normally accept.
- **Sensitive information in generated text:** When generating text, models may inadvertently include sensitive information from the training data. Adversaries can craft input prompts to coax the model into generating outputs that expose confidential details present in the training corpus.

A strategic and comprehensive approach encompassing aspects of technology, data, ethics, and organizational readiness is critical to overcoming these challenges:

- **Hallucination mitigation:** One promising strategy is using [retrieval augmented generation](#) (RAG) to mitigate hallucination in gen AI models—incorporating information retrieval mechanisms to enhance the generation process to ensure content is grounded in real-world knowledge. [Atlas Vector Search](#) is a popular mechanism to support RAG, which uses vector embeddings to retrieve relevant documents based on the input query.
- **Data quality and availability:** Take a step back before adopting AI to ensure the quality, relevance, and accuracy of data being used for AI training and decision-making can be accessed in real time.
- **AI education:** The key is to invest in training programs to address skill gaps, create a culture of learning and development, and promote awareness about vulnerabilities.
- **Develop new governance, frameworks, and controls:** Before going live, create safe and secure environments for testing.
- **Implement monitoring systems:** Measure and understand financial impacts, change impacts, scale, and complexity associated with AI.
- **Security and privacy:** Implement secure data access privileges and authentication measures to safeguard AI models and the data they rely on.

Relationship Management Support with Chat Bots



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.*

Traditionally, relationship managers (RMs) faced challenges accessing and analyzing client portfolio data. Tasks like finding coupon dates were time-consuming and required manual effort. To streamline this process, RMs often requested new reports, but finding the right information could be difficult.

Generative AI offers a solution by automating data retrieval and analysis. Through RAG, AI can provide answers and explanations, and even execute queries based on diverse data sources, including portfolio information, user manuals, and customer data. RMs can interact with the AI to refine results and ensure accuracy.

MongoDB's platform is well suited for this application. It can handle various data types, including structured and unstructured information, and it provides efficient data search and processing tools. This enables the development of AI-powered applications that deliver timely insights to RMs.

There are different approaches of building a RAG-based chatbot application.

A simple approach would be like the one explained in our article [Taking RAG to Production with the MongoDB Documentation AI Chatbot](#) with an architecture diagram as follows.

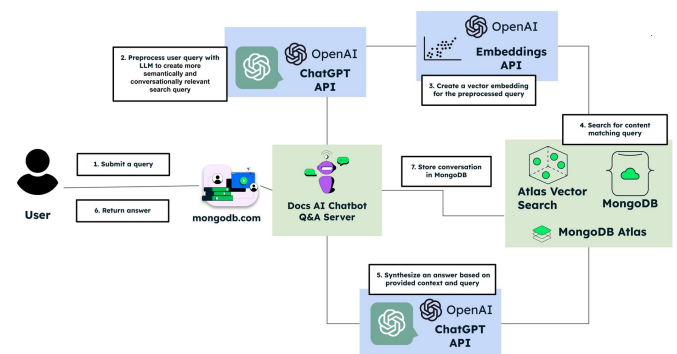


Figure 24: User query processing flow

The more advanced approach would be to leverage a composable RAG framework that is developed as part of the [MongoDB AI Applications Program \(MAAP\)](#).

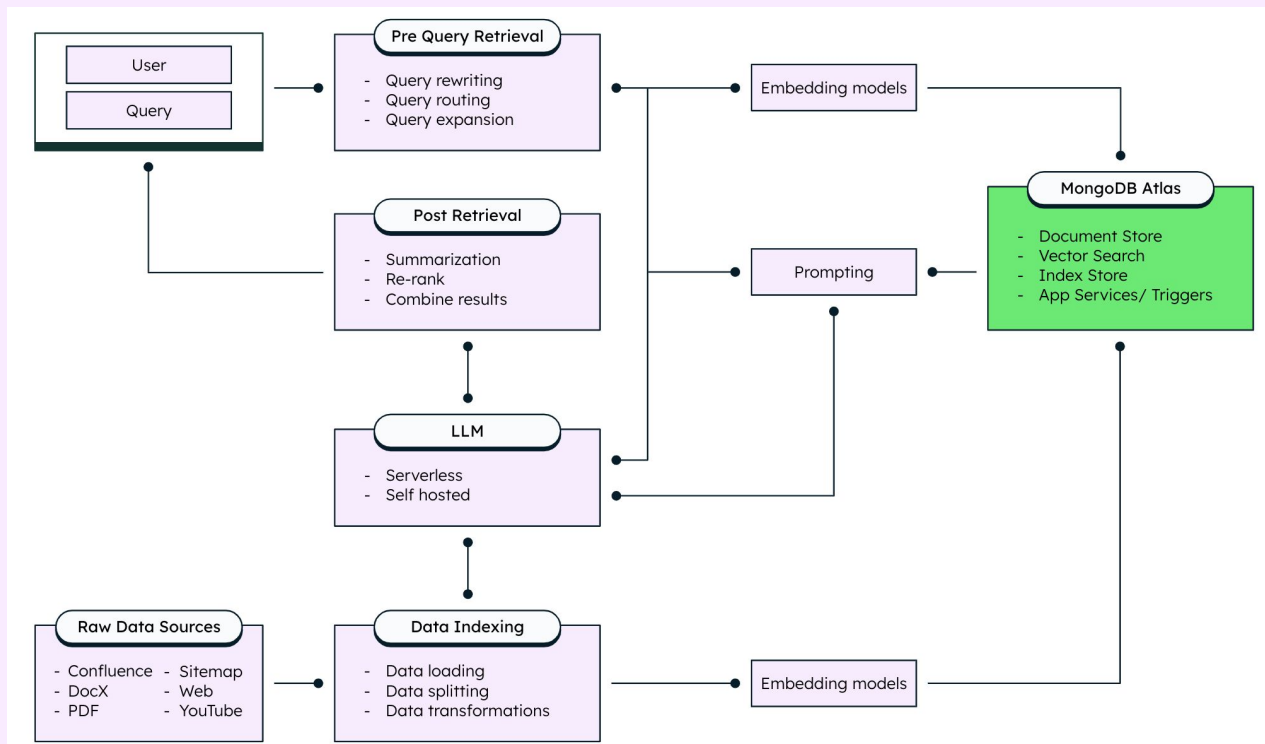


Figure 25: Query processing pipeline

Data loading from raw data sources

Begin by extracting data from databases or files. Examples of data can include customer profiles, transaction histories, and product information. The source of the data may be located in different databases or files and can be consolidated into MongoDB and serve as an operational data store to provide a customer 360-degree view.

Data indexing: chunking and vectorization

- **Data chunking** involves breaking down the documents (such as customer service records, emails, and transaction logs) into smaller, more manageable chunks (eg. individual interactions or sentences) and then creating an index that allows for quick searching and retrieval of relevant information.
- **Data vectorization** is implemented by using the embedding model (eg. Cohere, Anthropic, etc) to convert the above

mentioned text chunks into vector representations. The chunks can then be indexed using MongoDB vector indexing that will then enable an efficient retrieval of vectorized data using [Atlas Vector Search](#).

Pre-query Retrieval

Advanced RAG methods explore various transformations of user queries to enhance retrieval accuracy before it is passed to the LLM for response generation. As an example, If the RM asks for all recent fund transfers for a given customer, the gen AI application can pre-process the query by first checking if the RM has access privileges to the customer data. It can then refine the query to add a specific time period, and add the relevant documents retrieved from the vector search. This will augment the search, giving it context to optimize relevancy of the results.

Post retrieval

With the transformed query executed, the relevant documents will be retrieved (such as with regular, text, vector or even hybrid searches). The retrieved documents (such as the customer details and transactions) can be further filtered and reranked to provide an even more relevant set of documents to be used as a data augmentation when prompting the LLM later.

LLM

The LLM engine can be a single or a group of LLM engines deployed on an AI platform (eg. Fireworks.AI) to work together leveraging different specialization to provide a relevant, accurate and combined response to the users. Let us generalize and refer to these engines as **AI agents** so to not limit our imagination to just Language Models (be it large or small). These AI Agents not only can perform orchestration tasks but can agent can

also be trained to specialize in a specific domain for accuracy and efficiency. For example in the context of financial advisory support, an agent could specialized in Natural Language Understanding of financial market information, another AI Agent can specialist in Risk Assessment for evaluating client and product risk profiles while yet another can specialized in Compliance for regulatory adherence to ensure the investment decisions especially for corporate clients are also compliant. Together they can be “composed” using orchestration tools (eg. Langchain and LLamaIndex) to support the tasks required for the generative AI interactions.



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, 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 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.

Modernizing fraud systems improves detection accuracy, increases flexibility and scalability, enhances security, and reduces the cost of operation. MongoDB's developer data platform is a great fit for modern fraud prevention systems due to:

Data visibility: MongoDB's document model incorporates any kind of data no matter how often it changes, allowing you to create a holistic picture of customers to better predict transaction anomalies in real time. As well as process large amounts of data and analyze in real-time to identify suspicious activity. [Change Streams in MongoDB](#) enable the monitoring of database changes in real time meaning immediate detection of suspicious activity or transactions. By receiving instant notifications, fraud prevention teams can quickly respond to potential threats.

Real-time data processing: By leveraging MongoDB as the [operational data store](#) (ODS) you can handle large amounts of data to effortlessly capture, store, and

process high-volume transactional data in real-time. [MongoDB's aggregation framework](#) allows for complex data processing and analytics directly within the database, crucial for implementing sophisticated, rules-based fraud detection algorithms.

Integrating with modern AI/ML

technologies: MongoDB's flexible data model makes it easier to integrate with AI/ML platforms to develop models for real-time, predictive fraud prevention that are trained on the most accurate and relevant data available.

Regulatory compliance: Compliance with regulations like PCI, GDPR, CCPA, and PSD2 is crucial in the finance sector. [MongoDB's security features](#) ensure data handling meets these regulatory standards. Enhanced authentication and role-based authorization controls further fortify the system.

Besides the use of AI/ML techniques, there the option of leveraging vector data store capabilities such as [MongoDB Atlas Vector Search](#) to perform fraud and money laundering detection. This can be also be used in conjunction with the AI/ML approach to create a hybrid implementation to further enhance the accuracy of detection and minimize false positives. [Learn More](#)

Below is an [example architecture of an ML-based fraud solution](#) leveraging MongoDB and Databricks. Key features include data completeness through integration with external sources, real-time processing, AI/ML modeling to identify potential fraud patterns, real-time monitoring for instant analysis, and model observability for full data visibility into fraud behaviors. [Learn more](#)

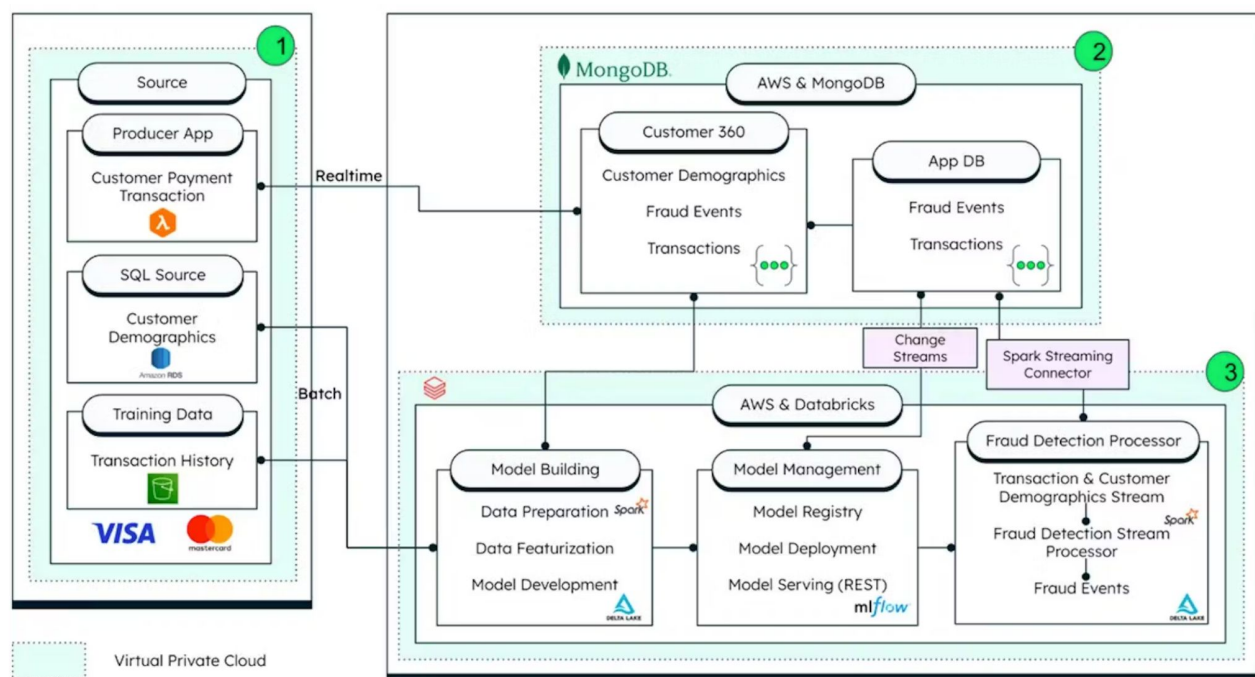


Figure 26: Example fraud detection solution leveraging MongoDB

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). 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.

Discover more about MongoDB's full full-text search and vector search capabilities.

To achieve search and summarization using RAG, a conceptual architecture leveraging MongoDB and generative AI can be used.

This architecture aims to centralize document storage, enhance search capabilities, and automate the extraction and synthesis of information. The following diagram illustrates this conceptual architecture.

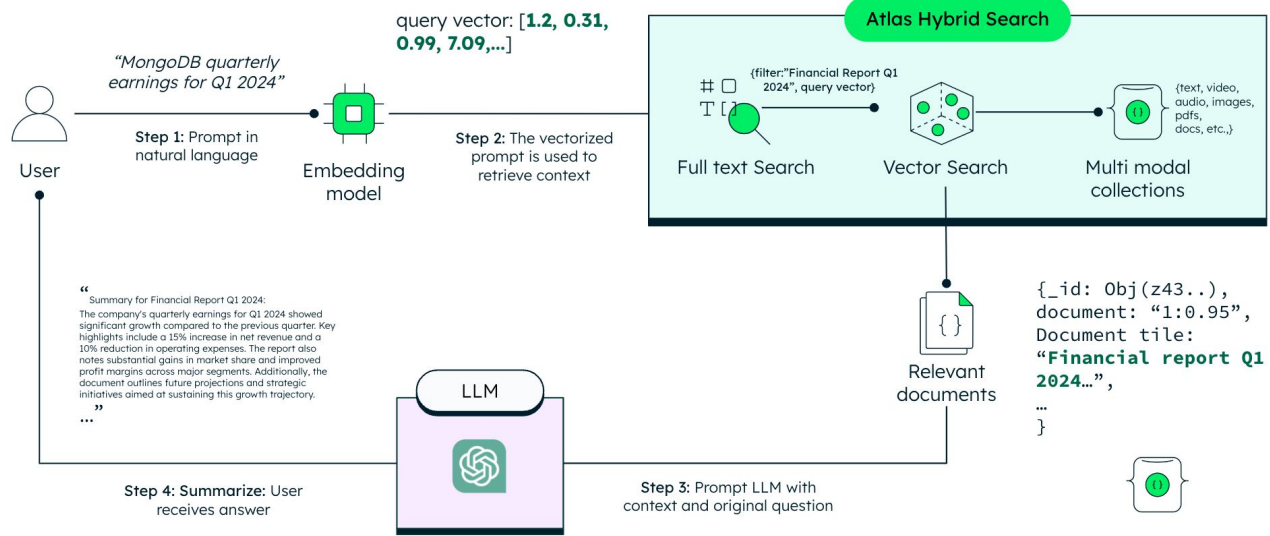


Figure 27: Atlas hybrid search workflow: query embedding, document retrieval, summarization, user presentation

ESG Analysis

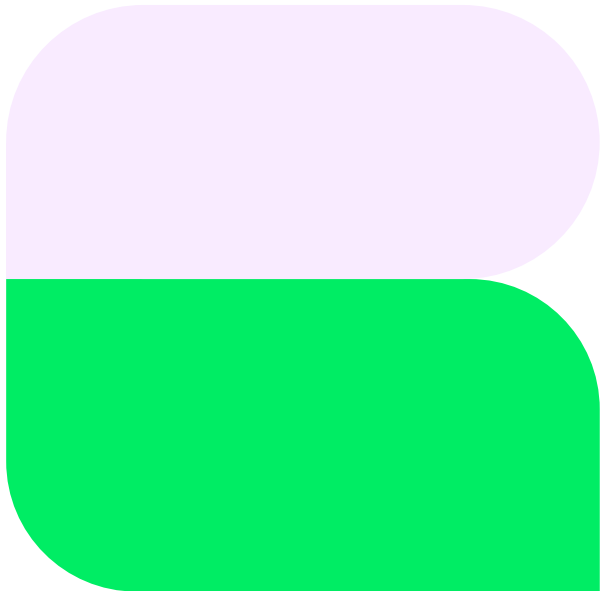
The profound impact of environmental, social, and governance (ESG) principles is evident, driven by regulatory changes, especially in Europe, compelling financial institutions to integrate ESG into investment and lending decisions. Regulations such as the [EU Sustainable Finance Disclosure Regulation \(SFDR\)](#) 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 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 a leaf where MongoDB can be deployed to support the ESG data analytics services.

Learn how MongoDB makes it incredibly easy to collect, analyze, and visualize ESG data in real time.

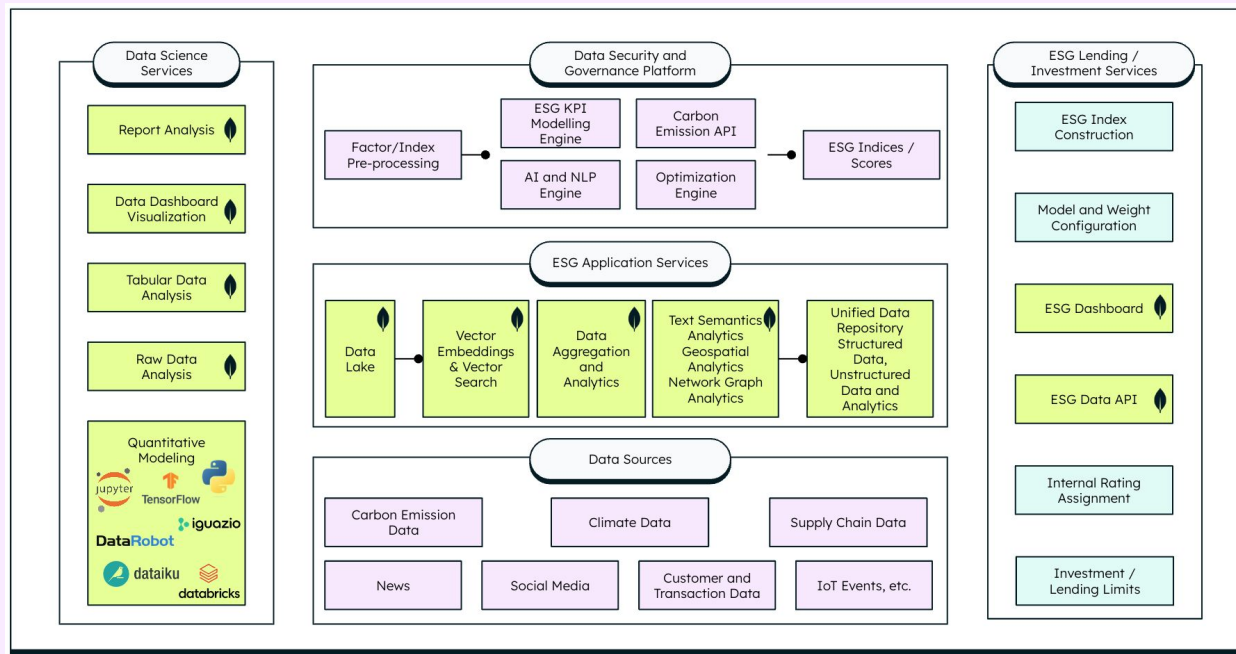


Figure 28: Blueprint for Enterprise ESG Solution Architecture Using MongoDB

Transforming Credit Scoring with AI

The convergence of alternative data, artificial intelligence, and generative AI is reshaping the foundations of credit scoring, marking a pivotal moment in the financial industry. The challenges of traditional models are being overcome by adopting alternative credit scoring methods, offering a more inclusive and nuanced assessment. Generative AI, while introducing the potential challenge of hallucination, represents the forefront of innovation, not only revolutionizing

technological capabilities but fundamentally redefining how credit is evaluated, fostering a new era of financial inclusivity, efficiency, and fairness.

The use of artificial intelligence, in particular generative artificial intelligence, as an alternative method to credit scoring has emerged as a transformative force to address the challenges of traditional credit scoring methods for several reasons.

Alternative data analysis: Unlike traditional models that rely on predefined rules and historical credit data, AI models can process a myriad of information, including alternative data, such as utility payments and rental history, to create a more comprehensive assessment of an individual's creditworthiness, ensuring that a broader range of financial behaviors is considered.

AI offers unparalleled adaptability: As economic conditions change and consumer behaviors evolve, AI-powered models can quickly adjust and learn from new data. This continuous learning ensures that credit scoring remains relevant and effective in ever-changing financial landscapes.

Fraud detection: AI algorithms can detect fraudulent behavior by identifying anomalies and suspicious patterns in credit applications and transaction data.

Predictive analysis: AI algorithms, particularly ML techniques, can be used to build predictive models that identify patterns and correlations in historical credit data, forecasting the greater likelihood of loan defaults.

Behavioral analysis: Many individuals encounter hurdles in the form of limited or nonexistent credit history, making it difficult to prove their creditworthiness due to the lack of historical data. AI algorithms can analyze behavioral data sets to understand financial habits and risk propensity. By monitoring real-time financial behavior, AI models can provide dynamic credit scores that reflect current risk profiles.

Lack of clarity in rejection reasons: A lack of transparency in rejection reasons leaves applicants in the dark, making it difficult for them to address the root cause and enhance their creditworthiness for future applications.

By harnessing the power of artificial intelligence, lenders can make more informed lending decisions, expand access to credit, and better serve consumers (especially those with limited credit history). However, to mitigate potential biases and ensure consumer trust, it's crucial to ensure transparency, fairness, and regulatory compliance when deploying artificial intelligence in credit scoring.

The most common objections from banks to not using AI in credit scoring are transparency and explainability in credit decisions. The inherent complexity of some AI models, especially deep learning algorithms, may lead to challenges in providing clear explanations for credit decisions. Fortunately, the transparency and interpretability of AI models have seen significant advancements. Techniques like [SHapley Additive exPlanations \(SHAP\) values](#) and [Local Interpretable Model-Agnostic Explanations \(LIME\) plots](#) and several other advancements in the domain of Explainable AI (XAI) now allow us to understand how the model arrives at specific credit decisions. This not only enhances trust in the credit scoring process but also addresses the common critique that AI models are "black boxes."

Gen AI has the potential to revolutionize credit scoring and assessment with its ability to create synthetic data and understand intricate patterns, offering a more nuanced, adaptive, and predictive approach. Gen AI's capability to synthesize diverse data sets addresses one of the key limitations of traditional credit scoring—the reliance on historical credit data. By creating synthetic data that mirrors real-world financial behaviors, gen AI models enable a more inclusive assessment of creditworthiness.

Adaptability plays a crucial role in navigating the dynamic nature of economic conditions and changing consumer behaviors. Unlike traditional models that struggle to adjust to unforeseen disruptions, gen AI's ability to continuously learn and adapt ensures that credit scoring remains effective in real-time, offering a more resilient and responsive tool for assessing credit risk.

In addition to its predictive prowess, gen AI can contribute to transparency and interpretability in credit scoring. Models can generate explanations for their decisions, providing clearer insights into credit assessments, and enhancing trust among consumers, regulators, and financial institutions.

Enhance credit applications and scoring with MongoDB in the following key aspects of the process:

Simplify data capture and processing:

streamlining this process by eliminating redundant requests and leveraging existing data could significantly enhance the user experience. Leveraging JSON for online credit application forms simplifies the data capture process and also the performance in data processing. JSON's structured data representation proves highly conducive for organizing the multifaceted information within credit applications, encompassing personal, financial, and employment details.

Enhance credit scoring with AI:

Leveraging [MongoDB's developer data platform](#) — an integrated suite of data services centered around a cloud database — we can create a comprehensive customer/user banking profile by combining relevant data points.

Below is an architectural diagram of the data processing pipeline for the predicting probability of delinquency and credit scoring. The data pipeline for credit scoring a customer involves the following steps: data collection, data processing, risk profile generation, model development, data transformation, and decision collection.

Explain the credit application

declination: when it comes to credit application declination, understanding the reasons behind it is crucial. Let's explore how MongoDB and [large language models](#) (LLMs) can shed light on XGBoost model predictions (the model used in this [tutorial](#)). The risk profiling ML pipeline employed provides a probability score that defines the risk associated with the profile for product recommendation. This message is communicated back to the user in a templated manner where only the final status of the application is communicated to the end user.

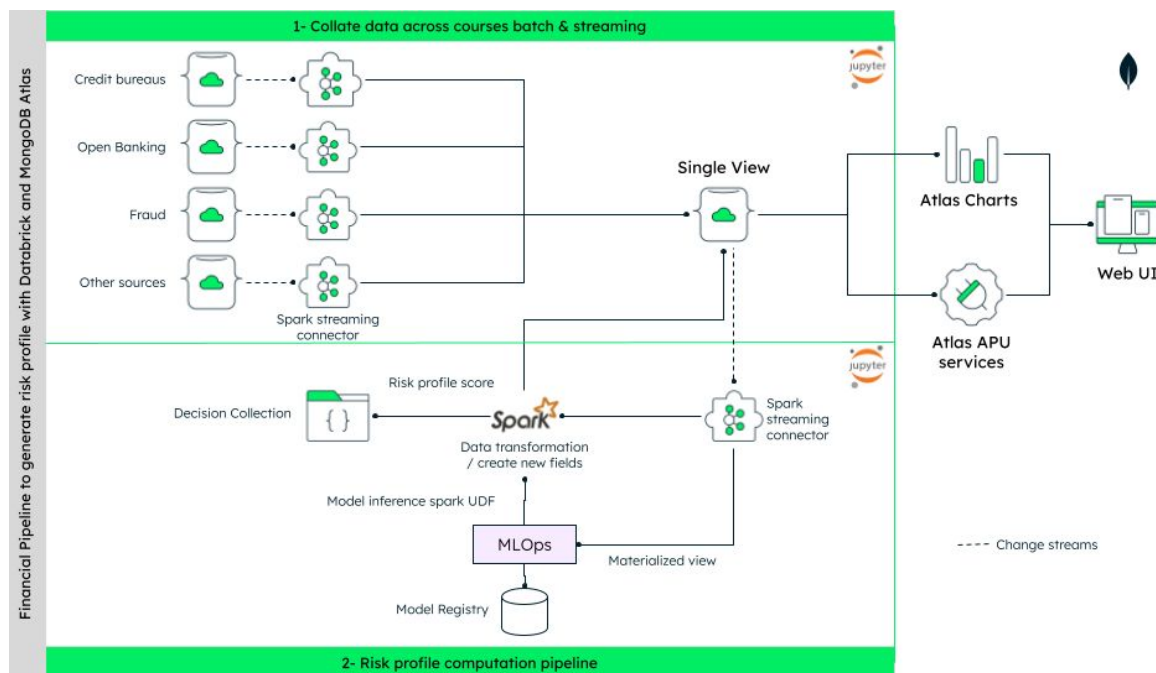


Figure 29: Architecture of the data processing pipeline for the predicting probability of delinquency and credit scoring

Below you will see the architecture explaining credit scoring using an LLM.

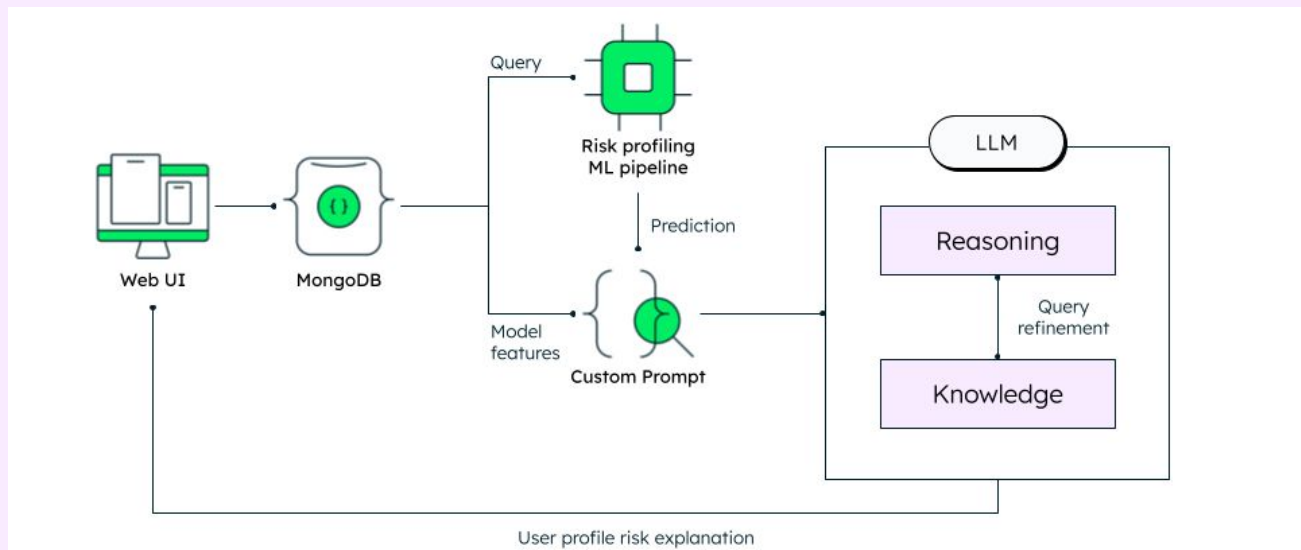


Figure 30: Architecture of the credit application declination

Recommend alternative credit products:

If the credit product applicant is declined, the credit institution should still try to cross-sell to the customer with a relevant product that meets their needs as they are

already engaged in the process. Below is an example of a data architecture that is used to achieve this.

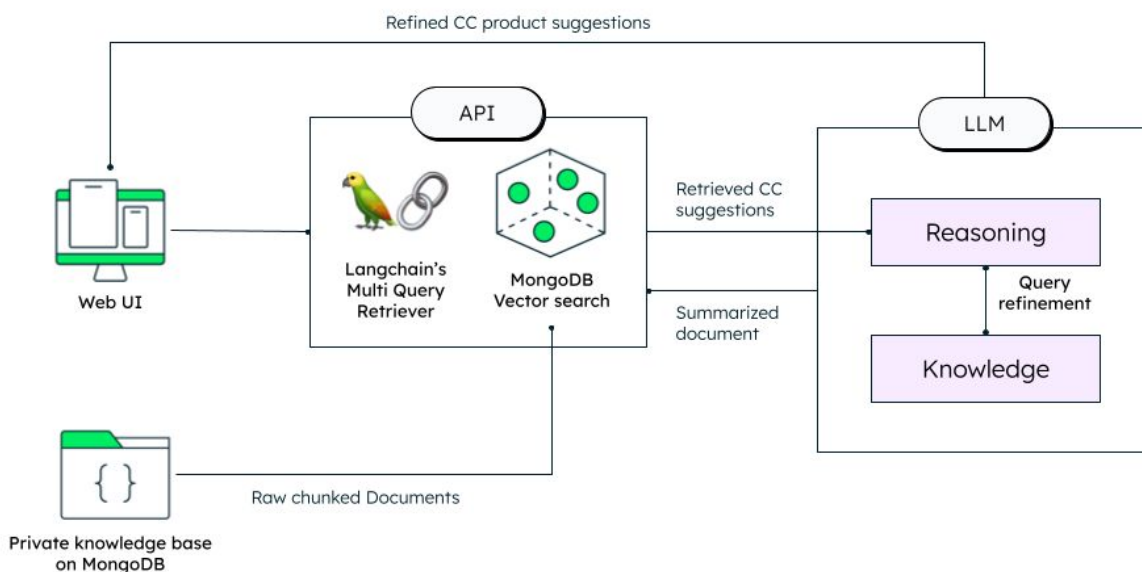


Figure 31: Architecture of the alternative credit product recommendations

Learn how with this credit card Credit card application with generative AI tutorial

Assessing Business Loan Risks

Business loans are a cornerstone of banking operations, providing significant benefits to both financial institutions and the broader economy. In 2023, the value of commercial and industrial loans at commercial banks in the United States reached nearly 2.8 trillion U.S. dollars. However, these loans involve significant risks, including credit and business risks.

Importance of Business Plans

A comprehensive business plan is essential in securing business loans, providing detailed insights into a borrower's strategy, market position, operational plans, and financial forecasts. However, the complexity and length of these plans pose challenges for loan officers, including time constraints and the need for specialized knowledge to interpret key metrics and risk factors accurately.

Generative AI in Risk Assessment

Generative AI can streamline the analysis of business plans by extracting essential information and highlighting key risks. This includes using gen AI-powered chatbots that simulate discussions with loan officers, providing real-time risk assessments based on the input data.

For instance, ChatGPT-4.0 can be prompted to perform a flood risk assessment. However, due to insufficient flood data, it first indicated the location was in a low flood risk zone but later claimed it had significant flooding. This inconsistency can confuse users and lead to incorrect risk assessments.

The screenshot displays a chat interface with a user asking for business risks, profitability projections, and repayment ability for a \$100,000 loan for a bakery at 1633 Broadway, New York, NY 10019, USA. The AI responds with an overview and then details business risks, including market competition. It then addresses a follow-up question about flood risk, stating that the location is generally not in a high flood risk zone but that FEMA flood maps and local history should be reviewed. Finally, it answers a question about recent flooding within 5km of the location, stating that the area has experienced significant flooding in recent years due to storms like Tropical Storm Henri and Hurricane Ida.

What are the business risks, profitability projections, and repayment ability of a \$100,000 loan for starting a bakery at 1633 Broadway, New York, NY 10019, USA?

Starting a bakery at 1633 Broadway, New York, NY, a prime location in Midtown Manhattan, involves careful consideration of several factors, including business risks, profitability projections, and the ability to repay a \$100,000 loan. Here's an overview of these aspects:

1. Business Risks

Market Competition

- **High Competition:** Midtown Manhattan is a bustling area with numerous established bakeries, cafes, and food outlets. New entrants face significant competition from well-known brands

The full risk assessment is more detailed

assume you are a loan officer analyzing the risk of this business. What is the business risk flood risk, profitability projection and repayability of a 100000USD loan for starting a bakery at 1633 Broadway, New York, NY 10019, USA?

Flood Risk

1. **Location Analysis:**

- 1633 Broadway, NY, is generally not in a high flood risk zone, but it's essential to review FEMA flood maps and local flood history.

has there been any flood within 5km of 1633 Broadway, New York, NY 10019, USA?

Searched 4 sites

2. **Flood Risk:**

- **Recent Flood Events:** The area has experienced significant flooding in recent years due to storms such as Tropical Storm Henri and Hurricane Ida, both of which brought heavy rainfall

Figure 32: Examples of how ChatGPT-4.0 could respond when asked to assess the risk of a business loan and how it responded to flood risk as a factor

Retrieval-augmented generation (RAG) and MongoDB

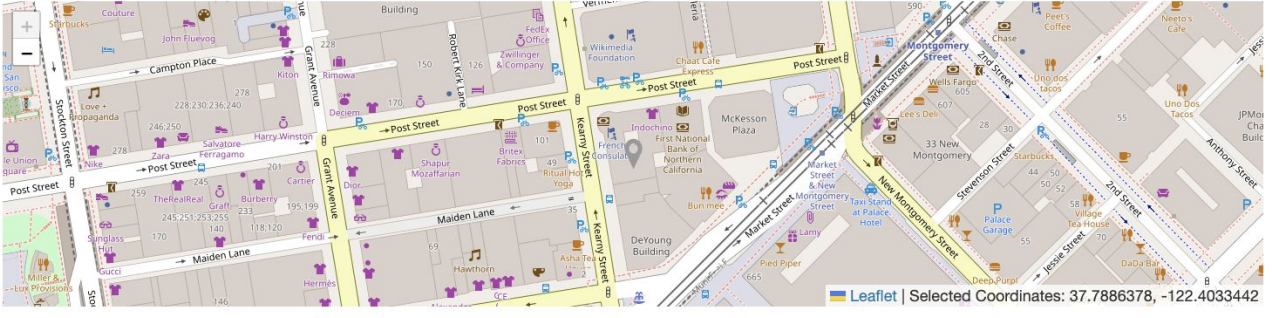
While gen AI offers significant potential, its effectiveness in loan officer workflows is hindered by the need for constant prompting and context augmentation. This manual process is time-consuming and often requires specialized skills. To overcome these challenges and enhance data accuracy, we propose a retrieval-augmented generation (RAG) approach utilizing MongoDB's multimodal capabilities. By incorporating geospatial data and other external datasets, RAG can provide a more comprehensive risk analysis. For instance, MongoDB's geospatial queries can identify historical flood occurrences near a business location, enriching AI-generated insights with crucial context.

Demonstration and Results

Let's examine several figures from a generative AI RAG demo built using LLama 3 on the Fireworks.AI platform, with MongoDB as an operational data store for geospatial queries. In this demo, users are asked for a business location using a map (where they can either choose one of the defaults from a dropdown list or use a geolocation pin to pick a location on the map) and a short business purpose and description (to emulate the idea of a business plan).

Please indicate the business location in the USA for your real estate by clicking on the map, or entering the address on the search bar.

Q 88 Kearny St Suite 500, San Francisco, CA 94108, United States



Leaflet | Selected Coordinates: 37.7886378, -122.4033442

Please provide a brief description of your loan purpose and business plan.

Vibrance Studios is a unique, membership-based art studio offering a variety of creative classes and workshops for adults and children. Our studio will provide a welcoming space for individuals to explore their artistic side, relax, and socialize. We'll offer classes in painting, pottery, sculpture, and more, taught by experienced local artists.

We're seeking a \$400,000 loan to cover startup costs. The funds will be allocated as follows: \$200,000 for studio design and build-out, \$100,000 for initial marketing and advertising efforts, and \$100,000 for working capital to cover instructor fees, supplies, and rent until we reach profitability.

One potential risk is competition from established art studios in the area, but we believe our unique membership model and diverse class offerings will set us apart. Additionally, we'll focus on building a strong online presence and community to attract and retain members.

Clear Example Submit

Figure 33: Choosing a location on the map and writing a brief plan description

Upon submission, the AI performs a RAG analysis. Using prompt engineering, it delivers a simplified business analysis considering the location and flood risk data from external sources. [McKinsey's survey](#) found that 65% of respondents report their organizations are regularly using generative AI, including RAG, in at least one business function. This demonstrates the increasing adoption and integration of RAG within business processes.

Assessor's response

As a risk assessor for a business loan company, I have analyzed the business description, address, and flood data for Vibrance Studios. Here is my assessment:

Business Description and Cost/Revenue Projections:

Vibrance Studios is a unique, membership-based art studio offering a variety of creative classes and workshops for adults and children. The business is seeking \$400,000 to cover startup costs, with \$200,000 allocated for studio design and build-out, \$100,000 for initial marketing and advertising efforts, and \$100,000 for working capital to cover instructor fees, supplies, and rent until profitability. Based on industry benchmarks, I estimate the average monthly revenue to be around \$40,000, with a projected annual revenue of \$480,000. The break-even point is estimated to be around 8-10 months after launch.

Local Competition and Demographics:

The location of Vibrance Studios in San Francisco, CA, is a densely populated area with a high demand for creative activities. The demographics suggest a diverse population with a strong interest in art and culture. However, there is moderate competition from established art studios in the area, which may impact market share.

Flood Risk Assessment:

The flood data indicates a high risk of flooding at the proposed location. There were 7 floods less than 5km away last year, and a total of 44 floods less than 5km away over the past 5 years. This suggests a recent and frequent history of flooding in the area. Although the number of floods over 5km away is higher, the proximity and recency of the floods less than 5km away are a significant concern. The flood risk is further exacerbated by the fact that the business is located in a flood-prone area, as indicated by the University of Oklahoma's crowdsourced database, Dartmouth Flood Observatory, and other sources.

Risk Analysis:

The high risk of flooding poses a significant threat to the business, particularly in terms of property damage and disruption to operations. The lack of flood insurance coverage was not mentioned in the business description, which is a concern. The business may need to consider investing in flood mitigation measures, such as flood-resistant construction and emergency preparedness plans. Overall, while Vibrance Studios has a unique business concept and a growing market, the high risk of flooding at the proposed location is a significant concern. The business will need to carefully consider and mitigate this risk to ensure long-term success.

Figure 34: Loan risk response using RAG

Clicking the "Pin" icon next to the location bar reveals sample flood locations near the selected business site. The pins mark flood locations, and a blue circle shows a 5km radius where flood data is queried.

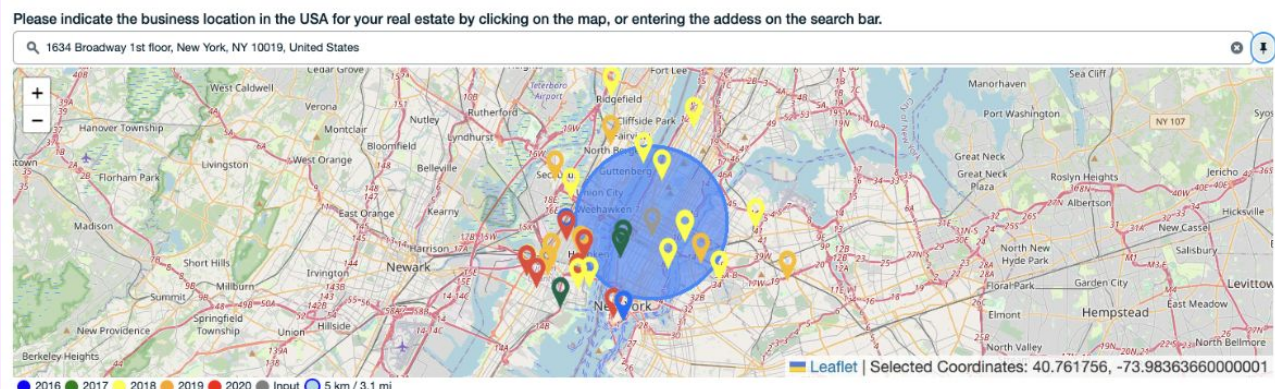


Figure 35: Flood locations displayed with pins

RAG and multimodal data: a robust approach to loan risk assessment

Integrating gen AI with a multimodal developer data platform like MongoDB Atlas offers a robust solution for detailed and accurate business loan risk assessments. The iterative nature of the RAG process ensures continuous learning and improvement, minimizing risks like data hallucinations. As AI models become increasingly multimodal, leveraging comprehensive data platforms will be crucial for maximizing their potential.

Solution: Better Business Loans with MongoDB and Generative AI.

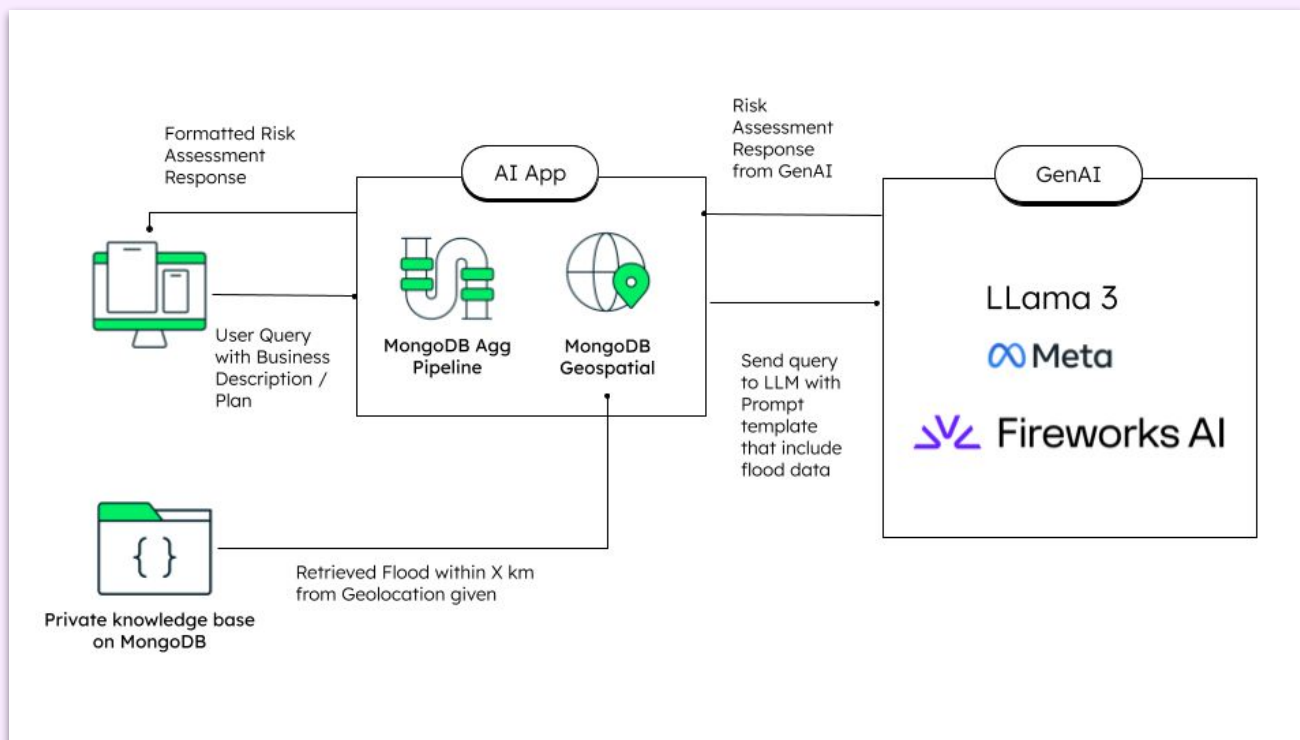


Figure 36: RAG data flow architecture diagram

Emerging use cases for AI in payments

A lack of developer capacity is one of the biggest challenges for banks when delivering payment product innovation. Banks believe the product enhancements they could not deliver in the past two years due to resource constraints would have supported a [5.3% growth in payments revenues](#). With this in mind and the revolutionary transformation with the integration of AI, it is imperative to consider how to free up developer resources to make the most of these opportunities. There

are several areas in which banks can apply AI to unlock new revenue streams and efficiency gains. The image below provides a high-level view on eight of the principal themes and areas. This is not an exhaustive view but does demonstrate the depth and breadth of current opportunities. In each example, there are already banks that have begun to bring services or enhancements to market using AI technologies or are otherwise experimenting with the technology.

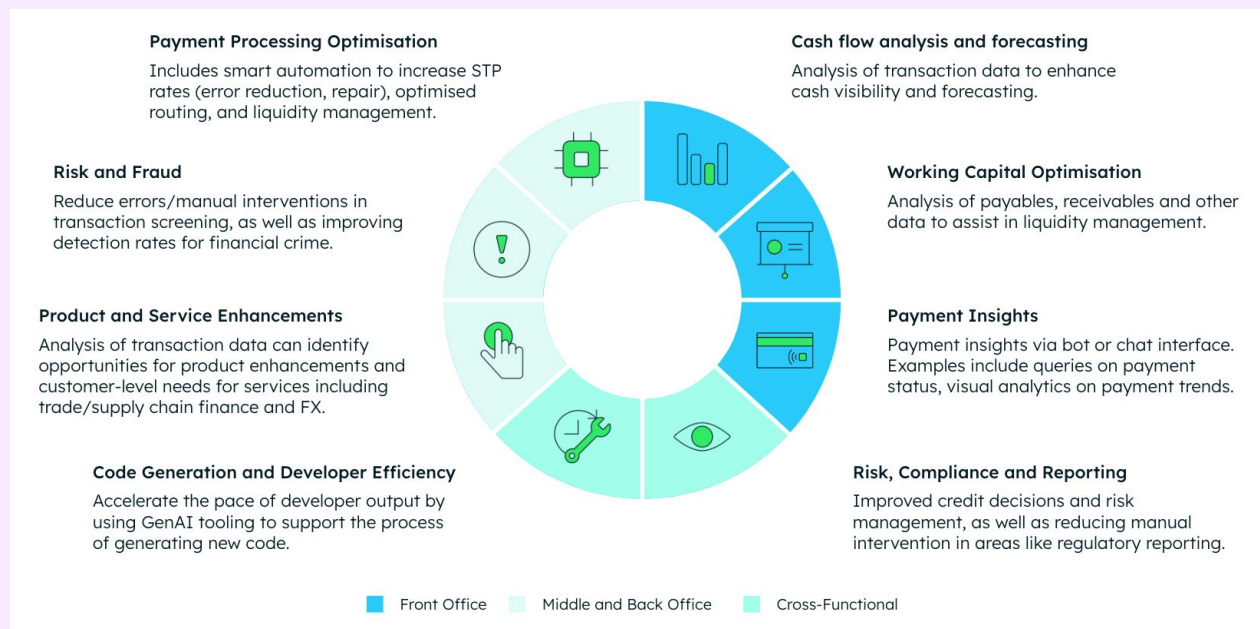


Figure 37: Benefits of payment processing optimization

Below are some areas [according to tech analyst Celent in which AI can be leveraged](#) across financial institutions in front, middle and back office operations.

Front office for new and richer insights:
AI has huge potential to support

value-added services making the most of payments data. Analysis of payables, receivables, and balance information in aggregate can be used to deliver a range of different insights and service improvements in real-time. Below are three areas where

financial institutions can use AI to enhance their customer-facing services:

- **Cash flow analysis and forecasting:** Using the information in payment messages to provide granular and real-time insights into the cash position for a client.
- **Working capital optimization:** Using AI in conjunction with payments data can support a range of use cases around liquidity and management with insights and recommendations.
- **Payment insights:** Gen AI can support the delivery of payment and account insights to customers.

Middle and back office for workflow and process improvements:

AI and advanced analytics are already commonplace across large areas of middle and back-office operations for payment processing, but many potential opportunities for further enhancement still exist:

- **Payment processing optimization:** One use case for machine learning is to further automate the payments repair process. AI technologies to improve transaction routing is another important opportunity. A third would be in message translation or conversion between older standards and ISO formats, e.g. to translate unstructured name and address in ISO 8583 messages to structured ISO 20022 messages.
- **Risk and fraud:** As with the use case for payments repair, applying AI technologies in transaction screening can reduce the need for manual intervention.

- **Product and service enhancements:** Personalizing services is a huge area AI can help.

Cross-functional capabilities to increasing agility and efficiency

AI technologies can also be used to support several cross-functional use cases within a bank:

- **Using gen AI to support code generation:** This is a huge opportunity for institutions to increase developer efficiency.
- **Risk, compliance, and reporting:** Technologies such as machine learning, NLP, computer vision, and gen AI can all be deployed in various combinations to support processes such as risk management and regulatory reporting.

Dive into Celent's report to learn more about the use cases above and understand how you can harnessing the benefits of AI in payments.



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, Dodd-Frank, PSD2, or various other 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.

Recent research highlights that generative AI not only identifies regulatory obligations and gaps but also provides accurate code recommendations. For example, generative AI can help a bank ingest all applicable regulatory requirements and internal controls, quickly alerting it to any discrepancies. By harnessing its comprehensive and sophisticated capabilities, generative AI can significantly reduce manual effort and minimize errors, all while strictly adhering to factual data.

It then determines an organization's obligations and evaluates its compliance. If obligations are unmet, AI maps them to specific controls or generates new ones. As AI models continue to train, their accuracy

in identifying gaps and interpreting new regulatory requirements improves. Eventually, generative AI may even write new controls, driving greater efficiency.

This seamless process is facilitated by using a multimodal database like MongoDB, which addresses the common issue of documents being scattered and lacking proper metadata and documentation. MongoDB offers several key differentiating features:

- **Converged operational and AI data storage:** MongoDB allows organizations to keep all their operational, unstructured, and vector data in one place, eliminating data silos and making it easier to access and manage.

- **Efficient data handling:** MongoDB simplifies working with data at scale. Centralizing data reduces manual effort, leading to quicker and more accurate compliance processes.
- **Enhanced search and generative AI capabilities:** MongoDB facilitates easy searching through documents and metadata, crucial for quickly identifying regulatory requirements and checking them against existing controls. Utilizing RAG capabilities with MongoDB provides better context to LLMs for generating code suggestions.

Below is an example of a conceptual workflow using MongoDB for code generation:

1. Regulatory requirements: Gather relevant regulatory requirements, determine the specific obligations from these regulations, and condense the regulatory text into concise summaries.

2. Translate to controls: Convert these summaries into actionable controls. *In this context, **control** is a specific measure often implemented through code that ensures compliance with regulatory requirements by automating necessary actions, such as enforcing multi-factor authentication (MFA) for certain transactions.*

3. Convert into executable code:

- Check the controls against the organization's existing internal controls, identify gaps, and create new controls where needed.
- Convert controls into executable code.

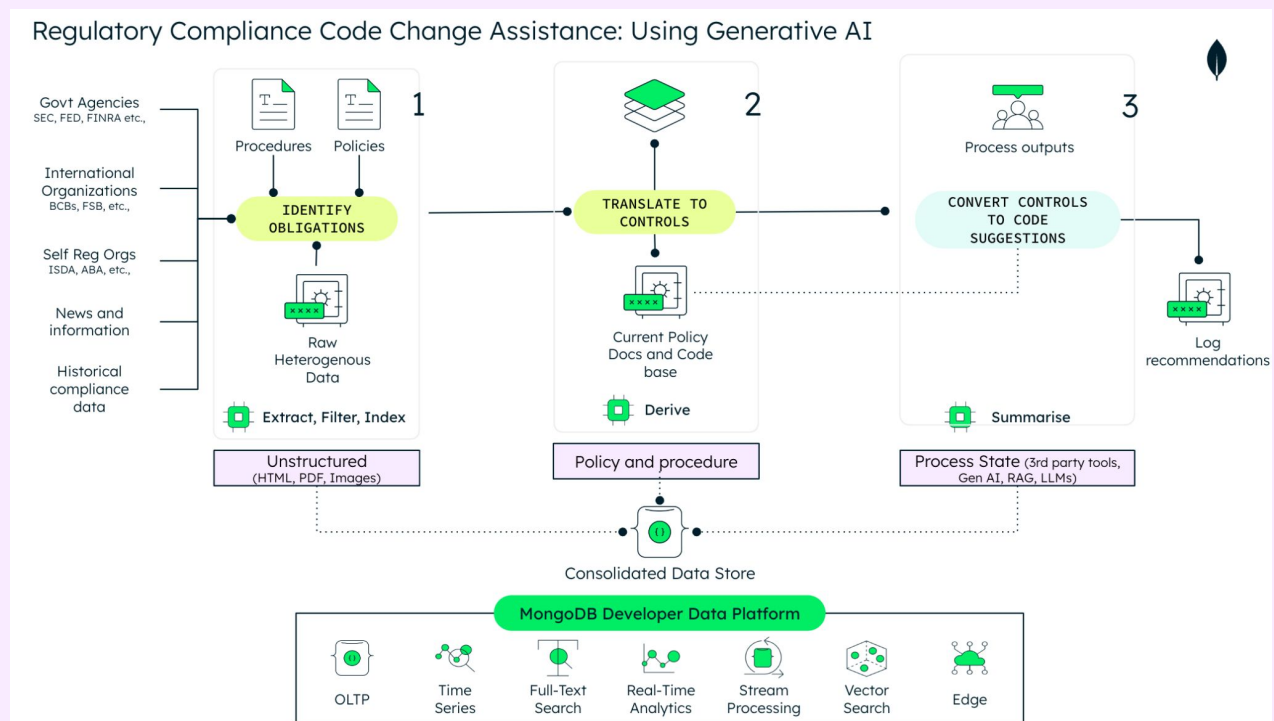


Figure 38: Automated compliance code adaptation with AI

In the financial services section, we emphasized the importance of regulatory compliance by referencing several key regulations such as AML, Basel III, PCI, and GDPR. To demonstrate the practical application of code generation, we used PSD2 as an example and engaged ChatGPT to see if this approach is feasible. Below, you will find the ChatGPT prompts and responses.

Step 1: Real PSD2 clause

Banks must apply strong customer authentication (SCA) for electronic payments initiated by the payer, unless an exemption applies.

Step 2: Control

Ensure that any electronic payment initiated by the payer requires multi-factor authentication (MFA), unless an exemption applies.

Step 3: Code suggestion

Using MongoDB and generative AI, the process might look like this:

1. **Identify payment transactions:** Store all electronic payment transactions in MongoDB.
2. **Check transaction type:** Create a control to check if the transaction is initiated by the payer.
3. **Implement MFA:** Generate code to enforce multi-factor authentication (MFA) for applicable transactions.

The full example is more detailed

please provide an example of converting PSD2 regulation to code



Certainly! Let's walk through an example of converting a specific PSD2 regulatory requirement into code. We'll focus on the Strong Customer Authentication (SCA) requirement, which mandates the use of multi-factor authentication (MFA) for online payments.

Figure 39: ChatGPT 4.0 Example of PSD2 regulation converted into code

Below is a simple example of what the generated code for MFA might look like.

```
from pymongo import MongoClient
from datetime import datetime

client = MongoClient("mongodb://localhost:27017/")
db = client["bank"]
transactions = db["transactions"]

# Function to enforce Strong Customer Authentication
def enforce_sca(transaction):
    if transaction["payer_initiated"] and not transaction["exemption"]:
        # Example MFA enforcement (pseudo-code)
        send_mfa_request(transaction["user_id"])
        return "MFA required"
    return "MFA not required"

# Insert a new transaction and enforce SCA
new_transaction = {
    "transaction_id": "78910",
```

Figure 40: MongoDB setup for multi-factor authentication

```
    "user_id": "user123",
    "amount": 50, # Amount in euros
    "payer_initiated": True,
    "exemption": False,
    "transaction_time": datetime.utcnow()
}

# Insert transaction into the database
transactions.insert_one(new_transaction)

# Enforce SCA for the new transaction
result = enforce_sca(new_transaction)
print(result)

# Example function to send MFA request (pseudo-code)
def send_mfa_request(user_id):
    print(f"Sending MFA request to user {user_id}")
```

Figure 41: Transaction handling and strong customer authentication (SCA) enforcement

Other Notable Use Cases



Below are a few more examples of where AI can be leveraged in financial operations:

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

Algorithmic trading: AI algorithms can analyze public market data and execute trades at speed, optimizing trading strategies.

Generating financial reports: AI can analyze financial data including transactions, invoices, and account statements to automate the process of generating reports. By using AI and ML techniques, relevant information can be extracted where required.

These examples highlight several avenues for integrating AI within financial institutions. Embracing AI in financial applications promises enhanced risk management, operational efficiency, and superior customer experiences.

Therefore, it is essential for financial institutions to grasp the profound technological implications, scale, and intricacies associated with AI, particularly in crafting a generative AI strategy. Adopting a strategic and holistic approach that addresses technological, data, ethical, and organizational dimensions is imperative for navigating this transformative landscape effectively.



Bendigo and Adelaide Bank uses gen AI and MongoDB to modernize applications

Bendigo and Adelaide Bank is a leading Australian financial institution with over two million customers and 7,000 employees. By prioritizing digital transformation, the bank successfully migrated 32 banking applications to AWS in just 30 days, leveraging MongoDB to accelerate development and deliver enhanced services to its customer base.

Contact Information



Wei You Pan

Global Director, Financial
Services Industry
weiyou.pan@mongodb.com



Luis Pazmino Diaz

Principal, Financial
Services Industry
luis.pazmino@mongodb.com



Peyman Parsi

Senior Principal, Financial
Services Industry
peyman.parsi@mongodb.com

FOR MORE INFORMATION AND RESOURCES

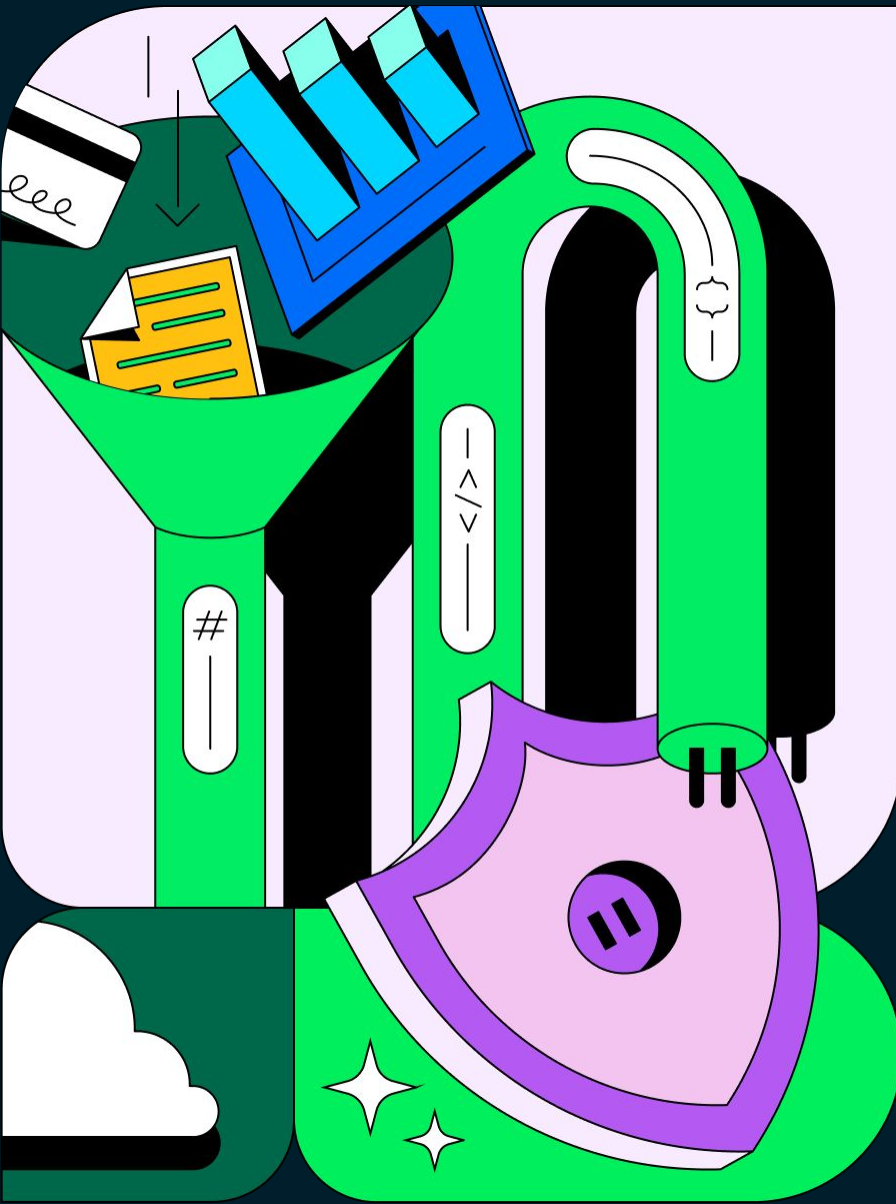
Visit MongoDB for Financial Services





Atlas for Industries

Insurance



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

Architecting with AI for the Insurance Industry



Delivering business value

When architecting for AI-augmented application workflows, it's important to keep your overall business objectives in mind. How will what you do help the insurance business you support? What problem are you solving?

IT portfolio managers who may be driving platform consolidation and modernization agendas inside of your organization may be approaching the space from a technology-first lens, evaluating one platform vs. another in terms of technical capabilities and cost. Data science and analytics folks may be wishing to test or prove data hypotheses. Business product owners may have priorities in terms of raising the bar when it comes to your business applications and their features, hoping to augment their data processing capabilities with the addition of AI.

For you, as the architect or CxO, your charge is to help the organization leverage AI to make meaningful strides in terms of business outcomes, and that means navigating a number of dimensions, not limited to:

- **Business objectives:** A clear understanding of the organization's goals
- **Data and workflows:** Knowledge of business processes and supporting data
- **AI application:** Applying AI to streamline data-intensive tasks

Our end goal is to be able to gather, understand, interact with, and generate data faster by applying the right AI capabilities to the right points in the organization.

Claim as an example

Common organizational goals for an insurer may include:

- **Operational excellence:** Emphasizing efficiency and effectiveness in operations to maximize returns and reduce waste
- **Customer centricity:** Prioritizing investments that improve customer satisfaction and engagement

How well your organization can process and resolve claims, for example, directly impacts the objectives above. Achieving this is directly coupled to how quickly, efficiently, and accurately we can process the data found within claim-handling application workflows.

Claim handling provides us great opportunities to exploit AI in order to accelerate data processing hotspots, so that the organization can return meaningful value from the technology investment.

So what data in a claim-handling workflow is hard to work with? Unstructured data sources such as damage photos, accident forms and reports, claim handler notes, traffic camera videos, and claim-handling guidelines and recommendations. Are any or all of these data sources cumbersome to work with for your employees? Do they need to open and read forms, examine and interpret images, and distill and write case-file notes before your claims can move forward in the handling lifecycle? If a catastrophic event strikes, do you have the staff to handle a massive, sudden influx of claims, all with these types of unstructured data?

Augmenting with AI

Architects are the ones tasked with deciding what type of AI is best suited to solve for insurance data processing needs.

Machine learning models trained on historical data can be used to make predictions and decisions within a business workflow, in lieu of a person.

Generative AI and large language models, however, give us core NLP (natural language processing) capabilities that are particularly well-suited to augmenting data processing abilities. These capabilities, as applied to the claim-handling workflow, include:

Entity extraction to help interrogate and retrieve relevant information from an unstructured source such as a PDF guideline, or a large body of text found within an accident form.

Text and image classification enables claim handlers to determine types of damage or characteristics found on a damage photo.

Text summarization speeds up the synthesis of large bodies of text, or across many sources.

Text generation helps generate case files and give succinct instruction to workers based on larger bodies of guidelines, for example.

Interactive chat makes the previous or additional existing information available more quickly to both your employees and your customers.

Architecting for applications

Your software delivery teams and the applications they support may be divided up by agile delivery domains. It's inside of those domains and applications that you'll need to apply AI in order to effectively move the needle on organizational and process outcomes. In short, your AI belongs in your applications.

The data that supports those applications resides in operational data stores. If we want both our applications and our AI to work with real-time data, it should be accessible within the same underlying data store. What serves our apps should also serve our AI.

Managing operational structured and unstructured data:

Your application architecture needs to provide a way to store, serve, and update structured data as part of your workflow and integrate the unstructured data. This may come in the form of both the raw data (PDF's, images, notes), and the vector-encoding representation of it. You'll want the vector-encoding data to be as close as possible to the structured data that already may exist in your application. Several reasons why include:

- **Performance:** Providing application-level SLA (service level agreements) utilizing efficient server compute
- **Security:** Providing consistent app-tier level security controls to determine who can access what data.
- **Ease of application delivery and maintenance:** The fewer components and dependencies, the lower the complexity and cost of building, deploying, and maintaining effective software solutions that leverage AI.

Applying AI to the claim-handling workflow and delivering better business outcomes

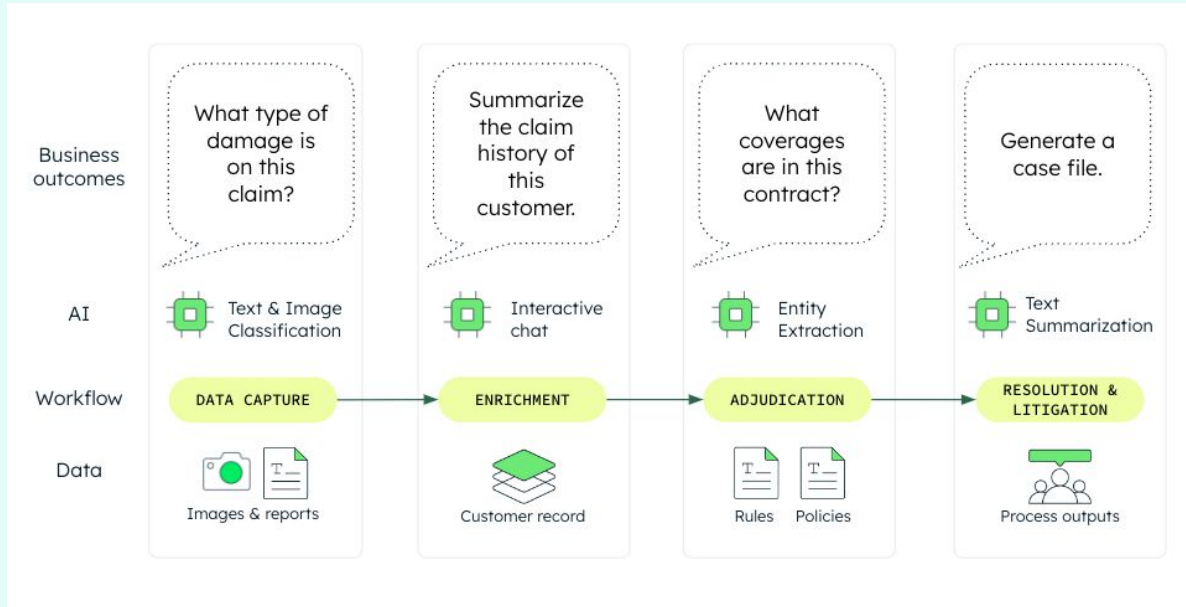


Figure 42: AI use cases aligned to the claim-handling workflow

The above diagram shows practical use of core NLP capabilities that include text and image classification, interactive chat, entity extraction, and text summarization. When applied to a claim-handling workflow,

for example, these capabilities can reduce data hotspots, resulting in lower processing times and costs, and improved customer experiences.

Before AI can transform our organizations, we first have to bring it into our applications

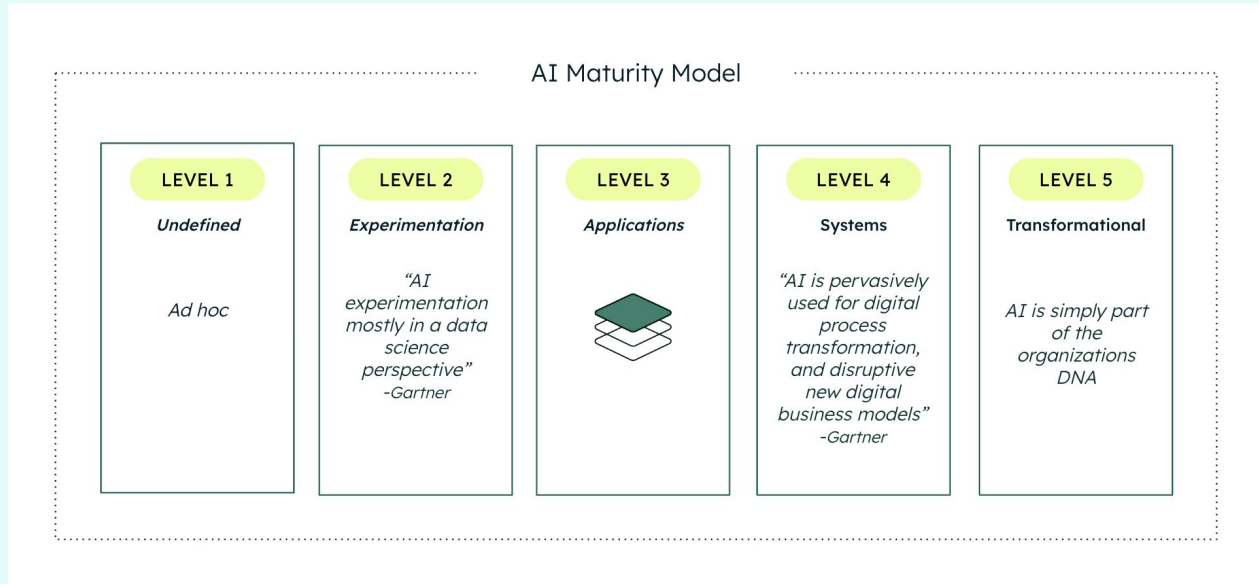


Figure 43: AI Maturity levels from undefined to transformational

The above diagram describes the different stages of AI adoption within enterprises, from early interest to a pervasive and structural integration into processes and decisions. Many organizations struggle with moving out of Level 2, analytics experimentation, to Level 3, deploying AI features within business applications, in order to deliver meaningful business value and outcomes.

The three layers of generative AI

Gen AI applications can be separated into three major layers:

Compute and AI models: Underlying processing power plus foundational and embedding models.

Tooling to fine tune and build applications:

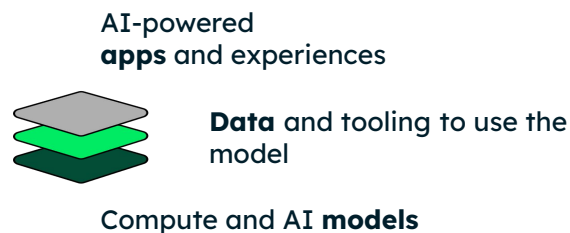
Tools that provide foundational models with context by feeding them proprietary data.

AI-powered apps and experiences: The interface and experiences that end-users interact with, and app frameworks that simplify the process of building AI experiences.

Foundational models are extremely powerful, but being trained on public datasets, they lack the domain knowledge and data context needed to adequately support enterprise applications.

This is where the second layer comes into play: the data and tooling that will glue everything together and enable generative AI-enhanced applications to be fully operational, moving your organization from the experimentation level of Level 2 into full-fledged production apps of Level 3 maturity.

MongoDB's operational database stores proprietary structured and vector data, and makes it available to LLMs when an application makes a request, effectively providing the context necessary for a foundational model to answer questions beyond its initial knowledge boundary.



Building applications that leverage AI

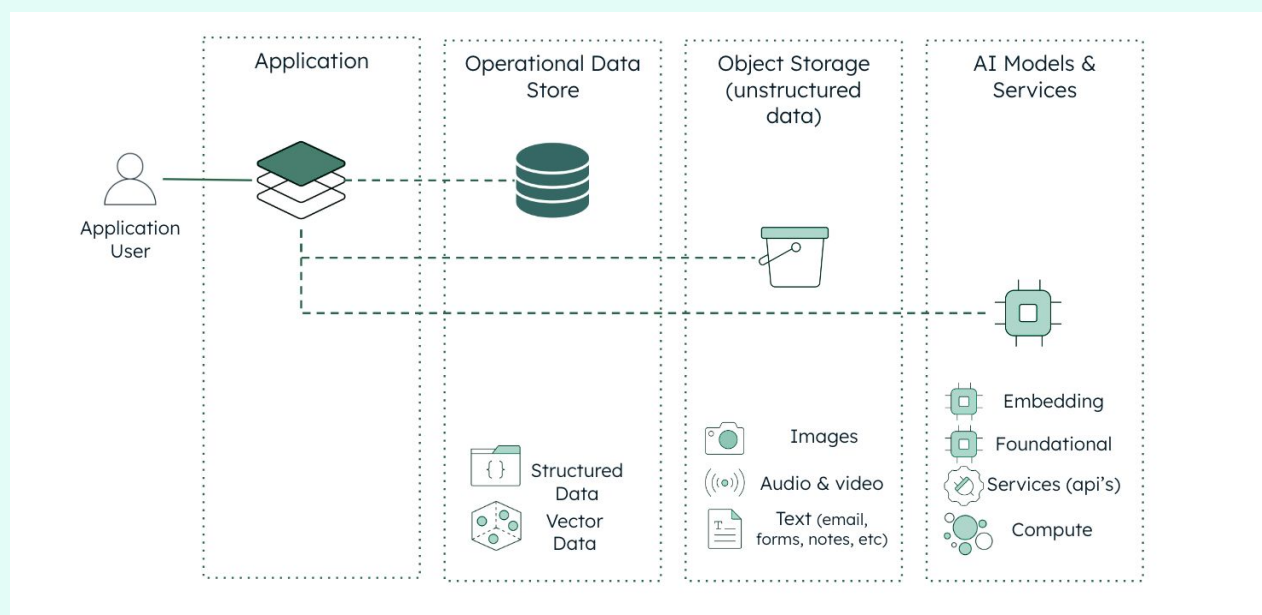


Figure 44: Applications with AI

The above diagram illustrates in greater detail the three layers of gen AI, showing the structured data flowing from the application into the operational database, raw unstructured data managed in object storage, accessible by applications where it can be

processed. Part of the processing includes the vectorization and subsequent persistence of those vectors in the operational data store, where it can be readily accessed by the application.

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 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 policies 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 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 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, supports modern 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 make sense of a trove of data faster and 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 generational 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

Claim 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 analyze more data more effectively and in less time.

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 and 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 virtual assistants is that they can free up your call center staff to work on more complex and high-touch cases. Handling routine inquiries can now extend to more complex scenarios than ever thanks to the power of vector-embedded content and large language models.

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 faster and significantly reduce errors or incomplete information.

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.

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.

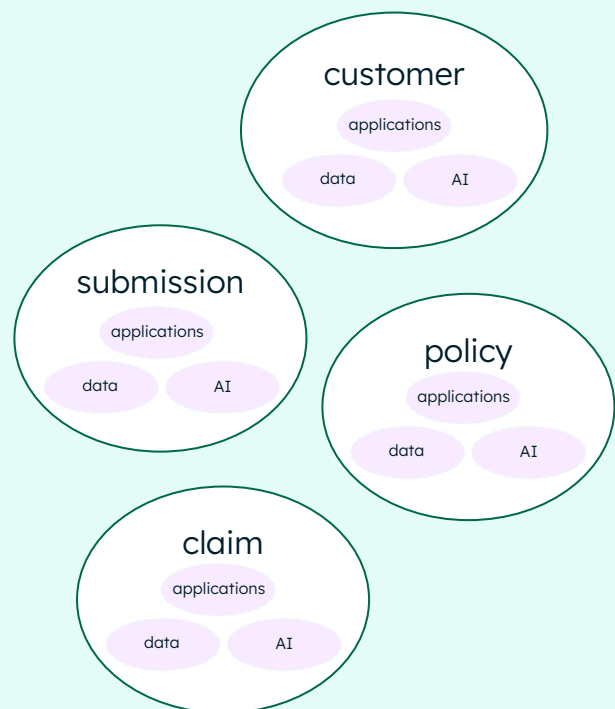
Domain-Driven AI

Insurance enterprises all over the world have embraced domain-driven design in addition to aligned software delivery teams and the applications they support by core processing domains. With a shift to microservices and event-streaming, AI capabilities can now supercharge this architecture and accelerate the ability to interact with and serve up real-time data. Domain-based operational data stores help segment and enable stronger data lineage, data quality, and data governance for more truthful AI interactions.

Working together: Applications, data & AI

Core insurance business capabilities organized and deployed by domain and domain capabilities augmented by AI:

- Domain- and task-specific AI, enhancing key application workflows and data processing hotspots
- Operational and vector data stored as close to applications as possible for:
 - Real-time context
 - Performance
 - Security & privacy
 - Domain agility
- APIs and events for efficient processing within and across domains.



Examples of Domain-Specific AI



Ferret.ai builds trust with AI & MongoDB Atlas, slashes costs 30%

[Ferret.ai](#) helps companies and individuals build trust by providing real-time insights on people and businesses using AI. Powered by MongoDB Atlas, Ferret analyzes vast amounts of public data to identify risks and opportunities, offering a comprehensive view of relationships.

“Through Atlas, it’s available as a fully managed service with best practices baked in. That frees developers and data scientists from the responsibilities of running a database so they can focus their efforts on app and AI innovation”

[Learn more](#)

AI Basseri

CTO and head of Engineering at Ferret.ai



VISO TRUST: Accelerating third-party risk assessment with AI

[VISO TRUST](#) is an AI-powered third-party cyber risk and trust platform that enables any company to access actionable vendor security information in minutes. By leveraging MongoDB Atlas for efficient data storage and retrieval, VISO TRUST’s advanced AI models can process vast amounts of security information, delivering actionable insights in minutes. This streamlined approach significantly reduces assessment time

and enhances overall cybersecurity posture. The platform uses sophisticated models and retrieval-augmented generation (RAG) techniques to provide highly accurate and contextually relevant intelligence. This not only accelerates decision-making but also ensures that companies are equipped with the most precise risk evaluations available. Assessments generated from analysis are 80% faster than manual methods with no loss in accuracy

[Learn more](#)

Solution Library References

The [Solutions Library](#) is a curated collection of tailored solutions, offering ready-made templates, best practices, and detailed guides to help developers quickly start projects. With industry-specific examples and sample code, the library enables insurance companies to efficiently implement common patterns, leverage MongoDB's features effectively, and ensure compliance with industry standards. Here are some examples of solutions for the insurance industry.

Claim management using LLMs and vector search for RAG

By converting claim data into vector embeddings, MongoDB's Atlas Vector Search accelerates information retrieval, making it quicker and easier to find relevant details. Large language models (LLMs) then analyze these embeddings to extract valuable

insights and context, optimizing claim processing. This combined approach enhances accuracy, efficiency, and overall claims management.

[Learn more](#)

Ask Leafy a Question

Ask


Suggested Questions:

Show me claims related to tire damage and summarise the coverages

For adverse weather related claims, what is the average loss amount?


Based on the given information, the average loss amount for adverse weather related claims is \$1539.

References




Customer ID:	Claim Date:	Loss Amount:
c110	2024-02-11	\$807

Claim Description:
A sudden hailstorm damaged vehicles on the road, breaking windshields and causing multiple accidents.



Customer ID:	Claim Date:	Loss Amount:
c100	2023-05-20	\$2591

Claim Description:
During a heavy thunderstorm, a vehicle lost traction and slid into a guardrail on the highway, causing a multi-car pileup.



Customer ID:	Claim Date:	Loss Amount:
c112	2023-09-08	\$1219

Claim Description:
At night, a vehicle hit a pothole at high speed, damaging its front axle and causing it to veer off the road.

Figure 45: Atlas Vector Search provides answers to user questions about insurance claims, including calculations and detailed claim examples

AI-enhanced claim adjustment for auto insurance

By leveraging AI and vector image search, this solution **automates auto insurance claim adjustments**. Accident photos are compared to a database of past claims, significantly accelerating damage estimates.

Centralized data management in MongoDB Atlas enhances efficiency and consistency throughout the claims process.

[Learn more](#)

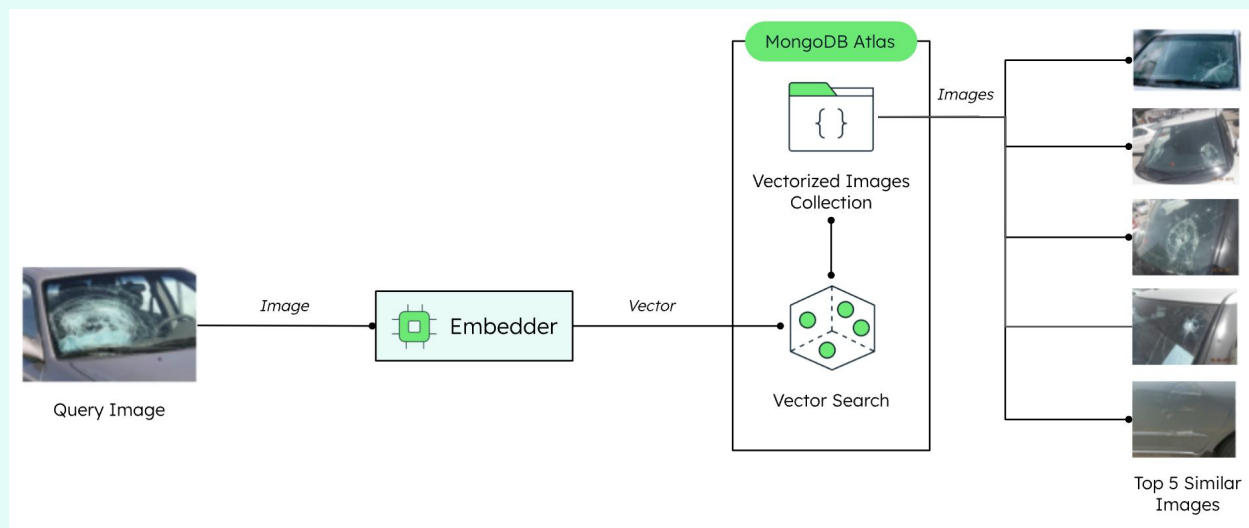


Figure 46: An image similarity query is performed, and the 5 top similar images are returned

Build a PDF search application with vector search and LLMs

PDFs are hard to search, making it tough for insurance workers to find what they need quickly. We're solving this by converting PDFs into a searchable format using tools like

Superduper, allowing users to quickly retrieve information and streamline insurance work.

[Learn more](#)

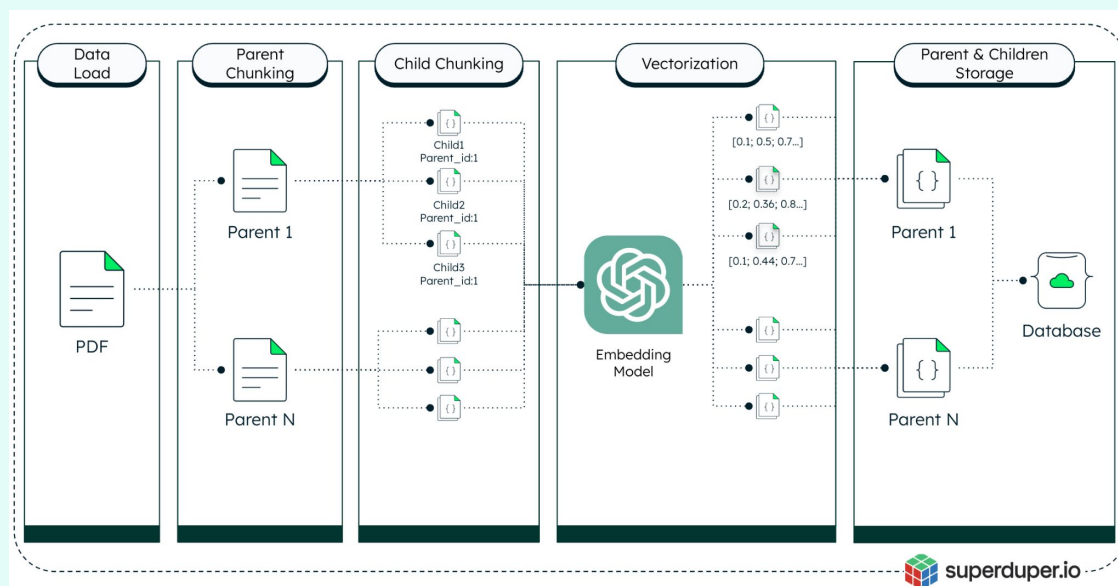


Figure 47: PDF chunking, embedding creation, and storage, orchestrated with Superduper

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

FOR MORE INFORMATION AND RESOURCES

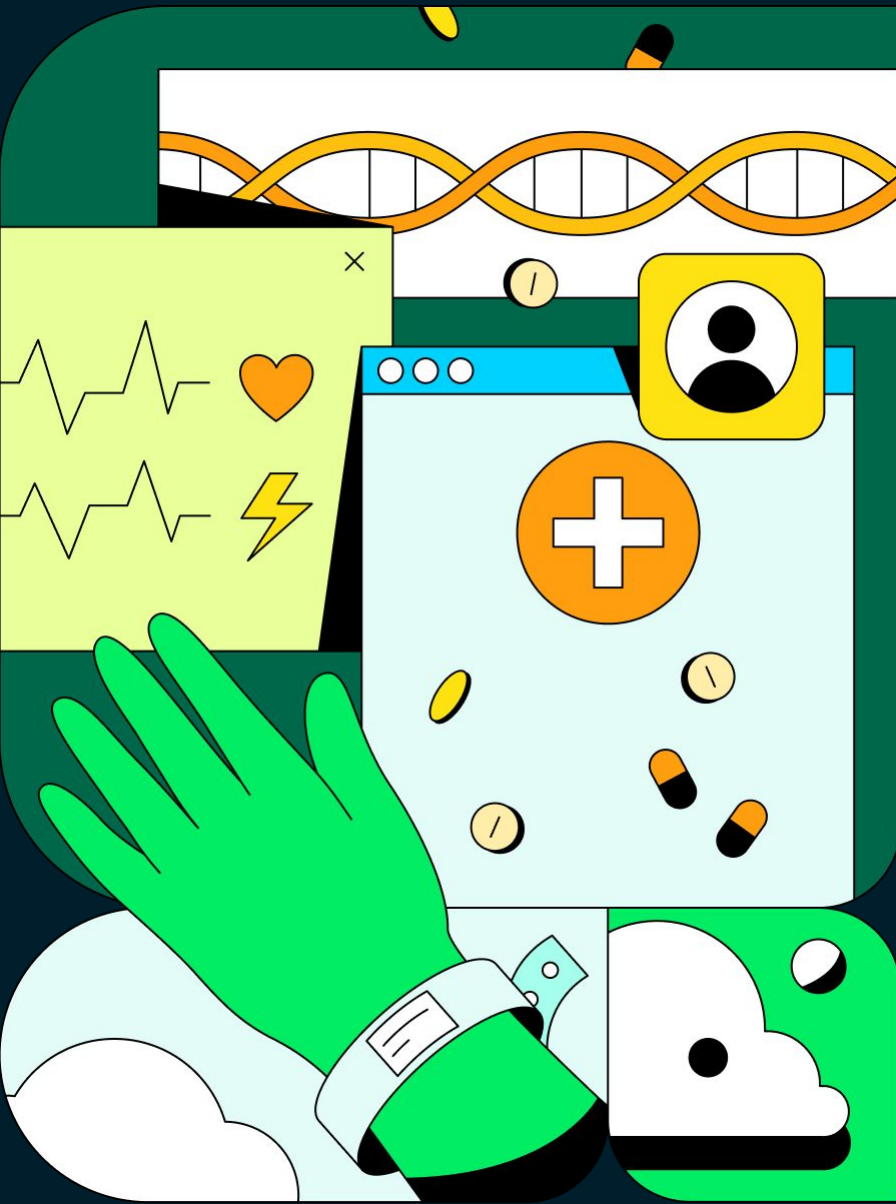
Visit MongoDB Atlas for Insurance





Atlas for Industries

Healthcare and Life Sciences



AI is transforming healthcare by enhancing decision-making, improving patient experiences, and boosting operational efficiency.

In life sciences, AI accelerates research, drives drug discovery, and personalizes treatments, leading to breakthroughs in innovation and tailored medicine.

Transformative Potential of Generative AI in Healthcare and Life Sciences



The healthcare and life sciences sectors have witnessed a significant surge in the application of artificial intelligence (AI) and machine learning (ML) over recent years. These technologies have been increasingly integrated into various aspects of healthcare, from diagnostics and treatment planning to operational efficiencies and patient engagement. The acceleration of AI and ML adoption is evident in the exponential growth of FDA-approved algorithms, reflecting a broader acceptance and reliance on these technologies to enhance healthcare delivery.

However, the advent of generative AI marks a new frontier in this evolution, bringing the potential to disrupt and revolutionize even more areas of the industry. Unlike traditional AI, which primarily focuses on analyzing and predicting, gen AI can autonomously read and interpret multimodal data, and generate new content, offering unprecedented possibilities in automation, personalized medicine, drug discovery, and beyond.

The challenge: Administrative burden and clinician burnout

Healthcare professionals have long been the cornerstone of patient care, responsible for inputting vast amounts of information into electronic health records (EHRs) with the promise that it would lead to improved patient outcomes. However, many clinicians feel they have yet to see the tangible benefits of these systems. Studies reveal that clinicians spend twice as much time on administrative tasks as they do in direct contact with patients.

Additionally, 57% of healthcare providers report that excessive documentation contributes to clinician burnout. This administrative overload not only affects the well-being of healthcare providers but also has a negative impact on patient care, with two out of three patients reporting a lack of empathy from their healthcare providers due to these constraints.

The solution: Leveraging gen AI to alleviate administrative burden

Gen AI has the potential to dramatically reduce the administrative workload on healthcare professionals, allowing them to focus more on patient care. Technologies such as ambient listening, combined with advanced patient records, could automate much of the documentation process, reducing the time clinicians spend on EHRs. By harnessing these capabilities, healthcare providers can reclaim valuable time, improve the quality of patient interactions, and ultimately enhance overall patient outcomes.

This technology could also facilitate the summarization of patient records and present them in visually rich interfaces tailored to the specific needs of each organization, specialty, and even individual healthcare providers. Gen AI can also assist in the documentation and analysis of insights from medical imaging and lab results, as well as automate prior authorization processes to enhance efficiency for both providers and payers. These advancements would significantly lighten the workload and reduce the burden on clinicians, allowing them to dedicate more time and energy to patient care.

Transforming patient communication and engagement

Gen AI is not only transformative for healthcare providers but also holds promise for enhancing patient communication and engagement. By offering personalized and customized communication strategies, gen AI can help create more meaningful patient interactions, build customized patient journeys, and promote preventive care. This increased engagement could lead to better patient outcomes and higher satisfaction rates, as patients feel more connected and understood by their healthcare providers.

Impact on life sciences: Accelerating the medication lifecycle

In the life sciences sector, gen AI is poised to impact the entire medication lifecycle, from discovery to post-marketing surveillance. In the realm of basic research, gen AI has shown success in simulating new molecules and interactions, providing researchers with novel ways to interact with vast amounts of scientific literature. This could accelerate the drug discovery process, leading to faster development of new therapies.

During clinical trials, gen AI can improve the accuracy and speed of developing clinical

study reports (CSRs) and other essential documents. By streamlining these processes, gen AI can reduce the time required to bring new medications to market, ultimately benefiting patients by providing quicker access to new treatments.

In the post-marketing phase, AI can enhance surveillance efforts by monitoring real-world data for adverse effects and other key indicators, ensuring that medications remain safe and effective after they have been approved.

Prioritization

With the rapid progression of AI, decision makers are now faced with the challenge of prioritizing these numerous opportunities. The possibilities are vast, but the path forward requires careful consideration and strategic planning. Keep reading to explore how industry leaders are already applying these groundbreaking technologies and discovering offering insights that can help navigate this complex and exciting landscape.



Figure 48: [McKinsey](#): Areas believed to benefit from generative AI (% of respondents)

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 interactions, 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 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 ML algorithms, natural language processing (NLP), 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 of which can reduce patient wait times. 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 reducing 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.

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 Atlas Vector Search can power search capabilities within patient portals, allowing patients to easily find relevant information and resources, thereby improving self-service.

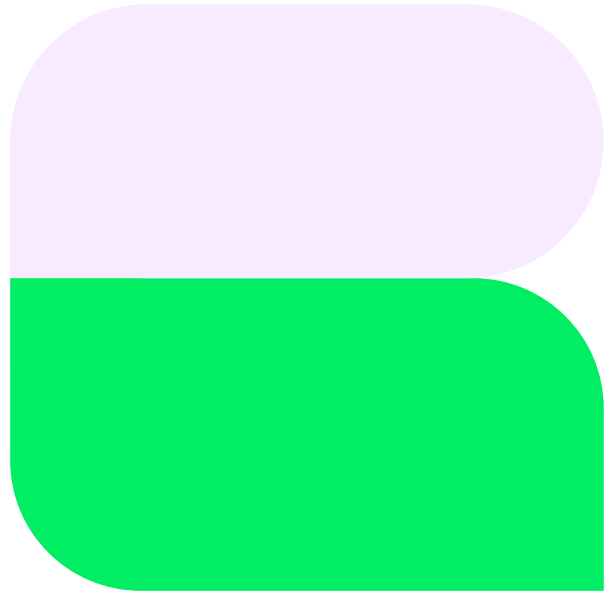
Enhanced Clinical Decision Making

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 data 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 developer data platform, is uniquely positioned to support the complex data needs of healthcare decision-makers. 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, 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 Atlas Vector Search further enriches decision-making processes by enabling semantic search across the database. This integrated approach enables the application of prefilters 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.



Enhancing Medical Imaging with Generative AI

Generative AI offers a transformative solution by automating the extraction, analysis, and summarization of information from medical images. By leveraging advanced NLP, embeddings, and ML techniques, generative AI can rapidly analyze large datasets, identify key insights, and generate accurate, comprehensive summaries. This significantly reduces the time radiologists spend on manual data review, allowing them to focus more on patient care and decision-making.

For instance, AI-powered tools can be integrated into radiology workflows to automatically generate impressions and diagnostic summaries based on imaging data. These tools not only streamline the reporting process but also enhance the accuracy of diagnostics by cross-referencing findings with historical data and known medical conditions. Moreover, chat-based AI assistants can provide real-time support to clinicians by answering queries and offering contextually relevant information derived from patient records and imaging results.

MongoDB's role in supporting AI applications

MongoDB plays a crucial role in enabling these AI-driven applications. By providing a flexible and scalable database architecture, MongoDB allows for the efficient storage and retrieval of vast amounts of unstructured medical data,

including reports, and metadata from DICOM files. Its support for vector search capabilities is essential for implementing retrieval-augmented generation (RAG) approaches, which enhance the accuracy and relevance of AI-generated summaries.

Medical visual question answering (MVQA)

Building on the broader capabilities of generative AI in radiology, MVQA specifically targets the intersection of medical imaging, clinical reporting, and NLP. In this sub-use case, generative AI is utilized to not only generate summaries but also to directly answer clinician-specific queries regarding medical images and reports.

As an example, a clinician might ask, "What are the signs of pulmonary embolism in this CT scan?" The MVQA system would analyze the CT image, cross-reference it with the clinical report, and provide a precise answer such as, "Presence of filling defects in the pulmonary arteries consistent with pulmonary embolism."

This application significantly enhances the comprehension of imaging reports, providing concise and clinically relevant answers, which improves diagnostic accuracy and speeds up decision-making processes.

MVQA combined with vector search streamlines diagnostics and supports a more targeted and effective approach to analyzing complex medical images.

Vector embeddings for biopsy image analysis

Another compelling use case involves applying vector embeddings to store and analyze predefined images with suspicious findings, such as those from large biopsy images. Similar to text analysis, where large texts are divided into smaller chunks, large medical images can be segmented into tile sets (e.g., 100x100 pixels).

A vector representation is generated for each tile, capturing its unique characteristics. By leveraging Atlas Vector Search, a specialized viewer tool can

rapidly identify and highlight suspicious tiles within the image. This process allows pathologists to quickly pinpoint areas of concern, significantly enhancing the accuracy and efficiency of their reviews.

Additionally, MongoDB's partnership ecosystem, including collaborations with AI pioneers, accelerates the development and deployment of these cutting-edge solutions in the medical field.

The following figure shows the recommended architecture to implement this use case using MongoDB Atlas.

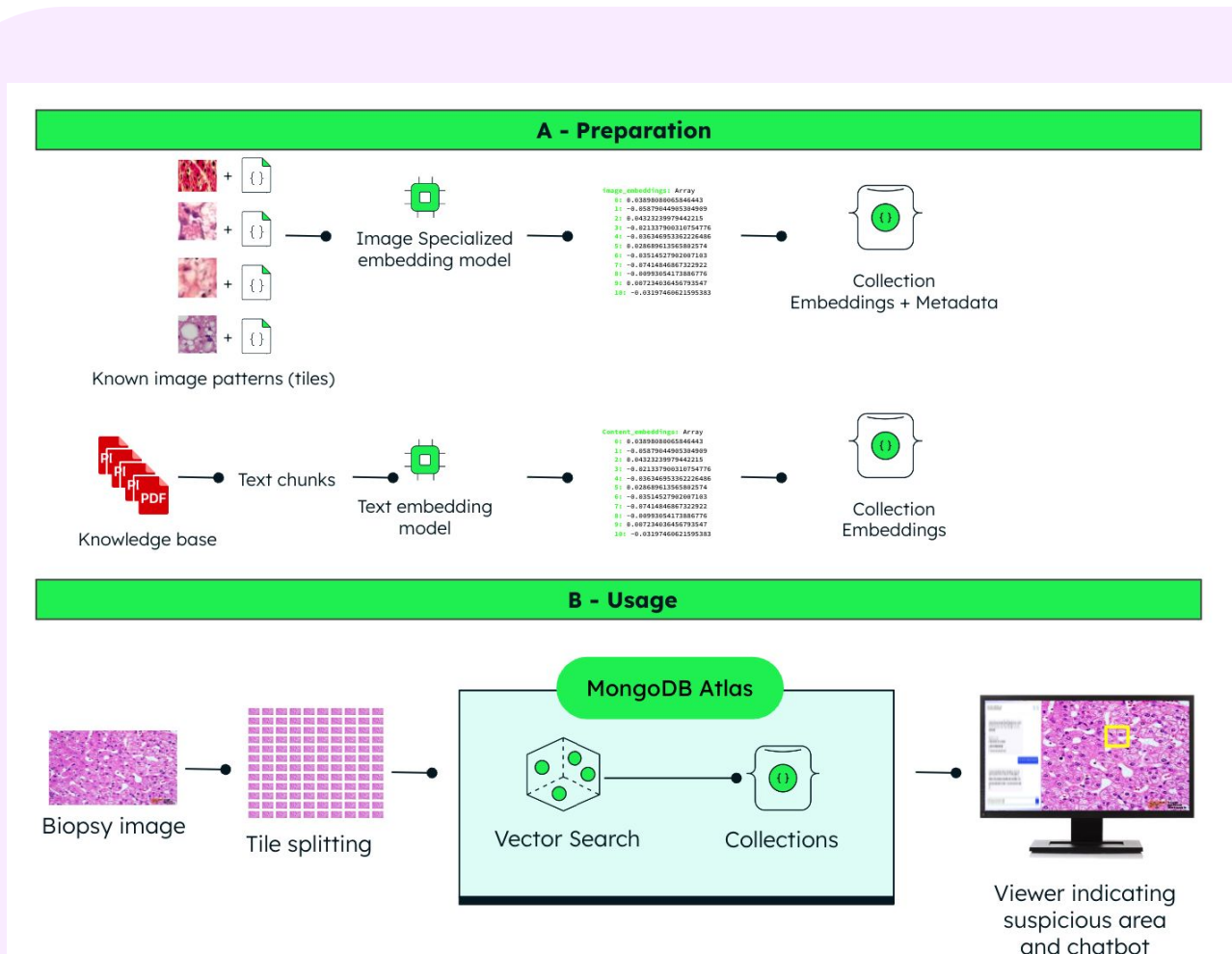


Figure 49: Gen AI in medical imaging with MongoDB Atlas for the detection of suspicious biopsies combined with a knowledge base chatbot

By integrating generative AI into the medical imaging workflow, healthcare providers can achieve significant improvements in diagnostic accuracy, reporting efficiency, and overall patient care. The synergy between MongoDB's robust data management capabilities and advanced AI technologies paves the way for a future where healthcare professionals can deliver faster, more reliable diagnoses, ultimately improving patient outcomes and reducing operational burdens.

This use case demonstrates how generative AI, combined with MongoDB, can revolutionize the way medical imaging data is handled in healthcare, offering a powerful solution to streamline processes, enhance accuracy, and support better patient care.

In life sciences, transforming vast amounts of data into actionable insights is critical throughout the entire medicine lifecycle—from basic research to market. With the rise of precision medicine and genomics, AI is playing a transformative role by tailoring treatments to individual patients based on genetic, environmental, and lifestyle factors. Generative AI accelerates drug discovery, analyzes genomic data for personalized treatment pathways, and optimizes clinical trials.



Revolutionizing Clinical Study Reports (CSRs) with Generative AI and MongoDB

The pharmaceutical industry faces immense pressure to expedite the regulatory approval process for new drugs and therapies. A critical component of this process is the creation of Clinical Study Reports (CSRs), which are comprehensive documents detailing the methodology, execution, and results of clinical trials. Traditionally, compiling a CSR is a labor-intensive task, often requiring several weeks to complete and involving multidisciplinary teams. This prolonged timeline not only delays the introduction of potentially life-saving treatments but also incurs significant costs associated with prolonged R&D cycles.

Challenge

The process of generating CSRs is complex, involving the integration of vast amounts of clinical data, including statistical outputs and detailed narratives. Manual methods are time-consuming and error prone, which can further delay regulatory approvals. Plus, the need for compliance with stringent regulatory standards, such as those set by the FDA and EMA, adds another layer of complexity to the document creation process.

Solution: Generative AI and MongoDB Atlas

Generative AI, integrated with MongoDB Atlas, offers a groundbreaking solution to these challenges by automating the CSR creation process. This approach can reduce the time

required to generate CSRs from weeks to mere minutes, allowing pharmaceutical companies to accelerate their time-to-market for new drugs.

With MongoDB Atlas, companies can leverage a flexible, scalable database environment that supports the dynamic and varied data structures inherent in clinical trials. This flexibility is crucial for managing the diverse data types involved in CSR generation, including text, tables, and complex statistical data. By using generative AI models, companies can automate the drafting of CSRs, producing high-quality, compliant documents that require minimal human intervention.

For example, these AI models can automate the importation and transformation of data tables, generate accurate narratives, and ensure that the final documents meet the compliance standards required by regulatory bodies. MongoDB's Vector Search capabilities further enhance this process by enabling the retrieval of relevant data with high precision, which the AI uses to generate consistent and accurate content.

Extended benefits to any medical writing

The same approach described here for CSRs can provide an end-to-end approach that covers a wide range of regulatory documents, including CSRs, clinical trial narratives (CTNs), summary clinical safety (SCS), and summary clinical efficacy (SCE). This comprehensive coverage ensures that companies can automate much of their regulatory submissions, reducing the risk of human error and speeding up the entire process.

Conclusion

By integrating generative AI with MongoDB Atlas, pharmaceutical companies can transform their approach to generating clinical study reports. This solution offers unparalleled speed, accuracy, and compliance, enabling companies to bring new treatments to market faster while maintaining the highest standards of quality and regulatory adherence. The result is a more efficient drug development process that ultimately benefits patients by accelerating access to innovative therapies.



Novo Nordisk accelerates drug approval with gen AI and MongoDB Atlas

Novo Nordisk, a global leader in healthcare, is transforming how it brings new medications to market using generative AI and MongoDB Atlas. Known for its pioneering work in diabetes care, Novo Nordisk produces 50% of the world's insulin and serves millions of patients worldwide.

NovoScribe: Revolutionizing regulatory submissions

With the introduction of NovoScribe, the company has significantly reduced the time required to generate clinical study reports (CSRs), which is a critical step in the regulatory approval process.

NovoScribe, built on Amazon Bedrock, LangChain, and MongoDB Atlas, has enabled Novo Nordisk to reduce the time to compile CSRs **from 12 weeks to just 10 minutes**.

This innovation is helping Novo Nordisk get new medicines to patients faster, enhancing both the speed and quality of their regulatory submissions. By leveraging MongoDB Atlas's capabilities, NovoScribe automates complex data retrieval and analysis, allowing the company to scale its operations efficiently and securely across multiple cloud platforms.

“We’ve reduced the time taken to create clinical study reports from 12 weeks to 10 minutes, with higher quality outputs and a fraction of the team. In terms of value, each day sooner a medicine gets to market can add around \$15 million in revenue to the company.”

[Learn more](#)

Waheed Jowiya

Digitalisation Strategy Lead at Novo Nordisk

Accelerating Drug Discovery with Generative AI

Current state and challenges

In the pharmaceutical industry, the drug discovery process is a complex, resource-intensive endeavor, often involving extensive experimentation, data integration, and analysis. Traditional methods, which rely heavily on manual research and iterative testing, can be slow and costly, delaying the time to market for new drugs and driving up overall development expenses.

A key challenge in this domain is managing and interpreting **vast volumes of diverse data**. Researchers must sift through enormous datasets of molecular structures, chemical reactions, and historical research to identify promising candidates. The process requires not only identifying molecules with desirable properties such as high efficacy and low toxicity but also predicting their behavior in complex biological systems.

The iterative nature of drug discovery, combined with these data challenges, results in high costs and significant time investment, with no guarantee of success. This underscores the need for more efficient, data-driven approaches to accelerate the discovery process and reduce development costs.

Leveraging advanced technologies to transform drug discovery

The drug discovery process is poised for transformation through the integration of advanced technologies such as vector embeddings, SMILES notation and MongoDB Atlas. By converting both textual information and molecular data into vector representations, researchers can unlock new efficiencies and insights, setting the stage for more effective NLP applications in drug discovery.

At the core of this transformation is the ability to convert diverse data types—ranging from textual descriptions in scientific literature to molecular structures and chemical reactions—into vector representations. These embeddings serve as compact, high-dimensional numerical representations that capture the essential properties and relationships of the data.

- **Molecular embeddings:** Molecules, including those represented by SMILES notation, are transformed into vector embeddings that encapsulate their structural and functional characteristics. This allows for the comparison of molecules based on their properties rather than their chemical structures. As a result, molecules with similar therapeutic potential can be identified even if they differ in their chemical makeup.

- **Reaction embeddings:** Chemical reactions can also be represented as vector embeddings, capturing the transformation of reactants to products. This facilitates the search for reactions that are functionally similar, helping researchers identify alternative synthesis pathways or optimize existing reactions.
- **Text embeddings:** Scientific literature, patents, and other textual data are converted into vector embeddings that represent the semantic content of the text. This enables the system to understand the context and meaning behind complex chemical terms and descriptions, making it easier to link related concepts across different documents.

By embedding all these data types into a unified vector space, researchers can perform powerful similarity searches, clustering, and predictive modeling, significantly enhancing their ability to discover new drug candidates.

MongoDB Atlas as a developer data platform

MongoDB Atlas provides the robust infrastructure needed to store and manage the diverse data types involved in drug discovery. With the ability to handle both structured and unstructured data, MongoDB Atlas supports the integration of molecular structures, textual descriptions, reaction pathways, and their corresponding vector embeddings into a single, scalable database.

This unified data platform allows for seamless querying and retrieval of data. Researchers can access molecular structures, related literature, and embeddings all within the same environment, ensuring that they have the

comprehensive information needed to drive their research forward. MongoDB Atlas supports the use of vector search technologies, enabling efficient and accurate exploration of the chemical space.

Enhanced discovery through Atlas Vector Search and NLP

Atlas Vector search allows researchers to explore chemical and textual data in a way that was previously impossible. When a researcher inputs a molecular structure or a piece of text, the system can rapidly identify other molecules, reactions, or documents that are similar in the vector space.

For instance, a researcher could input the vector representation of a promising molecule, and the system would return a list of similar molecules from the database, ranked by their similarity in the vector space. This enables the identification of potential drug candidates that might not have been considered through traditional search methods.

By leveraging NLP, researchers can run complex queries in plain language, such as “What are the most recent studies on molecules similar to this compound?” or “Show me reactions that convert this functional group to another.” The NLP models interpret these queries, search the vectorized database, and return the most relevant results, whether they are molecular structures, reaction pathways, or related literature.

This capability drastically reduces the time and effort required to find relevant information, allowing researchers to focus more on analysis and decision making rather than data retrieval.

LLMs for hypothesis generation and decision support

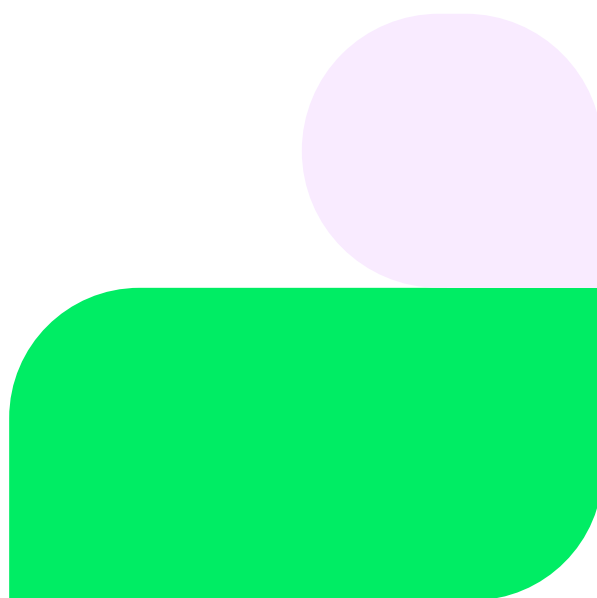
Large language models (LLMs) trained on extensive datasets of chemical and biological information can further enhance the drug discovery process by generating hypotheses, proposing new molecules, and optimizing reactions. Given a specific query or a set of parameters, LLMs can generate novel SMILES strings for potential drug candidates, predict their properties, and suggest optimal synthesis routes.

LLMs also play a crucial role in decision support. By analyzing the vast amount of data stored in MongoDB Atlas, they can provide

insights that might not be immediately apparent, such as identifying unexpected correlations between different molecules or predicting potential side effects based on historical data.

Outcome and benefits

- Time and cost savings: The integration of AI with MongoDB Atlas significantly reduces the time required for drug discovery and reaction optimization, leading to faster time-to-market and lower R&D costs.
- Improved accuracy: AI-driven predictions are more accurate, reducing the need for extensive manual experimentation and iteration.



Cutting-edge recommendation engine leverages advanced language models and vector search technology.

This AI-driven system integrates MongoDB's modern data platform to store and manage complex molecular data, allowing scientists to perform natural language queries for rapid identification of promising molecular structures. The platform also unifies data from various sources, including time series data from PLC devices, enabling faster and more refined molecule discovery processes.

The implementation of Exafluence's solution is expected to reduce the time-to-market for specialty life sciences products by 40%. The enhanced system enables faster and more accurate molecule discovery, streamlining the production of specialty chemicals across various industries. Additionally, the planned introduction of a mobile app will use real-time data to further improve the accuracy of molecular predictions, enhancing research decisions and overall operational efficiency.

[Learn more](#)

The screenshot displays the Exafluence AI-powered platform interface, titled "Planner 1". It shows a workflow for planning the synthesis of DEET.

Question:
Plan and execute the synthesis of any insect repellent, DEET, from search using

Agent:

- Action: WebSearch**
Action Input: DEET insect repellent SMILES
Observation: DEET is a personal insect repellent that has been shown to repel biting flies, biting midges, black flies, chiggers, deer flies, fleas, gnats, horse flies, ...
- Action: RXNPlanner**
Action Input: CC(C)(C)OC(=O)N(C)C
Observation: Input not provided completely. It should include molecule SMILES and user input separated by '||'

Visor:
Route picked by Agent

The route diagram shows the synthesis of DEET starting from diethyl phthalate and ethanol, reacting with sodium hydroxide, then dimethyl sulfate, and finally hydrochloric acid to yield DEET.

Final Answer:
To synthesize DEET, you can follow the following steps: 1. Combine diethyl phthalate (C₁₂H₁₄O₄) and ethanol (C₂H₆O) in a round-bottom flask. 2. Heat the mixture to a temperature of about 80°C and stir until it is homogeneous. 3. Add a solution of sodium hydroxide (NaOH) to the mixture and stir until the reaction is complete. 4. Separate the organic layer from the aqueous layer and discard the aqueous layer. 5. Add a solution of dimethyl sulfate (CH₃SO₃) to the organic layer and stir until the reaction is complete. 6. Separate the organic layer from the aqueous layer and discard the aqueous layer. 7. Add a solution of hydrochloric

Figure 50: Screenshot from Exafluence AI-powered platform showcasing the automated synthesis planning of DEET using integrated web search and reaction planning tools

Other Notable Use Cases



Behavioral health monitoring and sentiment analysis

Generative AI integrated into wearables or apps can monitor conditions like depression and anxiety by analyzing speech, activity levels, physiological signals, and sentiment. It provides early alerts and suggests interventions, enabling proactive care and improving mental health management.

Automating prior authorization, coding, and billing

Generative AI automates prior authorization, coding, and billing by assisting with form completion, accurate record coding, and reviewing insurance policies. This reduces errors, speeds up approvals and reimbursements, and improves efficiency for both providers and payers.

Personalized patient history visualization

Generative AI customizes and summarizes patient history views based on the specific needs of each healthcare professional. By highlighting the most relevant records, results, and treatments, and providing concise summaries, AI streamlines access to critical information.

Personalized treatment pathways in clinical trials

Generative AI analyzes patient data to create personalized treatment plans for clinical trials, predicting responses and optimizing trial designs. This improves efficiency, accelerates drug development, and enhances patient outcomes.

Contact Information



Francesc Mateu Amengual

Healthcare Industry
Solutions Principal

francesc.mateu@mongodb.com

FOR MORE INFORMATION AND RESOURCES

Visit [MongoDB Atlas for Healthcare](#)





AI Leaders in Action

Hear from AI pioneers sharing their expert
insights on the future of AI





Ben Ellencweig is a McKinsey senior partner who leads alliances, acquisitions, and partnerships globally for QuantumBlack, AI by McKinsey. Ben advises clients on harnessing the power of AI to deliver growth and bottom-line impact across sectors.

Quantumblack, McKinsey's AI arm, supports transformation initiatives using the power of technology, technical expertise, and industry experts. QuantumBlack Labs is the R&D and software development arm of QuantumBlack.

Ben Ellencweig,

McKinsey senior partner and global alliance leader for QuantumBlack, AI by McKinsey

Can you share a bit about your career journey and what led you to become a McKinsey Senior Partner and leader in QuantumBlack, AI by McKinsey?

My work at McKinsey goes back nearly 15 years at this point, starting out as an associate after getting my MBA in 2009. During that time, I've worked on a broad range of technical challenges, applying new technologies to industries ranging from automotive to software to media. More recently, I've been focused primarily on AI, and am currently the global leader of alliances and acquisitions for QuantumBlack, AI by McKinsey. I also led our acquisition of Iguazio, a leader in AI and machine learning technology, in 2023.

Before McKinsey, I spent nearly a decade in roles leading the deployment of new technologies in government and military settings. The constant theme throughout my career has been implementing cutting-edge technology in innovative ways that transform large organizations and create significant impact at scale. This is why it makes sense that I've become focused on AI—there's simply no other technology in our lifetime with the same kind of potential to improve society and our lives.

QuantumBlack has been instrumental in leveraging AI across various sectors. Can you highlight some of the most impactful AI projects or initiatives led by QuantumBlack?

QuantumBlack has led some truly incredible initiatives that have transformed major players in a variety of sectors.

- We helped a major pharmaceutical company develop an AI-powered drug discovery platform, which accelerated the identification of promising drug candidates by 60%.
- In manufacturing, QuantumBlack worked with a global automaker to implement an AI system for predictive maintenance that reduced unplanned downtime by 30%.
- We built an AI-driven risk assessment model for a financial services client that improved fraud detection rates by 40%, while reducing false positives.
- In healthcare, our AI solution for optimizing hospital operations led to a 15% reduction in patient wait times and 20% improvement in resource utilization.

There are also some high-profile projects that we can mention more openly, like developing an AI-enabled “dashboard” that led to a world speed record for a Formula E electric car racing team.

I’m especially proud of our ongoing work with One Ocean, a foundation that leads research efforts to support the development of an eco-friendly “blue economy.” Identifying which companies are truly having a positive environmental impact has always been difficult, so we helped One Ocean develop an AI-powered analytics platform that streamlines the process dramatically.



From your experience advising Fortune 500 executives and boards, what are the most common misconceptions about AI in the business world?

There are several! The most common misconception is one that shows up with nearly any emerging, transformative technology: the belief that it can be used as a plug-and-play solution, rather than one that requires strategic integration. Executives often look for immediate ROI from AI, without accounting for the necessary investments in data infrastructure, talent, and change management. I also see a tendency to underestimate the importance of high-quality, well-governed data as the foundation for effective AI.

It’s all too easy to overestimate AI’s capabilities in the short-term, expecting human-level reasoning across all domains, even if it’s working with low-quality or poorly structured data. In terms of gaining long-term value from AI, I’d say the biggest mistake is executives focusing solely on productivity through automation and cost-cutting rather than AI’s potential for business model innovation and value creation.

Given the varying rates of generative AI adoption across industries, what specific industries do you see as leading the charge in terms of deriving maximum value from the technology? What are the key factors driving this disparity?

Right now, technology, financial services, and healthcare are at the forefront when it comes to harnessing the power of gen AI. Tech companies are leveraging it for software development, content creation, and customer service. Financial services firms are using it for personalized financial advice, risk assessment, and fraud detection. Healthcare is seeing significant applications in drug discovery, medical imaging analysis, and clinical decision support.

Although these three sectors are very different, they have a few things in common that have contributed to their early adoption. All three are very data-driven with well-developed technological infrastructure, meaning they have access to the well-structured, high-quality data needed for meaningful AI action, and may not need to make big investments in hardware. Software and finance are also sectors where AI-savvy talent is already abundant. They have clear use cases (mentioned above) that are closely aligned with core business processes. Last but not least, these are competitive sectors that are used to making investments in order to maintain an edge, and stay ahead of the pack.

Data is a critical component of successful AI implementations. What are the unique data challenges and privacy considerations for different industries, and how can organizations effectively address them?

One that comes immediately to mind is healthcare: There's tremendous value to be gained, but you're also dealing with incredibly sensitive patient data that requires robust anonymization and consent management practices. Both healthcare and financial services face strict regulations on data usage, and finance in particular has to worry about cross-border data flows. Retail and e-commerce face similar but less extreme issues, as the more they use data to personalize experiences, the more privacy concerns they bump up against. Regulations like GDPR have already made these concerns broadly understood, but there's still plenty being worked out. Any sector that deals with classified or proprietary information, including manufacturing, government, and defense, must put serious effort into ensuring that data is protected.

Every sector faces issues of data quality and integration to some degree, and very large industries with a wide range of customers (like healthcare,) often struggle trying to manage a variety of datasets with different formats and quality levels. Strong data governance frameworks can go a long way to addressing these issues, and data quality and integration tools are getting better all the time. There's also been rapid improvement in privacy-preserving technologies like federated learning and differential privacy. Ultimately, good data management is a strategic and even cultural issue, and one that's essential for AI success.

How do you envision generative AI reshaping specific industries in the next five to 10 years? What are the key trends and disruptive innovations to watch for?

Predicting the future is always a challenging task, but if I had to guess I think some of the most visible changes will be in healthcare, with AI-assisted drug discovery, personalized treatment, and AI-powered diagnostic tools. You can expect more effective, more targeted medical treatments, and potentially a reduction or at least a leveling of costs, which all have exciting implications for our quality of lives. Several other industries will also start offering hyper-personalized products and services combining text, image and video, from financial services to education to shopping, in ways that we have never experienced before.

I'm quite excited to see how the intersection of AI and blockchain impacts legal services, and how IoT sensors can be leveraged through AI analysis to revolutionize manufacturing, realizing unprecedented levels of efficiency and predictive maintenance.

QuantumBlack emphasizes the integration of hybrid intelligence, combining human expertise with AI capabilities. Can you elaborate on how this approach has been implemented in specific projects or sectors? What benefits have organizations experienced from leveraging this blend of human and artificial intelligence, and how does it enhance decision-making and operational efficiency?

That's true, hybrid intelligence is something QuantumBlack has championed since its beginning and successfully implemented across a wide range of projects and sectors. To give one example, we integrated AI-powered predictive maintenance models with human expert knowledge in a large oil and gas project to optimize equipment maintenance schedules, which reduced downtime by 25% and brought significant cost savings. For a financial services client, we developed an AI system that assists human analysts in detecting complex fraud patterns, improving detection rates by 35% while reducing false positives. In healthcare, we helped one provider with a hybrid system for patient triage: AI provides initial assessments that are then refined by human clinicians, leading to patient routing that's 20% faster, and more accurate as well.

In each of these cases, the human operators welcomed this kind of AI support, seeing it as an extension of the technological tools they already used and a seamless way of augmenting their expertise.

What key skills and qualities do you believe are essential for future leaders in AI?

Given how rapidly the AI field is evolving, and will continue to evolve, leaders will need to be exceptionally flexible and adaptable, and able to learn continuously. This is in addition to other qualifications that might be more obvious, like an understanding of data science and machine learning, and an ability to think strategically in order to identify high-impact AI use cases and align them with business objectives.

Another aspect of AI that demands a unique skillset is its breadth. Because it is a general-purpose technology with the ability to touch almost any kind of human pursuit, AI needs leaders with a global and multidisciplinary mindset who are able to translate technical concepts for stakeholders in non-technical fields. Finally, it should be clear that AI brings tremendous ethical considerations with it, and that anyone seeking to lead here needs to not just respect ethics but do the work of understanding how they play out in real people's lives.





Asaf Somekh is the Founder & CEO at Iguazio (acquired by McKinsey)

Can you share your personal journey in the world of AI that led you to found Iguazio? What were the specific challenges you encountered in the industry that inspired you to create a platform to address them?

In late 2014, after the successful acquisition of Voltaire (which went public on NASDAQ in 2007 and was later on acquired by Mellanox, now a part of NVIDIA) I climbed Mount Kilimanjaro with one of my co-founders. Together, we decided to create Iguazio to address the challenge of operationalizing and de-risking AI. We saw firsthand the challenges that enterprises were struggling with when trying to implement machine learning in the organization. Many models were either unable to reach production or, worse, introduced errors and risks that had a negative impact on the business. These weren't just technical setbacks but rather financial losses, inefficiencies, and missed opportunities for innovation.

ChatGPT galvanized the adoption of AI in 2022, making the challenges of bringing AI from development to production and impact even bigger.

Asaf Somekh,
Founder & CEO
at Iguazio (acquired by
McKinsey)



Iguazio offers an AI platform that enables enterprises to develop, deploy, and manage ML and gen AI applications at scale. It serves as a unified place for data ingestion, storage, processing, model training, and deployment.

In early 2023, after eight years of building and scaling the company, Iguazio was acquired by McKinsey. Today, Iguazio is a part of QuantumBlack, McKinsey's AI arm. Together with McKinsey, we help global organizations, including Fortune 500 companies, with special interest from within the financial industry, implement and scale gen AI responsibly.

Could you explain how Iguazio's platform simplifies the development, deployment, and management of gen AI applications?

At Iguazio, our mission is to bring AI to life, making AI implementation not just possible, but scalable, effective and responsible in live business environments across industries.

We focus on two critical aspects that set our platform apart: operationalization and de-risking.

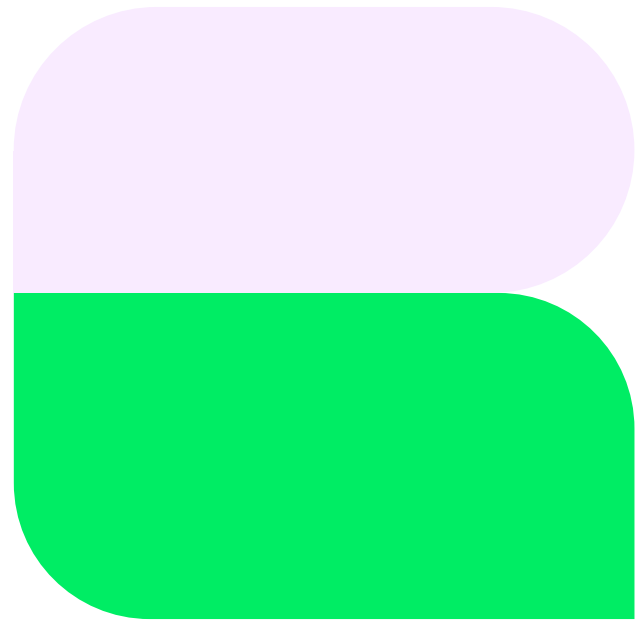
We've built our platform to automate and streamline the entire AI lifecycle, from data ingestion and model development to deployment and continuous management.

This means you can take a generative AI application from a proof of concept to full-scale production quickly and reliably. Our platform handles the heavy lifting by orchestrating four robust processes covering everything from development to production: data management, LLM training and customization, application deployment, and monitoring and governance. We also ensure that your AI applications are highly scalable, with optimized GPU provisioning, and flexibly deployed in any environment—on-premises, multi-cloud, or hybrid.

While generative AI offers incredible opportunities, it also comes with significant risks, such as AI hallucinations, IP infringement, biased outputs, data privacy concerns, and more. Iguazio's platform allows enterprises to easily add guardrails to mitigate these risks across the entire AI pipeline.

Putting this all together is essentially the "AI Factory" concept that we've been evangelizing even before the gen AI hype.

With the growing complexity of AI models and the need for continuous improvement, why do you believe the concept of an AI factory is essential for scaling and maintaining AI systems effectively? How does Iguazio's platform embody this idea to streamline the end-to-end AI lifecycle?



An AI factory operates on the principle of continuous, automated and scalable production—much like a traditional manufacturing factory but for AI applications. The AI factory approach ensures that gen AI and ML models are always up-to-date, performing optimally, do not pose risks, and are aligned with business objectives.

These models are governed with the data sets that were used to train each version, together with the application source code. This is done by applying clients' CI/CD frameworks but with these extra dimensions. This is critical for addressing the AI regulations that are being introduced globally, starting with the EU. The AI Factory also addresses the growing complexity in AI by integrating various processes—such as data ingestion, model development, validation, deployment, and monitoring—into a unified pipeline that reduces friction and inconsistencies. For organizations that want to maintain a competitive edge, this enables them to combine gen AI with traditional AI, rapidly iterate on AI models, and deploy new capabilities in response to changing market conditions or emerging data.

The Iguazio AI platform is built to streamline the end-to-end gen AI and ML lifecycle.

We provide an integrated environment that accelerates the deployment of foundation models, transformers, and application logic in production, while maximizing GPU performance. Comprehensive testing and built-in monitoring and management capabilities ensure the quality and reliability of user-facing applications. We built and maintain open-source AI orchestration framework MLRun, which is at the heart of our platform, for the community to further support these workflows. So enterprises can collaborate efficiently, streamline operations, and accelerate their time-to-market.

Iguazio and MongoDB have teamed up to create a powerful solution for gen AI applications. Could you explain the synergy between the two platforms and the benefits it offers to enterprises?

MongoDB is known for its flexibility, scalability, and ability to handle vast amounts of various types of data—key requirements for building and scaling generative AI applications. On the other hand, Iguazio brings to the table a sophisticated platform that automates and optimizes the AI lifecycle, from data preparation and model training to deployment and continuous monitoring and de-risking. What I also love about this partnership is that both MongoDB and Iguazio have always been focused on scalable production environments.

The integrated MongoDB-Iguazio solution allows customers to streamline data processing and the AI application life cycle. Together, they ensure gen AI applications reach production while eliminating risks, improving performance and enhancing governance.

All your data needs are covered by MongoDB from vector search to a document database to SQL queries, and Iguazio's AI Factory is layered on top.

This partnership and pre-made recipes provide enterprises with the capabilities to operationalize AI with minimal engineering, from prototype to production to monitoring. They can also scale their data while ensuring high performance, reliability, and accuracy. Customers can rest assured with enterprise security and compliance, including encryption, access controls, and compliance monitoring.

In addition, we support hybrid environments and provide unification of all data management needs, like logging, auditing, and more to streamline the process and ensure consistency and faster performance. Finally, we can accommodate diverse use cases, like real-time co-pilots for different professions, virtual agents, call center analysis applications, chatbots, and more. With McKinsey's expertise in strategy and consulting, we can also guide organizations strategically throughout their AI-based transformations.

How does Iguazio's platform support the seamless integration of LLMs and traditional ML models, and what are the key considerations in doing so?

This is a super important question. We see organizations build separate environments and teams for gen AI and traditional AI. Also, we see too many organizations trying to use LLMs for AI applications. These are the wrong approaches. Gen AI with its LLMs will sometimes underperform compared with a standard machine learning model or even simple heuristics. Therefore, the right approach is to build pipelines that can combine the adequate technique in each step of the application.

Iguazio's platform is uniquely designed to support various AI needs, like the integrated use of both LLMs and traditional ML models, all while addressing the complexities of managing and deploying both types of models within a unified framework.

Example use cases include integrating a customer-facing chatbot with a classifier model to check information about customer queries, generating synthetic data for ML model training, or using sentiment analysis for feature engineering. But there are many more.

Using LLMs and traditional ML involves the same MLOps (machine learning operations) practices. This includes ensuring you address automation and coordination of workflows like data preprocessing, model training and deployment, ensuring flexible access, distribution of GPU resources, and computational power when needed, real-time capabilities, monitoring for model performance, and establishing guardrails.

Given the rapid advancements in gen AI, where do you see the most significant opportunities and risks for enterprises over the next few years?

Over the next few years, I see several key areas where businesses can either thrive or face significant challenges, depending on how they navigate this new era. For example, McKinsey & Company has identified that gen AI has the potential to deliver an additional \$200-340 billion in value for the banking industry alone. The top use cases they see adding value across industries are using gen AI as a virtual expert, for content generation, for customer engagement, and accelerating coding.

But it's not just about automation and acceleration, the entire value chain can be transformed.

For example, in customer engagement, capabilities can include call center coaching, agents, virtual co-pilots assisting human agents, segment identification, and coordinated outreach campaigns.

For finance, future capabilities include advanced market intelligence and synthesis, and M&A target identification and strategy.

But, as I mentioned before, AI models also introduce risks, which will need to be addressed. For example, producing unfair and biased outputs, IP breaching, PII and data privacy breaching, poor LLM accuracy, hallucinations, toxic responses, and compliance violations. On top of that gen AI can become a very expensive exercise if the infrastructure is not used correctly. We're already seeing enterprises building gen AI apps where operating costs are higher than the value they bring.

For these reasons, it's important that enterprises build their architecture in a forward-thinking manner, maintaining flexibility and the ability to swap components in pipelines, involving open-source technologies, introducing guardrails, and supporting various LLM customization capabilities, like RAG and fine-tuning, to allow their architecture to change and evolve together with the rapidly evolving landscape.

Can you provide specific examples of how Iguazio's platform has helped enterprises in a specific industry, such as finance, healthcare, or retail to achieve tangible business outcomes?

Iguazio is trusted by enterprises across multiple industries: finance, telecommunications, manufacturing, mobility and more.

McKinsey acquired Iguazio in 2023 after a thorough market scan looking at hundreds of players in the field. Seeing that 90% of AI projects fail to deliver impact, and following extensive internal research into the solutions available in the market, they determined Iguazio is best placed to scale AI and gen AI across the organization, put thousands of gen AI apps into production, and, for the same investment, implement more projects, models, and impact for the enterprise.

One of our client is a fast growing European digital bank with 3.5 million customers and 120% YoY growth. Using Iguazio, they built a gen AI call center analysis app to improve call center operations, simplify agent training, improve customer experience, and reduce cost.

The gen AI application, deployed on-premises, summarized customer calls, analyzed sentiment and topics, and removed personally identifiable information (PII). The data was fed to downstream applications—like live agent support, customer profiles, auto-generated content, tailored recommendations, and customized offers.

For organizations just starting with gen AI, what advice would you give them to ensure they maximize their ROI while mitigating risks?

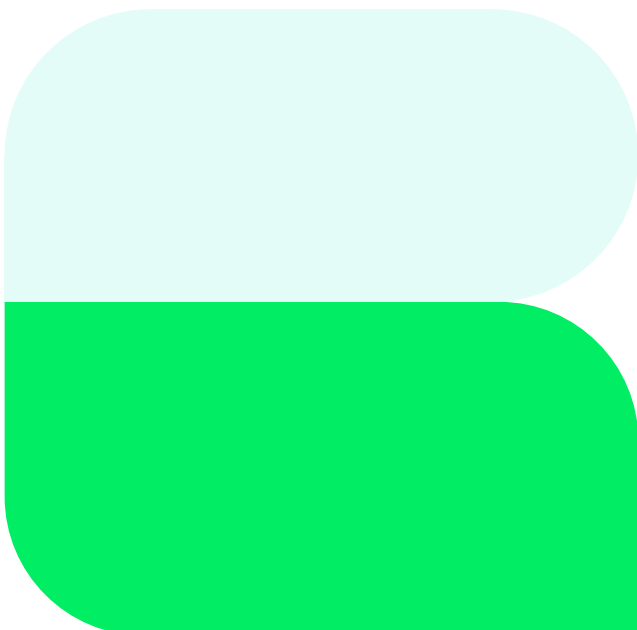
For organizations just starting with generative AI, there are several key steps to take to ensure maximum ROI and minimum risk. The first step is all business.

Begin by identifying and prioritizing the specific business problems you want to solve with generative AI. Whether it's improving customer service or analyzing activity.

Now, it's time to choose the right technology to support your business needs, such as a platform that can support your business goals and help you leverage AI to derive tangible business value. Depending on your needs and current infrastructure, it will often make sense to select a platform that supports the entire AI lifecycle—from data management to model deployment and monitoring. A platform like Iguazio's, which serves as a complete gen AI factory, allows for seamless and automated scaling and operationalization of AI models, accelerating time to market, and ensuring that your AI initiatives are both efficient and cost-effective.

Finally, establish clear governance frameworks to oversee gen AI model development and deployment. This includes setting up guardrails for regular model validation, bias detection, and performance monitoring.

This is a cycle, and you can and should always use feedback and monitor results, as well as update your models and applications as business needs change. These steps will ensure your AI projects are continuously bringing value to your business.





AI and Application Modernization Programs

AI modernizes apps faster for a better
user experience (UX)



MongoDB Atlas for Industries: Innovation Workshops



An innovation workshop is a carefully curated, virtual or in-person event where MongoDB industry experts and decision makers from an account come together to discuss the “art of the possible” relating to MongoDB-based solutions



Innovation Workshops

Accelerate mission-critical applications, drive innovation, and transform AI challenges into opportunities with unparalleled speed, reliability, and scalability



An Innovation Workshop is a carefully curated virtual or in-person event during which MongoDB industry experts and decision makers from customers and prospects come together to discuss the “art of the possible” relating to MongoDB-based solutions. Explore bespoke, modern app development and tailored solutions backed by our extensive industry knowledge and best practices. Discover exclusive workshops with MongoDB’s top experts.

Accelerate your modernization journey

[Learn more](#)



Industry-specific solutions

Gain insights into how MongoDB and adjacent technologies can provide end-to-end solutions for industry-specific challenges. Understand how similar issues have been addressed successfully by other leading companies in your field.



The art of the possible

Explore the art of the possible with MongoDB-powered solutions. Learn from real-world use cases that demonstrate how MongoDB can transform data management and modern application development within your industry.



Collaborative discussions

Engage in discussions with MongoDB experts and technology partners. Learn how others in your industry have leveraged MongoDB to solve complex problems and drive business growth.



Execution excellence

Understand MongoDB’s project execution capabilities, including successful collaborations with technology partners. Learn how we can help you achieve your business goals with efficiency and expertise.

MongoDB AI Applications Program



The MongoDB AI Applications Program (MAAP) helps organizations rapidly build and deploy modern applications enriched with AI technology at enterprise scale



MongoDB AI Applications Program

Rapidly build and deploy modern applications enriched with generative AI technology at enterprise scale



MAAP is a comprehensive program designed to accelerate gen AI application development for companies at every stage of their AI journey. It combines the technologies and services to progress customers from idea to roadmap to prototype to production. Molded to fit each organization's unique goals, timeline, and budget, customers are guided on the shortest and strongest path to success with a bespoke combination of resources.

Accelerate your AI journey



MAAP Ecosystem

Technology and expertise from the world's leading AI and tech organizations, across the entire gen AI stack, for a cohesive, fully supported application development experience.



Composable Architectures

Pre-designed architectures serve as accelerated frameworks for fully customizable and secure applications to accommodate ever-evolving gen AI use cases.



Expert Services

MAAP offers strategic guidance on roadmaps and skillsets, assists with data integration into advanced AI technologies, and helps organizations develop production-ready applications.



Education & Enablement

Comprehensive resources enable developers to quickly—and repeatedly—build modern AI applications faster while giving organizations the tools to expand their in-house AI expertise.

The MAAP to AI success

Access to trusted partners and a robust support system are critical for organizations that want to create a competitive advantage with AI in today's fast-paced innovation economy. Leaders are excited about gen AI, but are also unsure how to move from concept to production and want to control costs.

Many early adopters have struggled with the inefficiencies of legacy technologies that cannot scale or manage the real-time, multi-modal data required to power AI-enriched applications. This is compounded by a lack of in-house AI development expertise and the perceived risk of integrating disparate components without coordinated support. Due to these challenges, businesses aren't able to put AI-powered applications into production to deliver innovative, intelligent solutions to their customers.

Capturing the AI opportunity and cohesively bringing the cutting-edge innovation happening in the AI space directly to customers requires the right technology, investment, and collaboration from industry-leading providers across each layer of the gen AI application stack.

To directly address this need, MongoDB created MAAP, which offers customers a clear path to developing and deploying AI-enriched applications through reference architectures—an end-to-end technology stack that includes integrations with leading providers, professional services, and a unified support system. The cornerstone of MAAP is MongoDB technology. Applications are underpinned by MongoDB, which securely unifies real-time, operational, unstructured, and AI-related data without the need for bolt-on solutions.

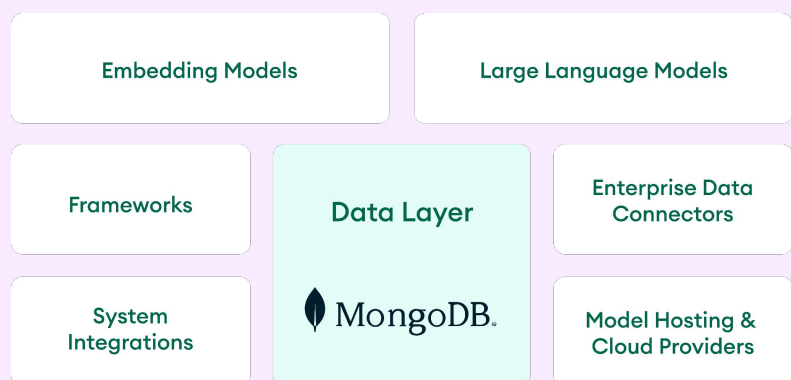
What's more is that customers can use any LLM provider and run anywhere (on all major cloud providers, on-premises, and at the edge). MongoDB offers seamless integrations with the entire ecosystem of AI frameworks and solutions.

Perhaps most importantly, thousands of customers already rely on MongoDB to power their mission-critical apps, and we have years of experience helping customers unlock the power of data. Ultimately, MAAP aims to help customers confidently innovate so that they can accelerate time-to-market, minimize risks, and realize the value of AI investments sooner.

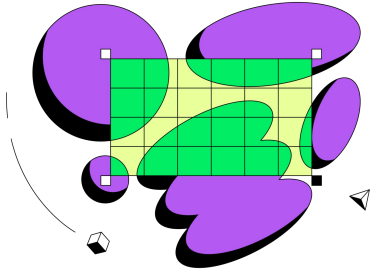


[Learn more](#)

The AI Stack



MAAP Ecosystem



While the [MAAP ecosystem](#) is just getting started, it already includes industry leaders from every part of the AI stack, including Anthropic, Anyscale, Arcee AI, AWS, Cohere, Credal, Fireworks AI, Google Cloud, LangChain, LlamaIndex, Microsoft Azure, Nomic, and Together AI.

The result is a group of organizations that will enable customers to build differentiated, production-ready AI applications—with simplified technology selection, reduced integration complexity, and optimized compatibility—while aiming to deliver substantial return on investment.

The AI Stack

The cornerstone of MAAP is [MongoDB technology](#). MongoDB sits at the heart of the AI application stack, the data layer, removing the friction of integrating, securing, and maintaining the essential data components required to build AI applications. MongoDB underpins these applications, securely unifying real-time, operational, unstructured, and AI-related data without the need for bolt-on solutions. MongoDB's open and integrated architecture provides easy access to the MAAP ecosystem and enables the extension and customization of applications to tackle any use case.

Embedding Models

Large Language Models

Frameworks


Data Layer

Enterprise Data Connectors

System Integrations

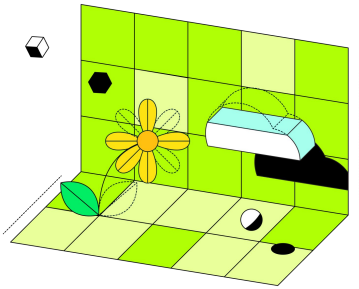


Model Hosting & Cloud Providers

 accenture cohere LlamaIndex ANTHROPIC CREDAL MongoDB® anyscale Fireworks AI NOMIC arcee.ai Google Cloud PeerIslands aws gravity9 pureinsights™ Azure LangChain together.ai Anyscale Capgemini QuantumBlack
AI by McKinsey CONFLUENT UNSTRUCTURED

The MongoDB AI Applications Program and its ecosystem of companies offers customers the right expertise and solutions for their use cases. It removes integration risk, making it easy for businesses to use the industry-leading technologies of their choice to unlock the value of AI with their data. Learn more about each member of the MAAP ecosystem in the [MongoDB Partner Ecosystem Catalog](#).

Composable Architectures



The [MAAP framework](#) provides access to a set of downloadable, pre-designed, pre-integrated RAG architectures with pre-configured UIs that serve as accelerated frameworks for building AI applications. Architectures are fully customizable and extendable to accommodate ever-evolving generative AI use cases, like [retrieval-augmented generation \(RAG\)](#) or advanced AI capabilities like Agentic AI and advanced RAG technique integrations. With MongoDB's open and integrated platform at its core, innovation with MAAP's composable architectures is unlimited.

Easily bring the power of leading AI platforms directly to your generative AI applications

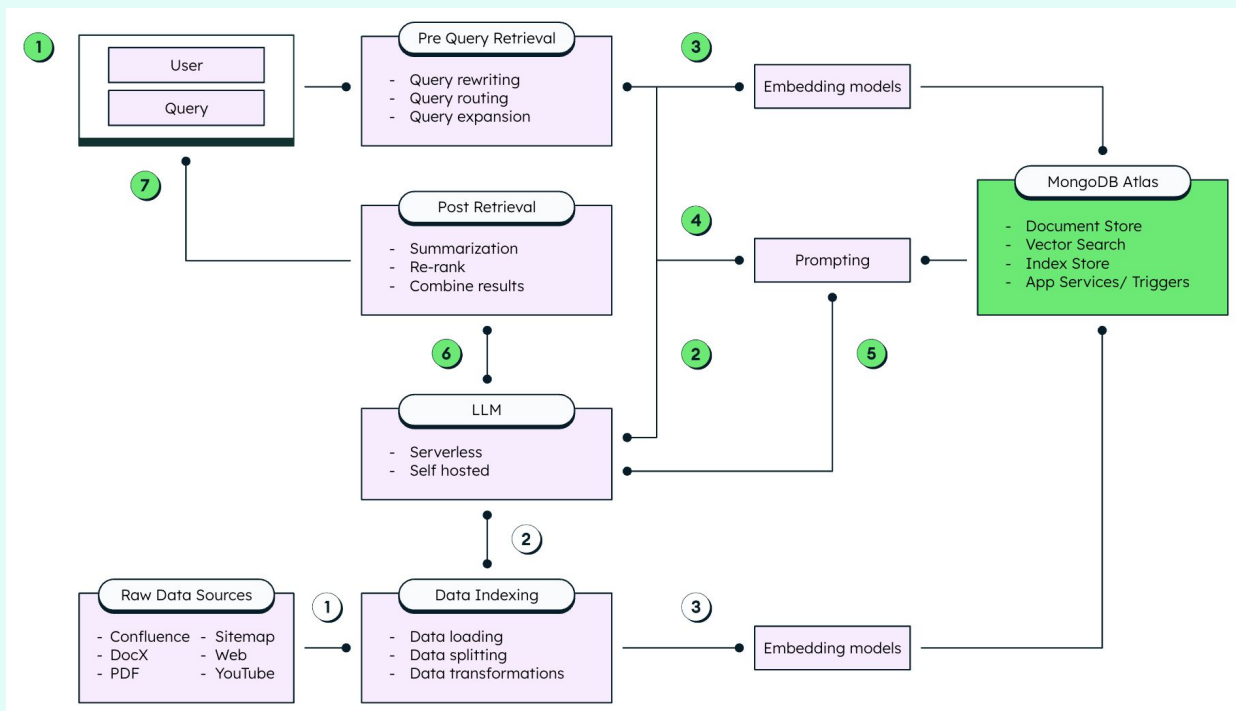
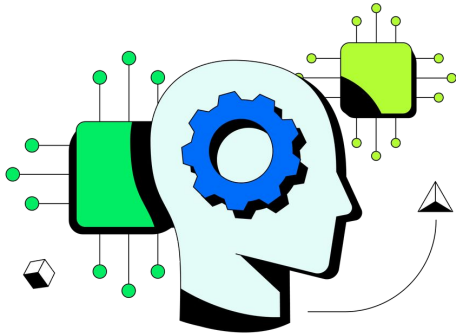


Figure 51: The MAAP framework is a set of libraries that you can use to build your RAG application using MongoDB and [Atlas Vector Search](#). Above is the reference architecture diagram of the framework with various components

Expert Services



MAAP expert services, combining the strengths of MongoDB Professional Services and industry-leading gen AI consultancies, will enable customers to rapidly innovate with AI. MAAP offers strategic guidance on roadmaps and skillsets, assists with data integration into advanced AI technologies, and can even develop production-ready applications. MAAP goes beyond development, empowering teams with best practices for securely integrating your data into scalable gen AI solutions, ensuring businesses are equipped to tackle future AI initiatives.

[Professional Services](#) can accelerate development of AI applications and get hands on, taking solutions into production. Learn more about [AI Accelerators](#) and [Application Development Services](#).

MAAP boutique and global system integrator partners include gravity9, PeerIslands, Pureinsights, and Accenture.

A major car maker in EMEA teams up with MongoDB Professional Services and Pureinsights to revolutionize noise diagnosis.

The Noise Diagnostics Application prototype—based on Atlas Vector Search and Google’s Vertex AI engine—uses car noise to assist mechanics perform root-cause diagnostics faster and more accurately, reducing dealership diagnosis time and vehicle immobilization. It also helps identify potential design issues and reassures clients about a car’s condition, reducing unnecessary visits to the dealership. This initiative not only cuts warranty costs but also enhances customer satisfaction and brand image for this leader in the automotive industry.

Built an Atlas Vector Search solution to search noise reports



83%

Of target noise reports appear in top 10 results

Evaluated cross lingual searches in 11 languages



73%

Mean accuracy

Evaluated 4 embedding models to ensure best performance with 32 automated tests



2

Weeks evaluation time

Education & Enablement

MAAP customers have access to a variety of [learning materials](#), including a dedicated [MAAP GitHub library](#) featuring integration code, demos, and a gen AI application prototype. These comprehensive resources enable developers to build intelligent, personalized applications faster, while giving organizations the tools to expand their in-house AI expertise. With MAAP, customers have access to integration and development best practices that they can use for future gen AI projects.

In addition to digital resources, our [MongoDB .local](#) series offers in-person opportunities to learn how to accelerate the delivery of production scale and safe AI-enriched apps.

Learn from experts and connect with community

MAAP Framework

The MAAP Framework is a set of libraries that you can use to build your RAG application using MongoDB, [Atlas Vector Search](#), and associated MAAP ecosystem partners.

MongoDB University

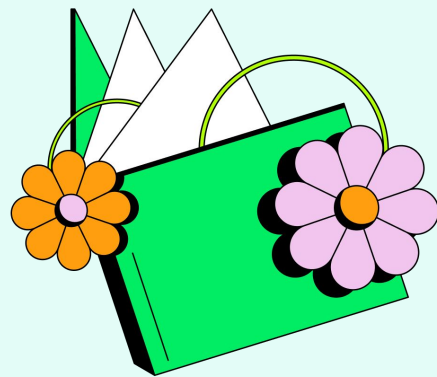
Learn to get the most from MongoDB products with our labs, courses and live training. Create a foundation with the [Introduction to AI and Vector Search](#) course or advance your learning with instructor-led training in [AI and Vector Search Basics](#).

Solutions Library

Drawing from experience with over 49,000 customers, the Solutions Library is curated with tailored solutions to help developers kick-start their projects. Get inspired by real gen AI solutions spanning diverse industries.

AI Resources Hub

View our library of articles, analyst reports, and case studies designed to help you build AI-powered applications.





AI Partnerships

How partners are leveraging MongoDB to build AI solutions

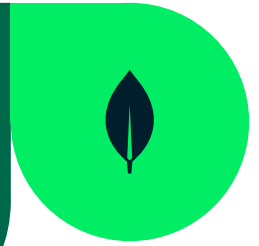
AI with MongoDB and hyperscaler cloud services



MongoDB Atlas on AWS, Microsoft Azure, and Google Cloud



Amazon Web Services (AWS)



Unlock the full potential of generative AI-powered applications with MongoDB Atlas and AWS. MongoDB Atlas on AWS allows you to build intelligent applications that are highly available, performant at global scale, and compliant with the most demanding security and privacy standards.

Build fast

Use MongoDB Atlas Vector Search and Amazon Bedrock to build and privately customize LLMs with real-time operational data.

Build efficiently

Get up to 60% faster query times and optimize costs when you isolate and scale gen AI workloads independent of the core operational database using MongoDB Atlas Search Nodes.

Fully Managed RAG in Minutes

MongoDB Atlas is a Knowledge Base for Amazon Bedrock, making it even easier to build generative AI applications backed by enterprise data.

With the click of a button, Amazon Bedrock integrates MongoDB Atlas as a vector database into its fully managed, end-to-end retrieval-augmented generation (RAG) workflow.

Build simply

Enjoy fully managed RAG and a unified interface and API for all data and application services. Eliminate the need for a bolt-on vector database and bespoke data pipelines.

Build safely

Integrations with [AWS PrivateLink](#) and other AWS services allow you to securely use proprietary data with gen AI across the organization.

Gen AI at Novo Nordisk



“Using Amazon Bedrock and MongoDB Atlas... we are the first in the industry to generate complete Clinical Study Reports in minutes rather than weeks.”

“We are doing it at scale, and with just a fraction of the resources we needed in the past. It is a game changer for healthcare around the world.”

[Learn More](#)

Louise Lind Skov

Head of Content Digitisation at Novo Nordisk



Microsoft Azure



MongoDB Atlas on Microsoft Azure empowers enterprises to build intelligent applications that drive real-world results derived from your data. Integrations with key Azure services, like Microsoft Fabric, offer a seamless, scalable, and secure platform to unify your data and launch experiences that harness the best of AI/ML and gen AI.

Architectural simplicity

Combine operational data, vector data, and metadata in a single database, then use that data with integrated Microsoft Azure services like Microsoft Fabric, Semantic Kernel, and more.

Connected platform

There are several ways to quickly and securely connect MongoDB Atlas to Microsoft Azure's suite of gen AI services, including via Fabric and Azure AI Studio.

MongoDB Atlas and Microsoft Fabric

A seamless integration with Microsoft Fabric enables you to run large scale AI/ML, analytics, and BI reports across your unified data estate on MongoDB Atlas.

Reimagine how teams work with data by bringing everyone together on a single, AI-powered platform designed to simplify and accelerate intelligent application development.

Dedicated scalability for Gen AI

MongoDB Atlas's dedicated search and vector search nodes allow you to dynamically scale AI workloads on Microsoft Azure's global infrastructure.

Security and compliance

Enterprise-grade security features and extensive compliance certifications from Azure and MongoDB.

An Intelligent Ecosystem

Enterprises need to take advantage of gen AI, AI/ML, and analytics to get the most from their data.

MongoDB Atlas, Microsoft Fabric, and Azure AI Studio operate as an ecosystem, driving actionable intelligence on historical data and real-time intelligence to power AI/ML and gen AI use cases.

Google Cloud



MongoDB Atlas and Google Cloud bring together the two halves of the gen AI equation—data and AI—to help organizations pursue the full potential of gen AI for their applications and software. The integration between Atlas and Vertex AI, Google Cloud’s unified ML and AI platform, empowers organizations to unlock the value of their data for AI, gen AI, and ML.

Creating and storing embeddings

Via the MongoDB API, Vertex AI vectorizes your data, generates embeddings, and passes them back to your MongoDB database for storage and retrieval, keeping your data secure and apart from the LLM dataset, yet readily accessible for model augmentation.

Streamline the gen AI Stack

MongoDB Atlas seamlessly integrates with Google cloud infrastructure, simplifying the connection between your data source and gen AI models.

MongoDB Atlas and Vertex AI

By harnessing the capabilities of MongoDB Atlas and Google Vertex AI, you can confidently deploy generative AI applications and overcome the barriers that hold many of your competitors back.

Speed application development, stay at the forefront of AI innovation, and deliver unparalleled user experiences—all while scaling effortlessly, reducing AI hallucinations, and keeping your data private.

Building RAG workflows

Using the Vertex AI platform, developers can perform CRUD (create, read, update, delete) on MongoDB Atlas.

Develop in natural language

Your teams can query MongoDB Atlas in natural language within Vertex AI with Google Gemini. Automatically turn human language into MongoDB-specific query syntax.

Serving Up AI for Delivery Hero



“With Atlas Vector Search we can compose sophisticated queries that quickly filter across product data, customer preferences, and vector embeddings to precisely identify hyper-relevant product recommendations in real time.” “We needed to move to an up-to-second real-time recommendations system, and that is what MongoDB Atlas Vector Search enabled us to do.”

[Learn More](#)

Mundher Al-Shabi

Senior Data Scientist, Delivery Hero



AI with MongoDB and System Integrators

Leverage the power of MongoDB alongside the expertise of Capgemini, Accenture, Pureinsights, gravity9, and Peerislands to achieve faster time-to-market, improved data management, and enhanced scalability for AI initiatives



Capgemini



Capgemini and MongoDB have joined forces to provide gen AI-powered solutions that accelerate businesses' AI journeys. Together, they've developed over 10 industry-specific accelerators to help organizations harness the power of AI more efficiently.

Genyoda

Gen AI-powered digital assistant revolutionizing the way insurance companies interact with their customers.

Customer analytic record

Enables successful, enterprise-wide hyper-personalization for banking, wealth management and fintech customers with humanized AI chatbots to streamline issue resolution and optimize customer interaction.

Card-not-present fraud prevention

Tackle real-time fraud in card-not-present transactions through a comprehensive system utilizing 3D secure protocol, behavioral biometrics, user behavior analytics, and machine learning algorithms. This event-driven fraud detection and prevention solution is powered by MongoDB.

Database migration-as-a-service

Database Convert & Compare (DCC) is a powerful tool developed by the Capgemini to optimize activities like database migration, data comparison, validation, and much more in a database migration roadmap. When migrating from RDBMS to MongoDB, DCC achieves 70% automation and 30% manual retrofit on a database level.

Health druid

Revolutionizing the healthcare industry with AI-powered insights, enhancing the patient journey through personalized and compassionate care.

Virtual store assistant

Transforming retail journeys with intelligent virtual assistants for tailored shopping experiences across platforms and channels for higher conversion rate.

Trusted Vehicle

Accelerate time-to-market with a secure and scalable platform of next-generation driver and fleet-management experiences..

“Building data-driven applications in which AI is embedded is a new frontier for many organizations, one that presents new operational and application development challenges, and requires a new way to think about data at the application level. MongoDB has some really good foundations to start making that happen, and when combined with an Operational GenAI accelerator like RAISE gives you an entirely new collaborative tool chain that is designed for building these next generation digital applications.”

[Learn More](#)

Steve Jones

EVP, Data Driven Business & GenAI,
Capgemini

Accenture



Together, MongoDB and Accenture provide unparalleled expertise to help customers modernize their environments and adopt a cloud-first approach throughout their organization. As the cloud becomes more critical to managing data at scale, our partnership helps enterprises unlock data from legacy data warehouses and lakes to build new applications faster.

Lift and shift

Transition your organization from a self-managed cloud to a multi-cloud continuum, meeting modern storage needs while still preserving traditional enterprise capabilities.

Data modernizer tool

Experience faster MongoDB adoption with accelerated data modeling while enabling seamless migration of applications to the cloud.

Creating a culture of experimentation

Thomas Edison's 10,000 failures led to 10,000 successful ways that didn't work. Organizations need this spirit of experimentation to innovate. Mark Porter, MongoDB CTO, and Michael Ljung, Global Software Engineering Lead and Chief Software Engineer for Accenture, [discuss how to encourage a culture of experimentation](#) and normalize failed experiments.

Legacy modernization

Reimagine legacy monolithic applications to harness the true potential of the cloud and accelerating your shift to modern, scalable architectures.

Smart data mover

Leverage a comprehensive toolkit designed to seamlessly migrate data from on-premise data warehouses to cloud data repositories.

Migration made easy

“Accenture sees growing demand from companies for solutions that can easily migrate applications and data migrations to the cloud. Our Smart Data Mover solution accelerates this process for moving data into MongoDB and other target platforms. We continue to invest in the Smart Data Mover application and look forward to expanding our capabilities using MongoDB as a target platform.”

[Learn More](#)

Shail Jain

Global Managing Director, Data & AI at
Accenture

Pureinsights



Pureinsights transforms the way organizations interact with information, leveraging cutting-edge AI technologies like Generative AI, Vector Search, and NLP to build intuitive, human-centered applications that go beyond traditional search. From information retrieval to innovative use cases like audio diagnostics, we help businesses uncover new insights and drive innovation.

Pureinsights, MongoDB's [premier services partner for search and AI](#), has built one of the best Retrieval Augmented Generation (RAG) architectures for a large European car manufacturer using MongoDB's tech stack and by leveraging our comprehensive ecosystem for AI. Pureinsights' expertise in search and AI, along with the [Pureinsights Discovery](#) platform, have demonstrated the capability to help customers take their ideas for search and AI from business idea, to working prototype, to phased production rollout – all while delivering business value.

“Pureinsights exemplifies the innovation and dedication to customer needs that define our leading partners, earning them this year's Services AI Partner of the Year award.”

Alan Chhabra, Executive

Executive Vice President at MongoDB

Accelerating AI with MongoDB

As a [MongoDB BSI partner](#) and [MAAP launch partner](#), Pureinsights is a preferred service provider for MongoDB's GenAI Build initiatives. The purpose of these initiatives is to accelerate the planning and prototyping of AI-powered applications on MongoDB platforms which will result in faster time-to-value for customers. Pureinsights staff have over 15 years of experience with commercial or open-source data, search and AI platform.

Gen AI Initiatives

- [GenAI Build Essentials](#) is an initial project assessment engagement to scope out a plan to build a Generative AI (GenAI) solution with Retrieval-Augmented Generation (RAG) capabilities on MongoDB Atlas.
- [GenAI Build Implementations](#) are a follow-up to Essentials. This service leverages Pureinsights' and MongoDB's technical and development expertise to build a working prototype of the GenAI solution implemented on MongoDB Atlas with retrieval-augmented generation (RAG) using synthetic data based on your existing datasets in a MongoDB environment.



[MAAP Partner](#)



[gravity9](#) is a leading technology partner, driving businesses beyond outdated legacy systems into the future with AI-led, cloud-based platforms designed for continuous innovation. Our expertise in application modernization harnesses advanced cloud technologies, artificial intelligence, and microservices architecture to streamline operations, elevate user experiences, and foster ongoing growth. Through our sleek micro-UI platform, we empower organizations to thrive in a rapidly evolving market, ensuring agility, scalability, and a lasting competitive edge.

gravity9 is proud to be one of only six global strategic MongoDB implementation partners, offering comprehensive solutions that combine MongoDB's powerful data platform with gravity9's expertise in Generative AI and data analytics. gravity9 harnesses the power of MongoDB Atlas to deliver advanced AI solutions and develop applications based on Large Language Models (LLMs), Retrieval Augmented Generation (RAG), graph knowledge base and GraphRAG. Our expertise also extends to Agentic AI solutions for workflow automation, driving efficiency and innovation. By integrating these cutting-edge techniques and leveraging exclusive access to MongoDB resources and insights, we help businesses embrace the future of AI and data-driven growth with confidence.

Partnering for Success with MongoDB

gravity9's deep partnership with MongoDB enables them to deliver a flexible, best-practice approach to technology modernization. Backed by rigorous training and certification, gravity9's team is fully equipped to design, implement, and optimize MongoDB solutions tailored to each client's needs. Whether migrating, optimizing, or building new applications, they ensure seamless integration and rapid deployment, maximizing the value of each investment.

Digital Transformation Experts

gravity9's highly skilled professionals are dedicated to transforming digital applications, rapidly revitalizing outdated systems with tailored, innovative solutions. Their approach unlocks new possibilities for businesses across various industries by delivering efficiency, flexibility, and enhanced functionality.

PeerIslands



Since 2018, [PeerIslands](#) has been leading the AI revolution with an exceptional team and state-of-the-art technology, propelling AI-driven software solutions into the Enterprise and SMB marketplace. Central to our mission is a handpicked team of top-tier, Top 1% developers, leveraging advanced AI platforms and our distinctive ‘Human in the Loop’ approach. We go beyond conventional software development to turbocharge your enterprise applications, reimagine outdated systems, and harness the power of your data.

We work very closely with MongoDB and complement a great database product with end to end solutions that customers want. With a talent pool of over 150 MongoDB certified Polyglot engineers and our active collaboration with Enterprise clients, PeerIslands has been named MongoDB's Boutique SI of the Year for 3 of the last 4 years. Our software delivery expertise, combined with MongoDB technology and Solution Architects, has been a recipe for success for clients seeking to modernize their applications and migrate their data to a modern, scalable, and flexible data architecture.

PeerIslands' Polyglot developers enhance MongoDB's capabilities by providing comprehensive solutions that drive client success. Our team boasts extensive experience delivering critical applications, demonstrating our commitment to excellence.

Payments Platform Modernization

PeerIslands is currently working closely with a leading Payment Services Provider who supports the backbone of top financial institutions that rely on services such as wire transfers, real-time payments, and bill payments. They are collaborating with the client on a multi-year modernization journey aimed at transforming critical business applications using MongoDB Atlas. The combination of PeerAI and MongoDB Atlas has played a crucial role in accelerating the client's time to market and facilitating the transition from traditional databases.

Digital Transformation Experts

- Dramatically speed up the transition from legacy platforms, reducing costs and providing a [70% time-to-market advantage](#) for new products and services.
- With MongoDB at the core, their platform features a data model that supports their leadership's vision for AI-enabled services for their customers.

Accelerate
Business

Accelerate
Technology

Accelerate
Value

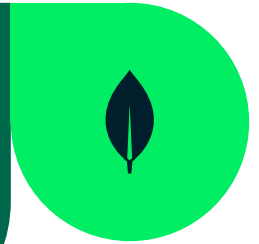
Unlocking the Power of AI With SaaS



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 can 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 that AI and gen AI applications don't just remain in the lab, but have 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:** enables 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**.

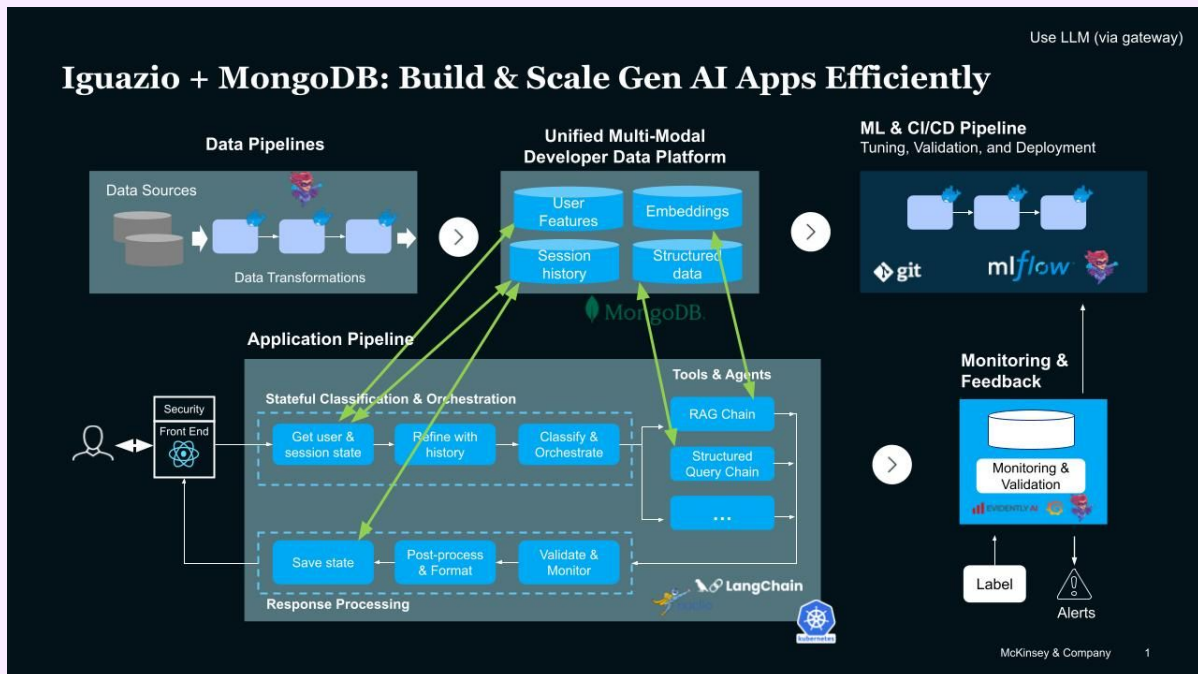


Figure 52: How to build and scale gen AI applications efficiently with MongoDB and Iguazio

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

Build safe, reliable, enterprise-grade gen AI solutions with MongoDB and Anthropic



[Anthropic](#) is a pioneering AI research and product organization, dedicated to developing reliable, interpretable, and trusted AI systems. As a public benefit corporation, Anthropic is committed to ensuring the world's safe transition through transformative AI. Their flagship product, [Claude](#), is a family of trusted AI models designed for enterprise applications across every industry, which securely connects to company knowledge to help every team benefit from trusted AI.

Reliability

Claude powers business-critical use cases, exhibiting lower hallucination rates and higher accuracy.

Safety

Claude offers robust security and compliance features, meets SOC 2 Type II and HIPAA standards, and ensures enterprise-grade protection.

Trustworthiness

Claude combines best-in-class jailbreak resistance and misuse prevention to mitigate brand risk for enterprises.

Flexibility

The Claude model family offers a range of AI solutions across the price-performance spectrum, providing enterprises with options for any use case.

Bring enterprise data to gen AI and use a family of flexible foundation models

Anthropic and MongoDB offer collaborative support to help you navigate the complexities of building safe, reliable, enterprise-grade generative AI solutions. Together, we provide comprehensive solutions for enterprises—so you can launch and scale generative AI applications safely and reliably, customized with your data.

- **Tailored AI for Your Needs:** Deploy and scale generative AI applications tailored to your use case and leveraging your data.
- **Compliance Built-In:** Ensure your generative AI apps are secure and compliant out of the box.
- **RAG Made Easy:** Abstract the challenges of building a Retrieval-Augmented Generation (RAG) workflow with your data.

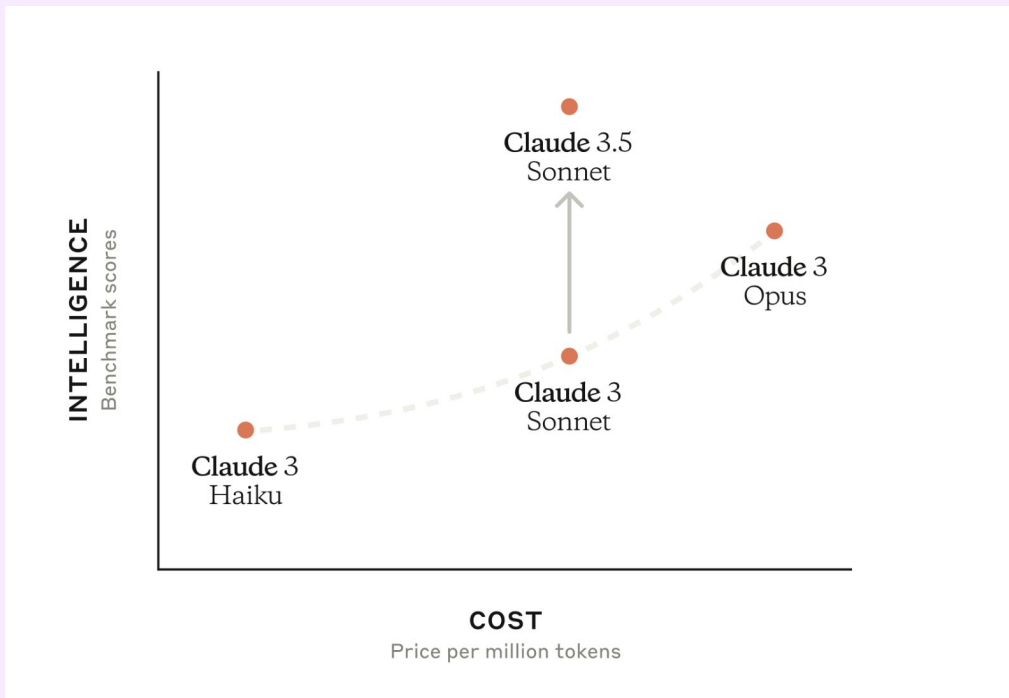


Figure 53: Cost vs intelligence of Claude models.

The Claude model family offers state-of-the-art performance across a wide range of tasks, allowing users to select the optimal balance of intelligence, speed, and cost for their specific applications. Claude’s capabilities include:

- **Advanced reasoning:** Claude can perform complex cognitive tasks that go beyond simple pattern recognition or text generation
- **Vision analysis:** Transcribe and analyze almost any static image, from handwritten notes and graphs to photographs
- **Code generation:** Start creating websites in HTML and CSS, turning images into structured JSON data, or debugging complex code bases
- **Multilingual processing:** Translate between various languages in real-time, practice grammar, or create multilingual content

Agnostiq & MongoDB: High-Performance Computing for All



[Agnostiq](#), founded in 2018, is transforming how AI developers and researchers run compute intensive applications that utilize GPUs and high-performance computing (HPC) resources. With [Covalent](#), a workload orchestration and infrastructure management platform, developers code their projects locally and run on serverless hardware, without needing extensive machine learning operations expertise. Agnostiq's commitment to interoperability and resource neutrality ensures accelerated computing is accessible to every start-up, research institution, and enterprise.

Train, fine-tune, and serve AI at scale

Developers and researchers can accelerate the entire AI development and deployment process - from training, fine-tuning, benchmarking, data synthesis, to inference - in a cost-effective way.

Simplify agentic workflows

Take AI agents to the next level by expressing complex interactions as a seamless Python workflow and even deploying each agent in its own separate compute environment.

Python abstracted infrastructure

Developers simply add a few lines of Python code to access GPU and HPC infrastructure, without the burden of learning Kubernetes, Slurm, or other DevOps tools.

Run efficient AI & HPC infrastructure

Covalent automatically pools, assigns, and scales resources to workloads dynamically, in order to meet changing demands for compute and optimizing infrastructure efficiency.

Building an Environment Agnostic Workload Orchestration Platform with MongoDB

- **Data Scalability:** MongoDB Atlas provides an ideal foundation for modern AI and HPC applications which require serverless compute, autoscaling resources, and distributed workloads
- **Ease of Use:** MongoDB gives the small, agile development team the freedom to build and manage data workflows without the need for a specialist DBA
- **Multi-Cloud Capability:** MongoDB Atlas allowed Covalent to reach multi-cloud compatibility faster than standard tooling and give users the freedom to move data between AWS, Google Cloud, and Microsoft Azure.
- **Open Source Optionality:** Covalent was originally released as an open source project, and the MongoDB NoSQL database was the perfect complement for quickly maturing product roadmap.

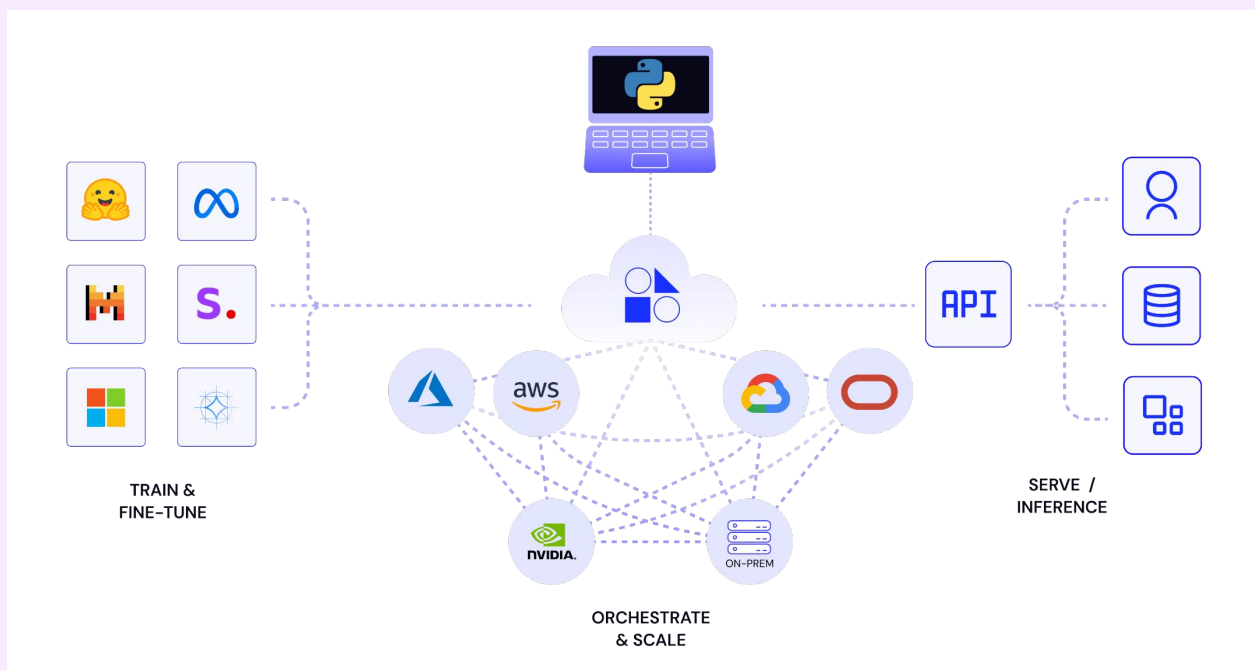


Figure 54: Covalent by Agnostiq accelerates AI development and deployment with compute orchestration, dynamic resource allocation, and auto-scaling infrastructure

Agnostiq chose MongoDB as their default NoSQL database for the free, open source version of Covalent. Without any DBAs as a small agile team, MongoDB gave Agnostiq the freedom to build and manage data workflows without the need for a specialist.

As their customer base grew along with the demand for cloud computing access, Agnostiq moved to MongoDB Atlas, gaining the freedom to move data seamlessly between AWS, Google Cloud, and Microsoft Azure. This gave Covalent the flexibility to reach multi-cloud compatibility at a faster rate than with standard tooling.

Covalent provides a workflow management service by registering jobs, dispatching IDs, and collecting other metadata that allows fellow researchers and developers to reproduce the original work.

MongoDB is used in the front-end, allowing a high volume of metadata and other assets to be published and cached in accordance with an event-driven architecture. This near real-time experience is key to a product aimed at delivering a unified view over distributed resources. MongoDB Atlas further provides the autoscaling required to grow with the user base and the number of workloads while keeping costs in check.

“MongoDB Atlas helps us provide an ideal foundation for modern HPC and AI applications which require serverless compute, autoscaling resources, distributed workloads, and rapidly reconfigurable infrastructure.”

Santosh Kumar Radha, Head of Product at Agnostiq

Credal.ai: Build secure AI assistants for Enterprise operations



Credal helps enterprises build secure AI applications: We provide a secure AI platform that integrates seamlessly with internal data systems, ensuring privacy, compliance, and ease of use for enterprises to build and manage AI applications.

Credal uses MongoDB Atlas for performant, high scale vector search for enabling AI powered applications. Credal has multiple deployment options including Cloud, Managed Single Tenant and On-prem with pluggable MongoDB Atlas instances managed either by Credal or the customer.

Give every employee the ability to create the perfect AI Assistant for their work.

Customer genie

Answer questions about any customer in a shared Slack channel, from disparate data sources.

Security questionnaires

Speed up deals and customer time to value by offloading security approvals to AI.

Product bot

Slash time spent on internal knowledge sharing and back and forth, democratize product expertise.

Sales coach

Analyzes call transcripts & recommends on missed questions, discovery opportunities, and objection handling.

Robust Security and Compliance for Your AI Initiatives

- **Security & Compliance built in:** Fully sync permissions from source systems (eg Google Drive, Atlassian products, and other SaaS tools). Wide deployment options including Cloud, managed single tenant, cloud-prem, and on-prem to meet your security needs. Automatically redact sensitive data like PII/PHI from AI to further improve compliance and security posture, while letting your users experiment, iterate and be productive. Of course, Credal takes care of enterprise security fundamentals such as SAML/SCIM, audit logging, and RBAC out of the box.
- **ROI from the SaaS tools your enterprise relies on:** Credal connects to the most important tools in your enterprise, such as Google Drive, Microsoft Sharepoint, Microsoft Onedrive, MongoDB, Box, Notion, Salesforce, Zendesk, Confluence, Jira and Slack. Non-technical users can easily sync data from these tools for use in their assistants. Credal automatically honors the permissions on all your enterprise data - regardless of the end user.
- **Trusted by Giants:** Credal helps companies scale AI and manage complex data, serving clients like Wise, MongoDB, and unicorn startups like Lattice.



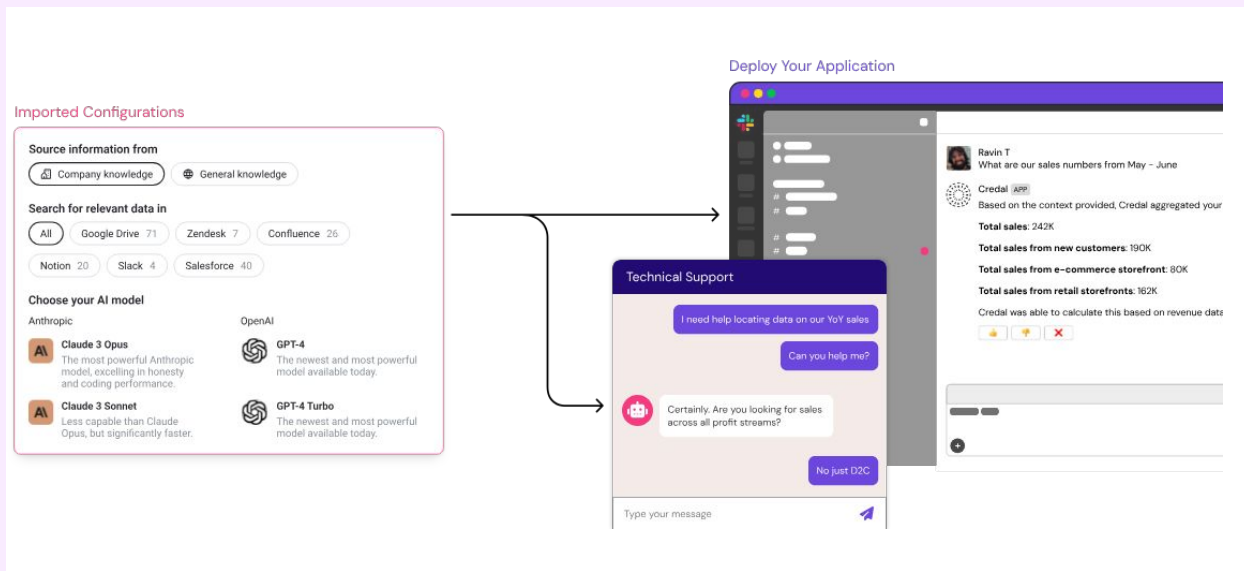


Figure 55: Bringing your data to LLMs and bring LLMs to your operations.

MongoDB integration

Credal seamlessly integrates your organization's MongoDB data with AI, offering direct developer access for flexible customization. It also supports dedicated collections for curated data with complex retrieval strategies in MongoDB.

Credal offers flexible deployment options: cloud, managed single tenant, or on-premises with pluggable MongoDB Atlas instances.

Leverage your existing investment in MongoDB to integrate AI into your operations: Credal also lets developers and data engineers use MongoDB as a data source unto itself, to unlock new use cases that are only possible with your application data.

Point and click AI Copilots

Credal's AI copilots empower users to set up dedicated assistants for a wide range of use cases, from customer support to contract review.

Copilots can assist with any task that combines AI and data. Copilots are designed to be experts on the data and context you provide. They combine AI with your data to provide accurate, context-aware responses, while citing their sources.

Copilots are useful to the full spectrum of non-technical builders to highly skilled developers. Our API documentation is available at docs.credal.ai.

Anyscale enables MongoDB users to deploy and seamlessly scale their AI workloads



[Anyscale](#), founded in 2019 by Professor Ion Stoica and the Berkeley RISELab team, developed Ray, a leading AI Compute Engine for scaling AI workloads. The Anyscale Platform provides a fully-managed version of Ray, optimized for performance, scalability, and developer collaboration.

Optimized Infrastructure & Scaling

Anyscale scales effortlessly across heterogeneous clusters (GPUs/TPUs) on any cloud, ensuring high reliability and optimized performance for production workloads as data and user loads grow.

Any AI Workload, Any Framework

Anyscale's Ray Engine and high level libraries supports end-to-end AI workflows, from GenAI to deep learning, embeddings, and data processing. Use your data for any AI use case with any framework (TRT-LLM, vLLM, PyTorch, TensorFlow, etc.).

Leading Performance

Anyscale's Ray enhances performance, scalability, cost-efficiency, and utilization through optimizations ranging from hardware tweaks to intelligent instance management, making it the ideal platform for running Ray.

Powerful Developer Tooling

Anyscale Workspaces provide scalable development environments, simplified dependency management, and powerful observability and debugging tools, streamlining the dev-to-prod lifecycle so teams can build, test, and deploy faster.

Anyscale + MongoDB: A Powerful Pair for AI

- **Seamless Integration:** [MongoDB Atlas](#) integrates Anyscale's AI Compute Platform effortlessly, enabling users to scale AI workloads from Gen AI embeddings to model training and inference. Both [MongoDB Atlas](#) and Anyscale are available on AWS Marketplace for identical VPC deployments.
- **Scalable Efficient Multimodal Data Processing:** Process and query diverse data types, including images, text, and structured data, in real-time, using MongoDB's flexible schema and Anyscale's scalable compute.
- **Gen AI and RAG Applications:** With Anyscale for fast and efficient LLM inference and MongoDB Atlas scalable vector indexing for contextual search, users can build and deploy super scalable, AI-based retrieval-augmented generation (RAG) flows and agentic systems..
- **Simplified Developer Experience:** MongoDB's Developer Data Platform and Anyscale's distributed compute platform combine to give developers the power to deliver AI use cases across their organization without managing infrastructure

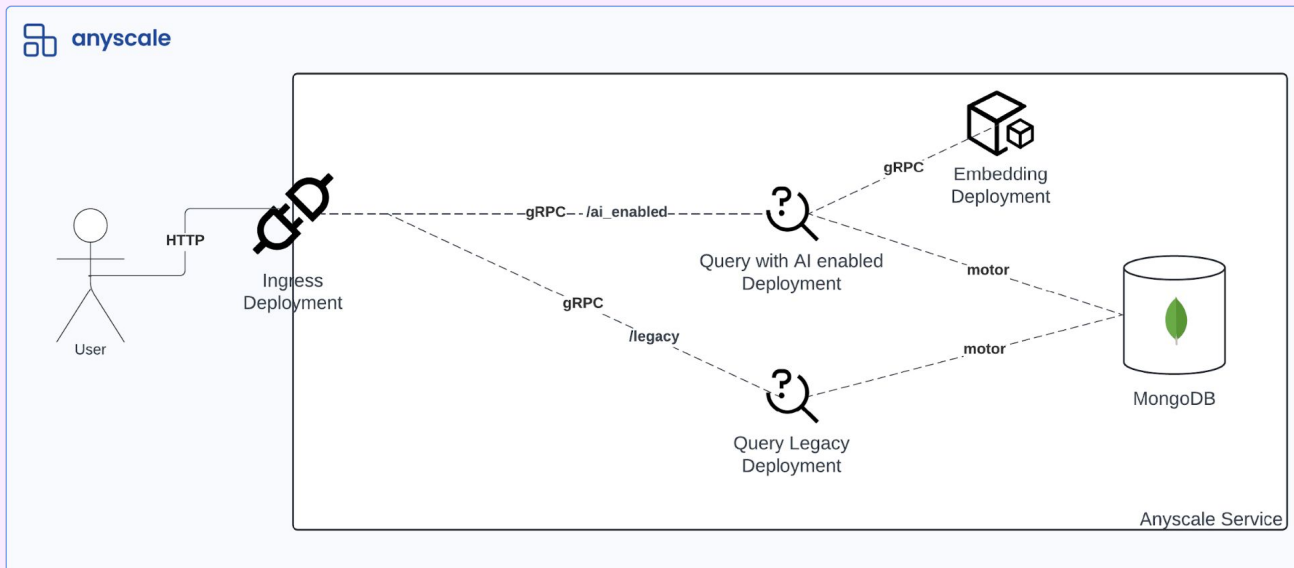


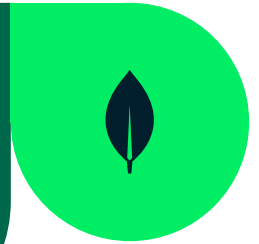
Figure 56: Multi-Modal Search Application - application is composed of multiple components - including legacy search and the AI-enabled search - that work together to provide a hybrid search experience

Multi-Modal Search is a cutting-edge use case that leverages Anyscale's distributed AI compute engine and MongoDB Atlas's flexible NoSQL database to create a seamless search experience across diverse data types, including text, images, and structured data. Anyscale provides the infrastructure to easily scale and optimize compute resources, enabling rapid processing of multimodal data, while MongoDB Atlas efficiently stores and indexes these datasets.

The combine solution makes it possible to perform fast and accurate similarity searches across large volumes of complex, unstructured data. With this architecture, users can quickly retrieve relevant content from a variety of formats, such as images, documents, and even video, without being burdened by infrastructure management. This makes it ideal for building scalable, AI-driven search applications across multiple domains while optimizing for resources utilization and scalability.

- Enable vector search across multiple fields and dimensions via [MongoDB Atlas Vector Search](#)
- Supporting the storage of multi-modal data like images, text, and structured data in MongoDB Atlas
- Run performant LLM Batch Inference compute jobs with Anyscale's Platform
- Enable highly available and scalable deployments with Anyscale Services
- Optimally scale and maximize utilization of costly compute resources with Anyscale's platform and intelligent infrastructure management

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 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.
- **Streamlined Integrations:** MongoDB's robust API set simplifies integration with Basikon's pre-built integrations and external systems, ensuring a smooth flow of data.
- **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.
- **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.

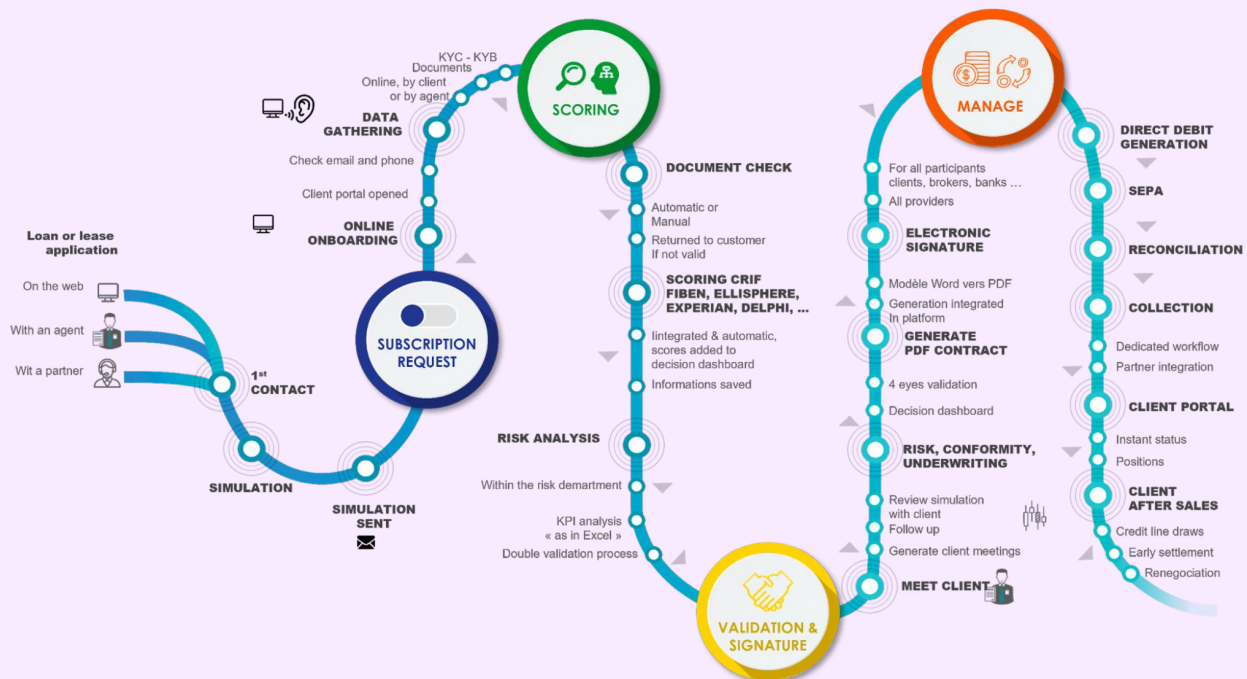


Figure 57: Basikon handling the complete customer life cycle

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 the Figure above, 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

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, Compliant 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.

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.

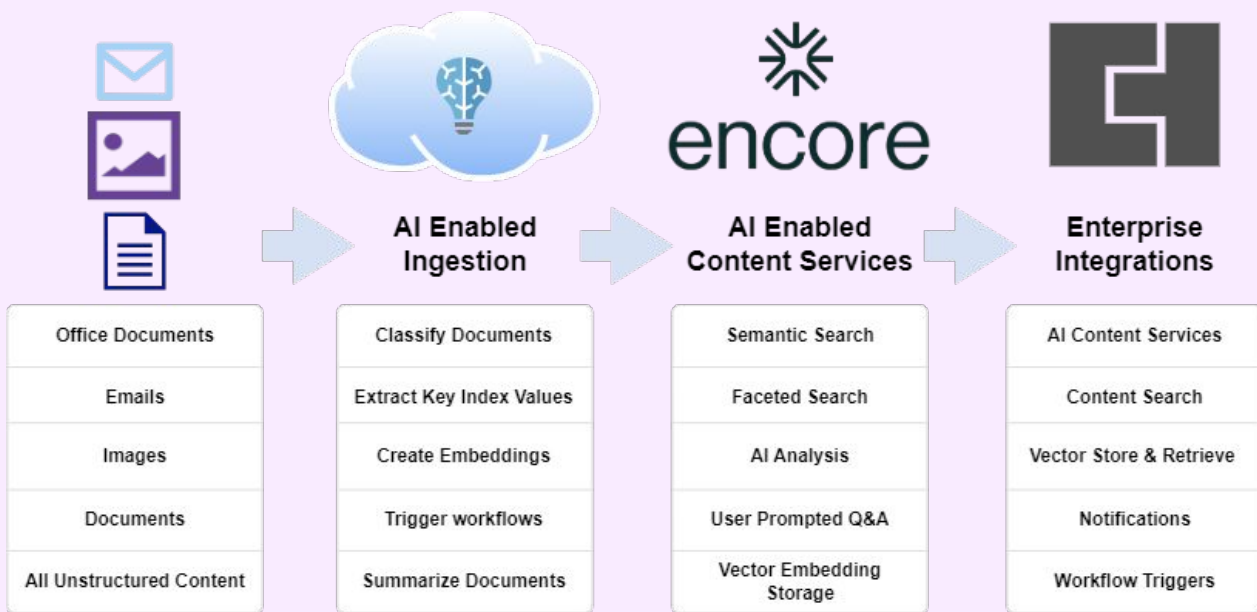


Figure 58: The value of Encore

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 easier. 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 searching for 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.



Figure 59: 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 the Figure above, 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 easily.

Imagine your chatbot being able to learn and improve over time, 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, and 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.

Adapting to future advancements

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

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.

Human-in-the-loop approach

Devnagri integrates human feedback 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.

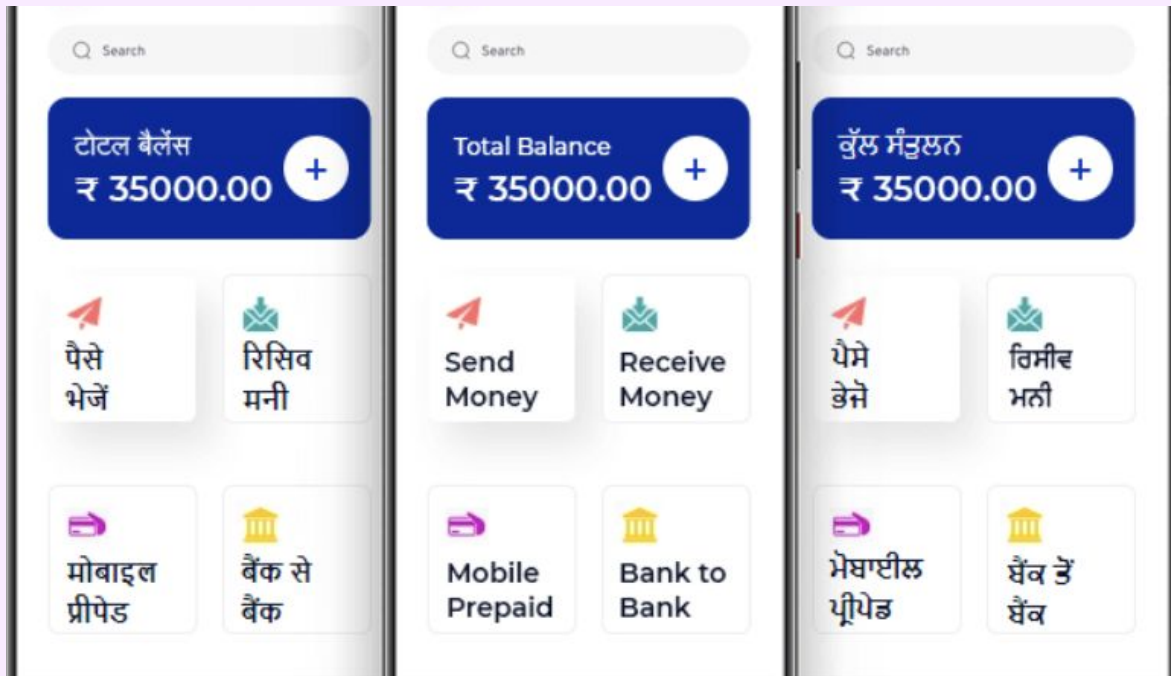


Figure 60: A visual of Devnagri's real-time translation engine

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.

CRM Integration

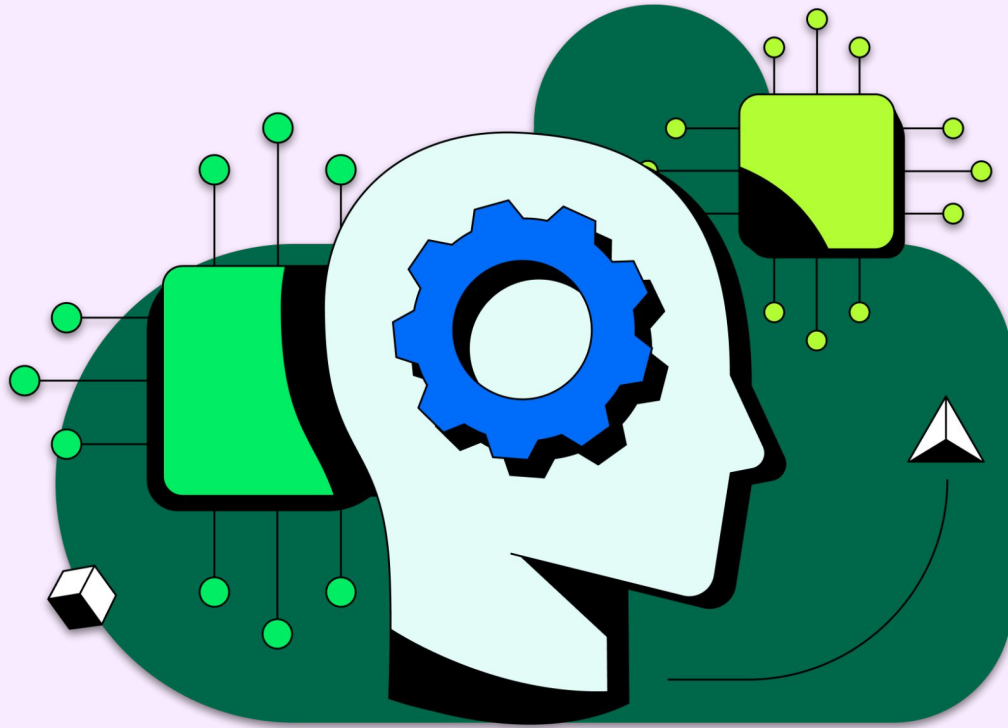
After each meeting, WINN.AI extracts and summarizes relevant information, updating the CRM system with follow-on actions, eliminating manual data entry, saving time, and reducing errors.

Contextual Understanding

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

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.



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.

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



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.

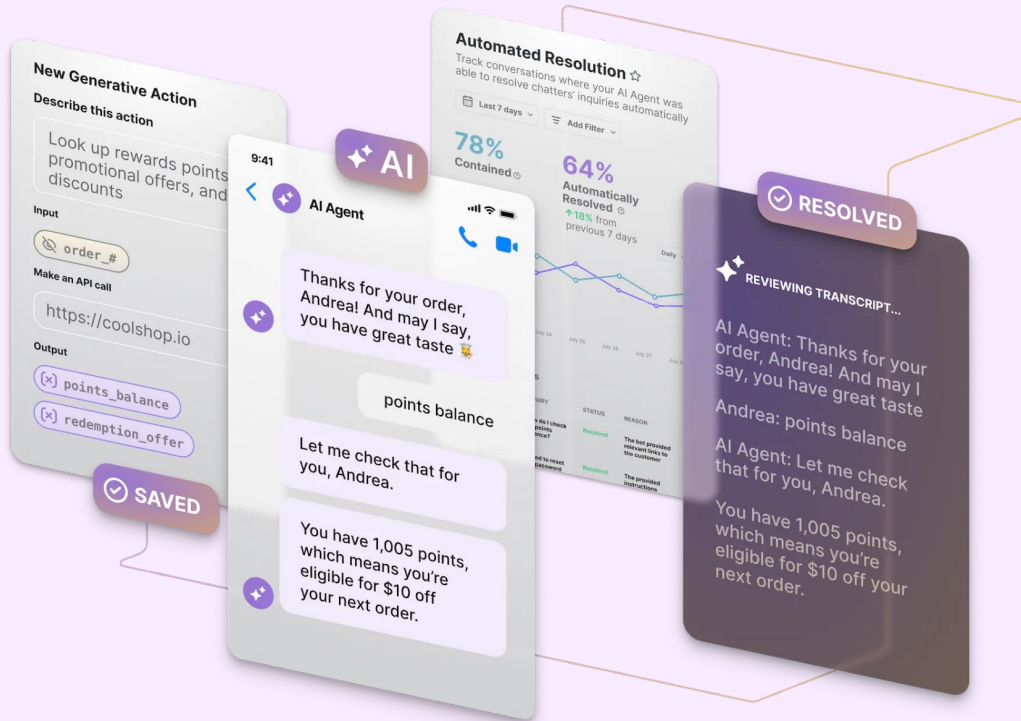


Figure 61: Ada's AI customer service

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: Gen AI 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.

Hyper-Personalized Encounters

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

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.

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.
- **Support for Machine Learning Models:** MongoDB provides the necessary data for training and fine-tuning XOLTAR's sophisticated machine learning models.
- **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.
- **Real-Time Notifications:** With Atlas Triggers, MongoDB notifies downstream consuming applications so they can react in real-time to new treatment recommendations and regimes.

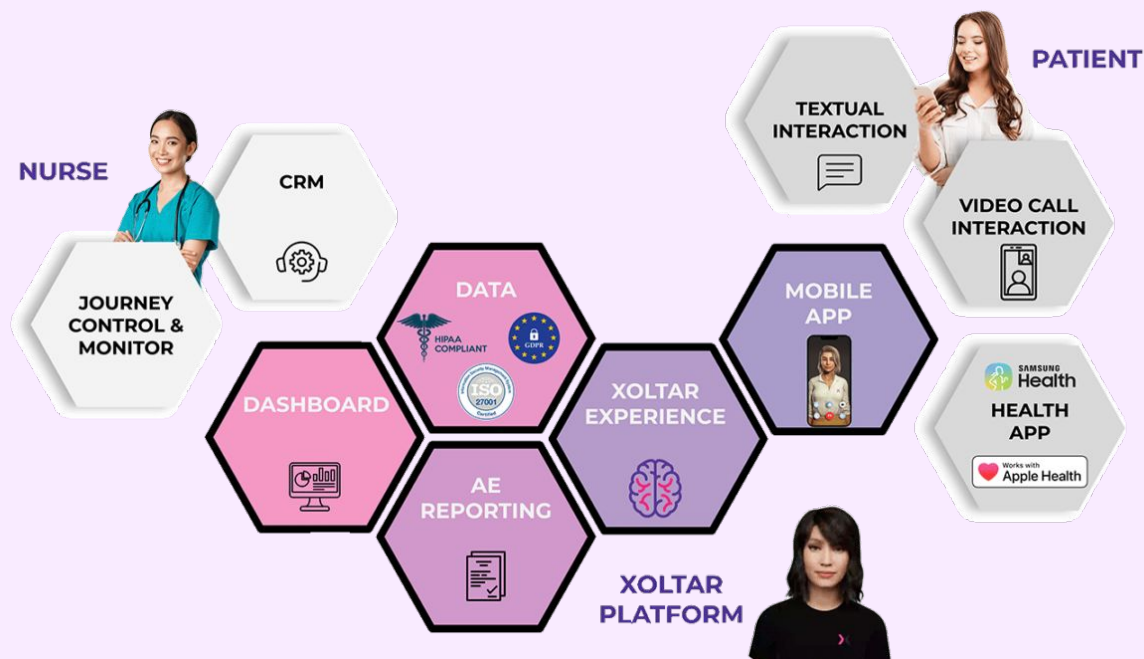


Figure 62: How the XOLTAR platform works

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.

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.

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.

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

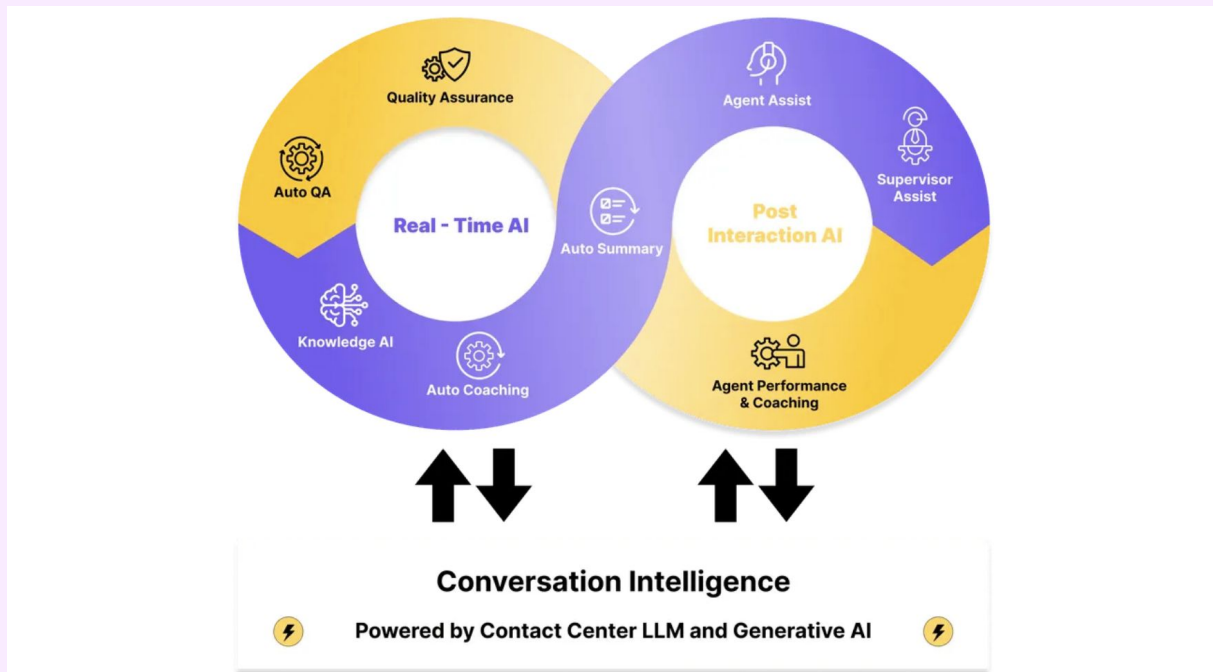


Figure 63: Observe.AI's conversation intelligence

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**.

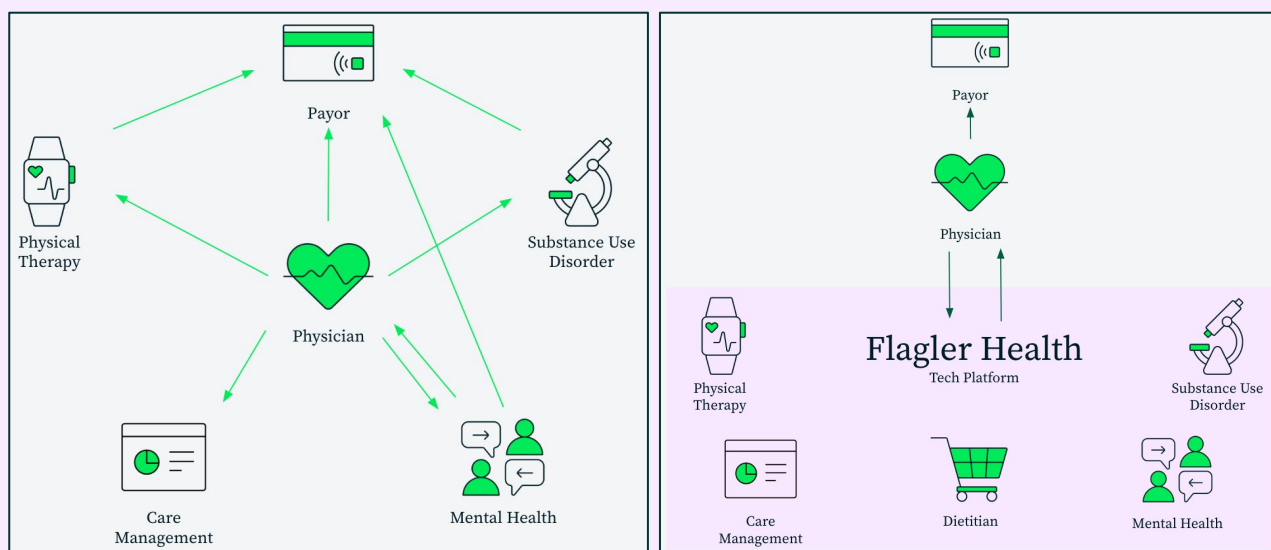


Figure 64: Current state without Flagler Health (left) and what Flagler Health can offer (right)

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.

Dataworkz: Generate Faster Data Insights with Gen AI Apps & Proprietary Data



The [Dataworkz](#) gen AI applications platform provides an all-in-one RAG as a Service to rapidly build, deploy, operationalize and scale gen AI 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 gen AI 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 gen AI app adoption

Implement additional use cases, connect new data sources and use RAG APIs to embed gen AI 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 gen AI 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**.

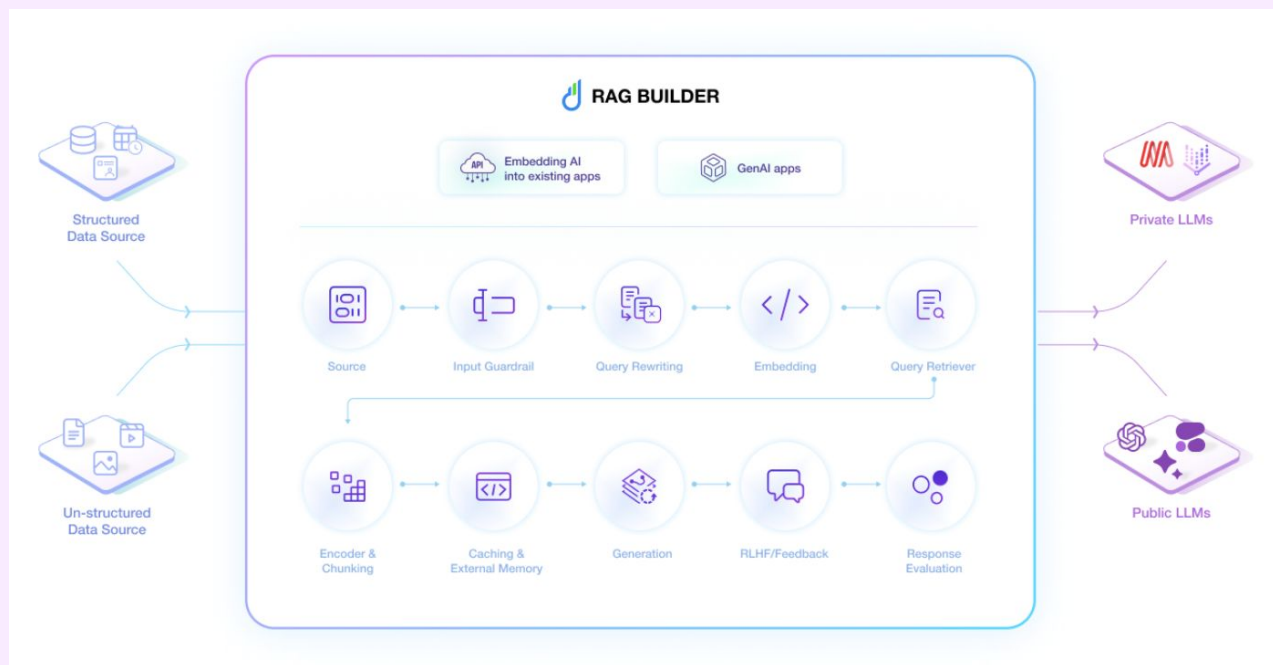


Figure 65: A leading-edge platform to rapidly build, deploy, operationalize and scale gen AI applications

Why Dataworkz: Unique capabilities for gen AI applications using RAG

Build

Visually Create Gen AI Applications:

Develop gen AI 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.

Observe

End-to-End Traceability: Get full visibility of your gen AI 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.

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.

Scale

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

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

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.

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.

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.

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.

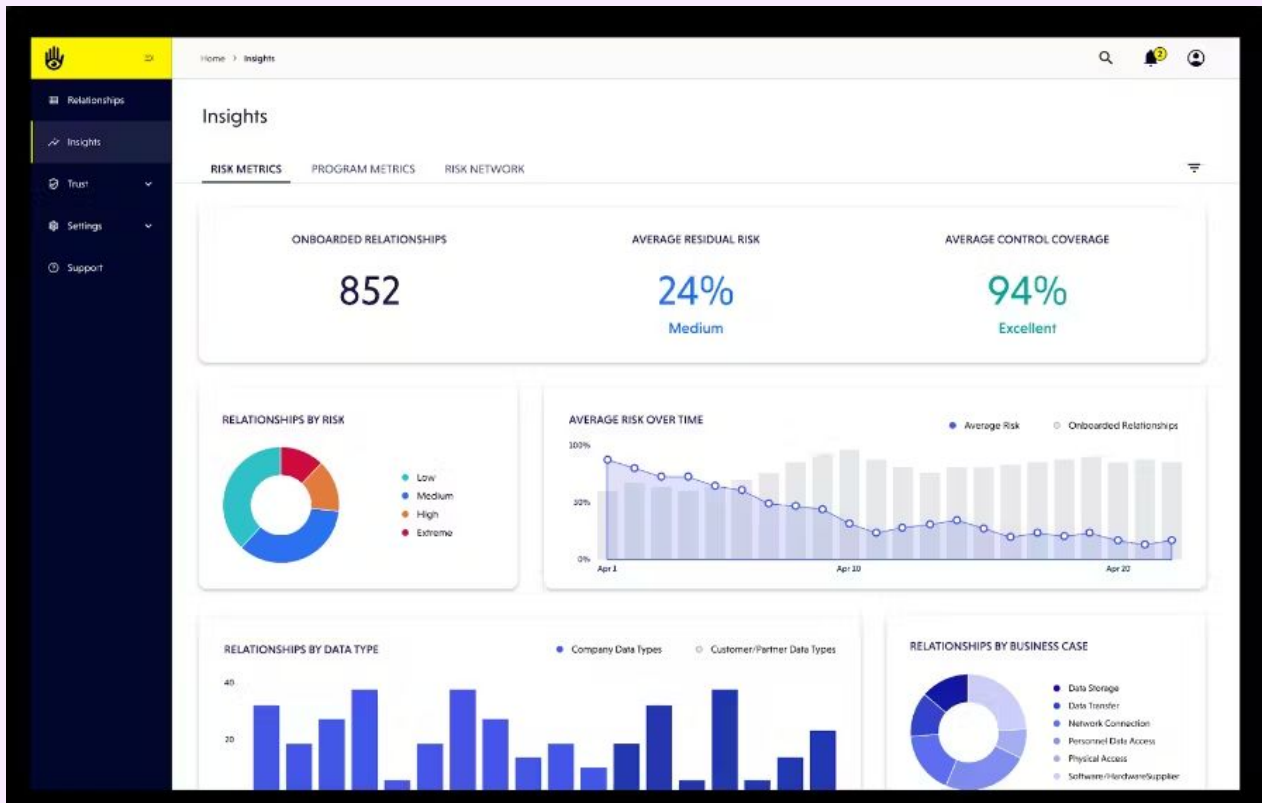


Figure 66: Insights dashboard

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**.

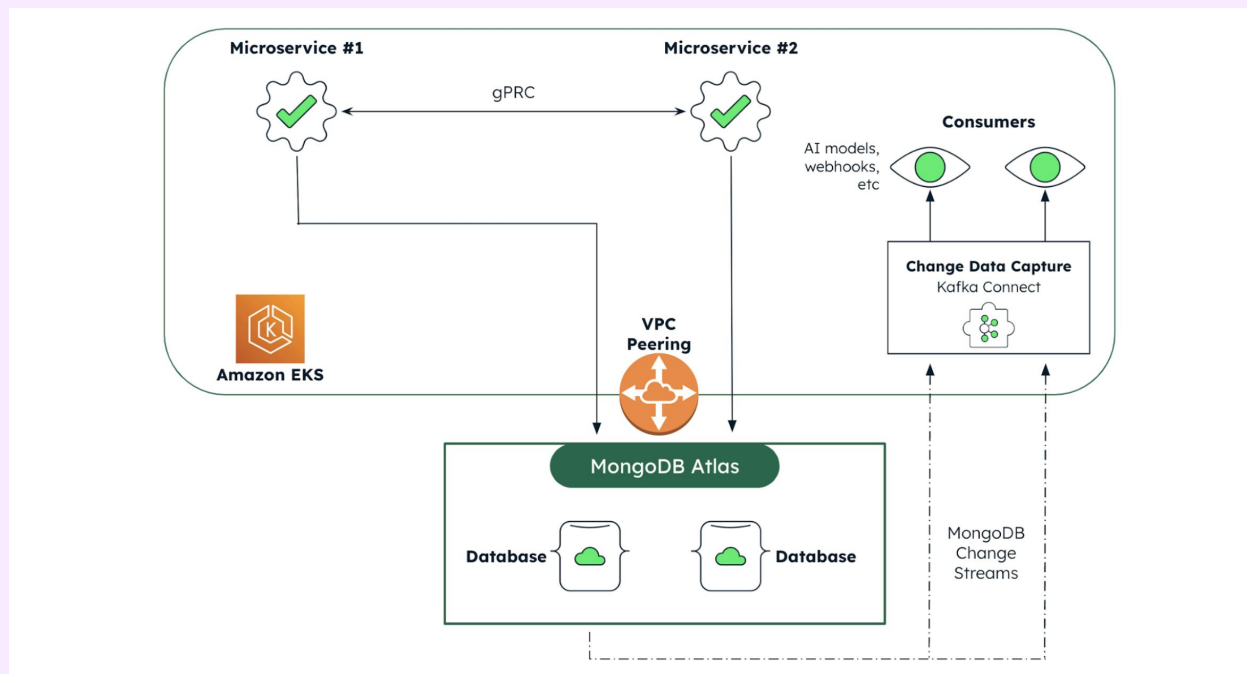


Figure 67: Event-driven microservices architecture for DevRev’s AI-powered CRM platform

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, 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 **real-time**.
- Utilizing Cloneable and Atlas Vector Search to generate vector embeddings from images and device data enables users to **efficiently search and analyze** field-collected events, thereby **enhancing decision-making and insights**.

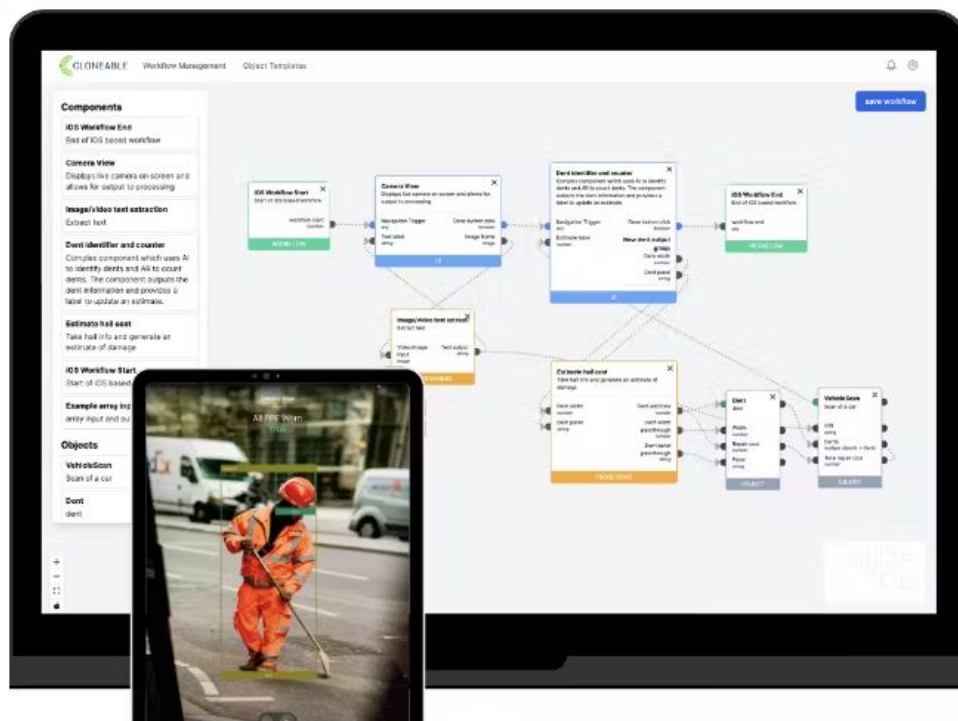


Figure 68: Cloneable components

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 Gen AI



[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.

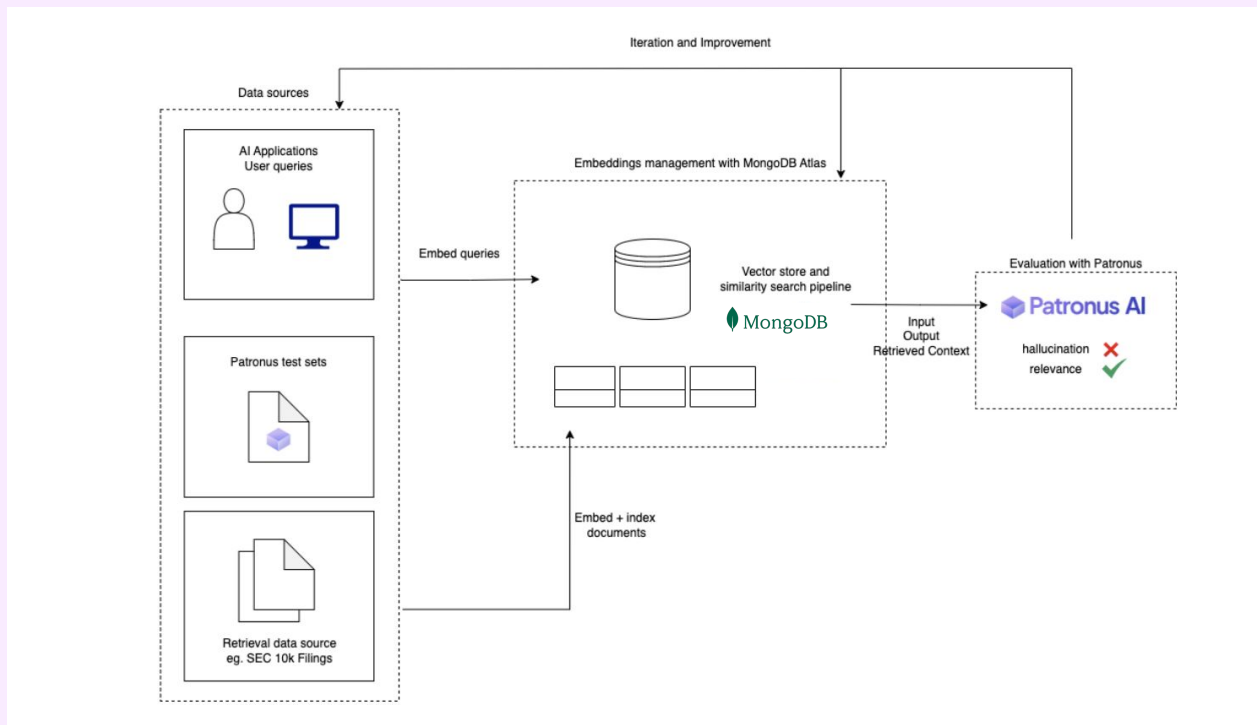


Figure 69: Reference architecture and workflow

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.

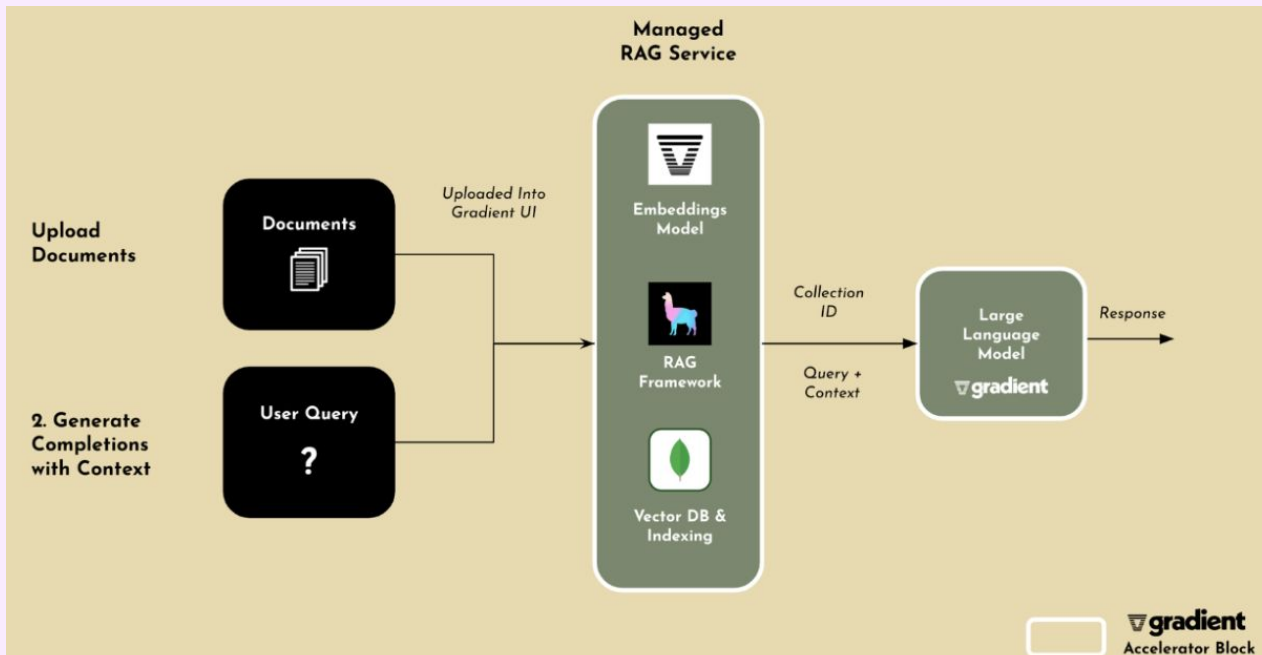


Figure 70: Managed RAG service

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.

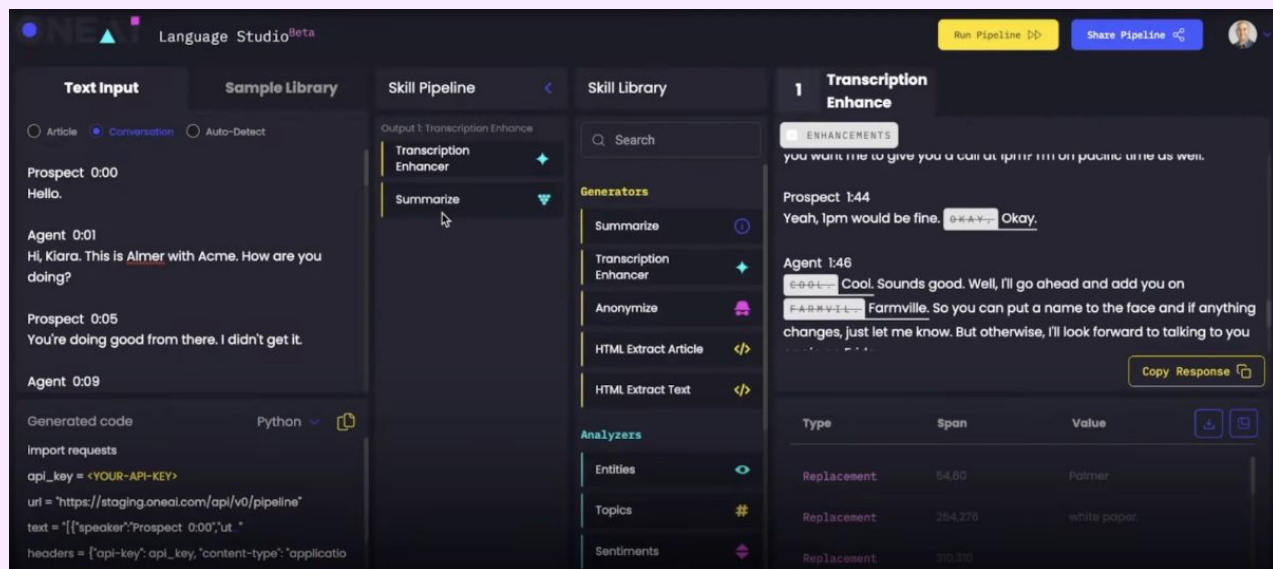


Figure 71: The One AI Language Studio

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 at 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.”

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**.

High level architecture

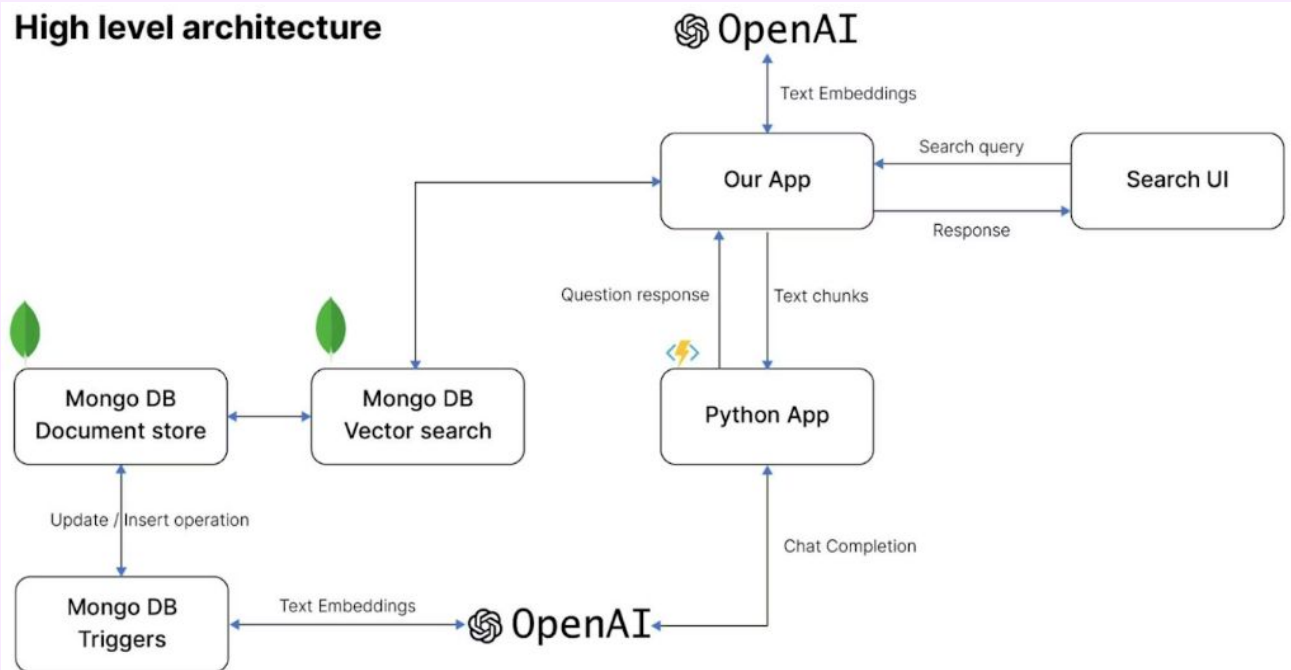


Figure 72: Reference architecture

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.

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

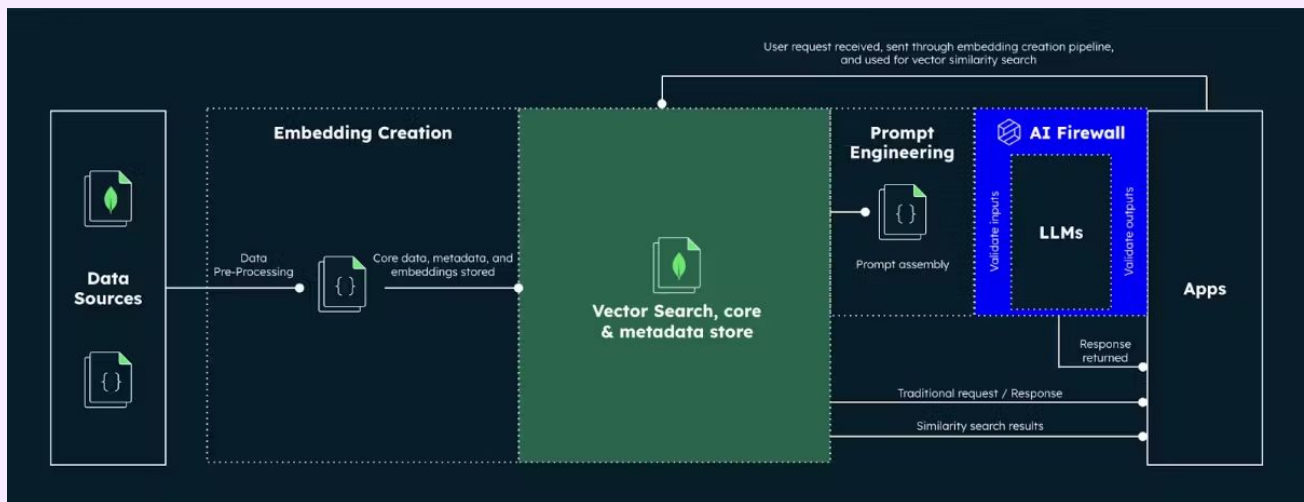


Figure 73: High level architecture

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.

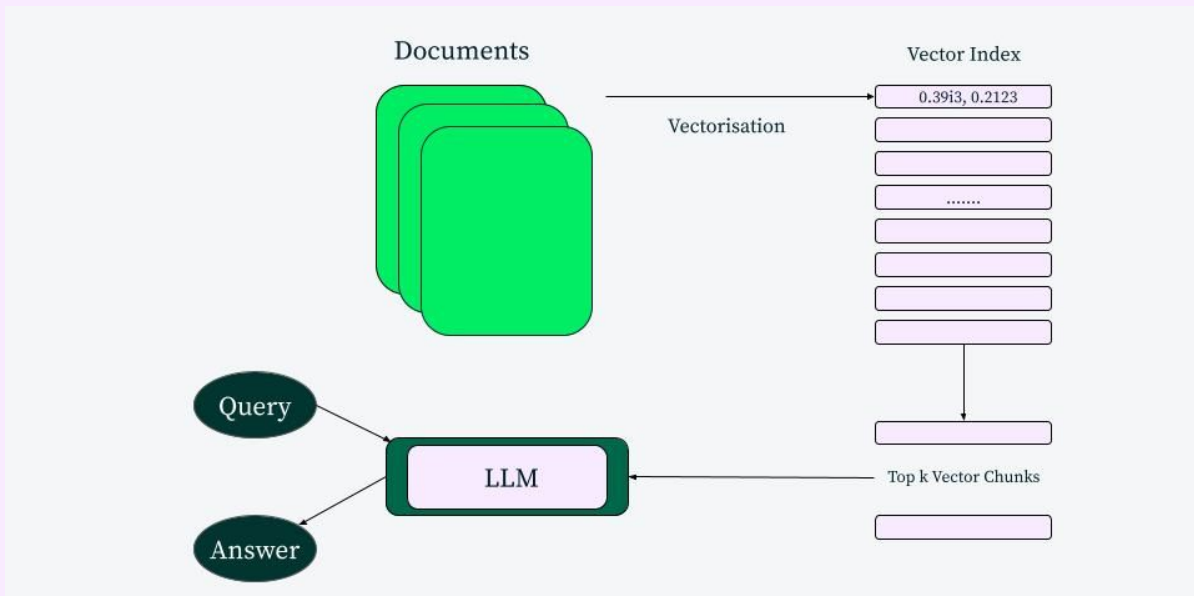


Figure 74: Bringing your data to LLMs

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 140 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.

LangChain: build, test, and monitor enterprise-ready LLM applications and agents



[LangChain](#) and [LangGraph](#) are open-source frameworks for building context-aware reasoning applications and reliable agents. [LangSmith](#) is a unified developer platform for building, testing, and monitoring LLM applications, whether you use LangChain or not.

Join 1M+ builders who standardize their development using LangChain's frameworks and platform together or separately to accelerate their AI application development.

Flexibility

Build context-aware, reasoning applications with LangChain's flexible framework that leverages your company's data and APIs

Develop, Debug, and Monitor

With [LangSmith](#), see what your agent or chain is doing and what context is being provided to the prompts. See what's happening in production.

Reliability

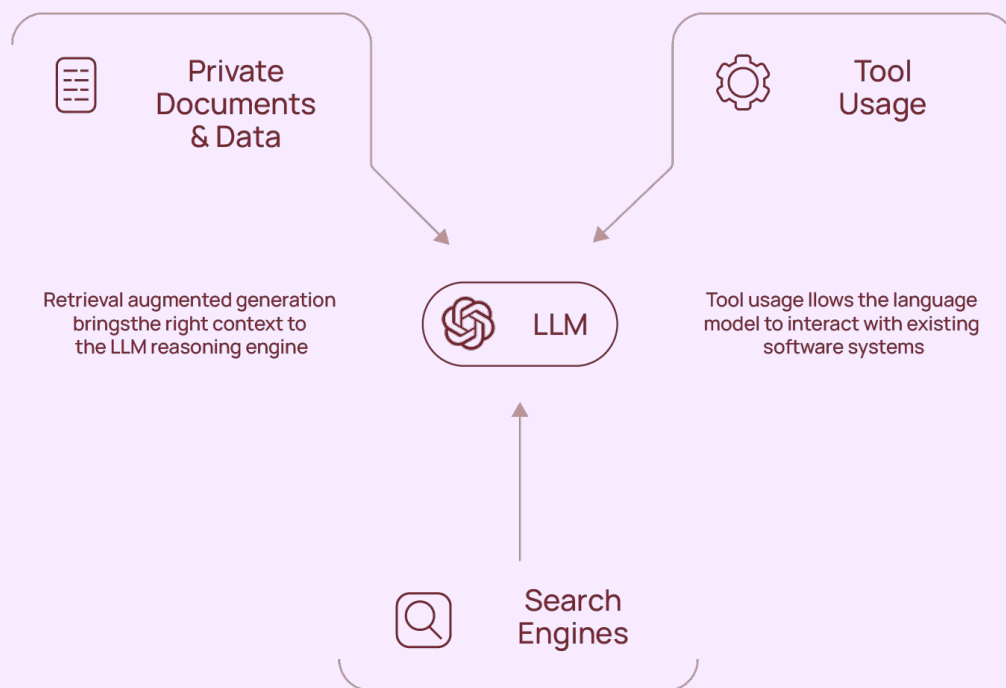
Use [LangGraph](#) as your controllable agent orchestration framework for handling complex tasks reliably.

Evaluate and Test

Layer in human feedback on [LangSmith](#) runs or use automatic evaluation. Stress-test quality over large datasets.

AI Efficiency, Simplified: LangChain and MongoDB

- **Boosting Retrieval Performance:** With MongoDB's efficient data storage and LangChain's flexible frameworks, you can increase operational productivity, improve discovery and personalization, and enhance scalability and performance of your LLM applications.
- **Enhanced Search:** The LangChain integration makes it easy to vectorize your operational data from and to MongoDB Atlas, allowing for more efficient and accurate retrieval of documents, even when the queries are complex or ambiguous, which can significantly improve the performance and accuracy of search applications.
- **Creating Tailored Solutions:** MongoDB's vector database and LangChain's application framework make it possible to build and deploy solutions that leverage your proprietary data to address many use cases.



Augment the power of LLMs with your data

The most powerful LLM applications and agents integrate with public data (via LLMs and search engines), external systems (via tools), and most importantly, your organization's private documents and data.

MongoDB's deep integration with LangChain's [Python](#) and [TypeScript](#) frameworks provides the powerful semantic search capabilities of MongoDB Atlas's vector search engine over your private data.

Start with a retrieval-augmented generation workflow just over your private data, expand your functionality with real-time search engine powered features, and finally give LangGraph agents access to your APIs to interface with other systems.

Systems composing LangChain, MongoDB, and an LLM together enable you to extract valuable insights from complex and recent data, highlighting the potential of these technologies in developing intelligent applications. For more information, please refer to this [detailed blog](#).

Build generative AI applications over your enterprise data with LlamaIndex and MongoDB



[LlamaIndex](#) is a developer platform that accelerates building agentic applications over your enterprise data. It has two main products: a development framework in Python/Typescript that is one of the most popular orchestration tools for building agentic applications. Another is LlamaCloud, which provides an end-to-end RAG platform freeing up dev teams to focus on business facing logic. LlamaCloud includes LlamaParse, an industry-leading document parsing solution that handles complex documents, formatting them to be easily understood by LLMs.

Accelerated time to production

LlamaIndex enables 5-10x faster development, from start to production deployment, reducing development costs (e.g. 2 engineers delivering production apps in a few weeks), and enabling rapid iteration on multiple use cases.

Scalable and Enterprise ready

LlamaIndex can scale to handle large volumes of enterprise data. It provides enterprise grade security, including in VPC deployments

Higher performance and accuracy

Customers see a significant boost in retrieval performance, as much as 2x over naive retrieval over complex documents (e.g. Docs with Tables, Charts, Figures)

LlamaIndex and MongoDB Atlas - better together

- **Seamless Integration:** [MongoDB Atlas](#) as a Data Sink (VectorDB) is supported in LlamaCloud and LlamaIndex framework. This works for SaaS as well as in VPC deployments and takes just a few clicks.
- **High performance results:** Atlas Vector Search Index is automatically created and vector search queries can be run via the advance retrieval algorithms provided by LlamaIndex!
- **Scalable Applications:** Customers can build AI-driven applications, such as agents or assistants, with scalability and flexibility, benefiting from MongoDB's robust data management and LlamaIndex's enterprise ready offerings

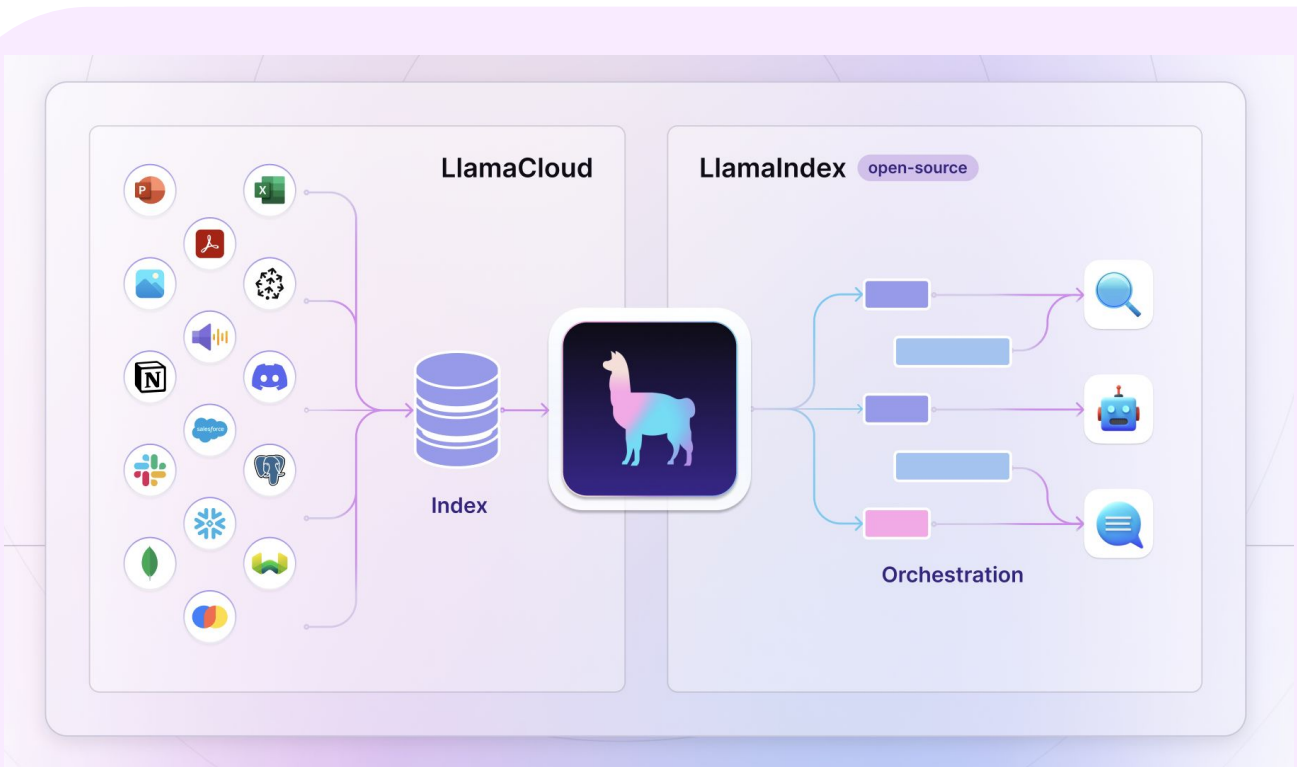


Figure 75: LlamaCloud indexes data from various sources and enables advanced retrieval algorithms. This can be connected to the open-source LlamaIndex for agentic workflows, Q&A, chatbot etc

LlamaIndex provides the fastest way to build production-ready AI applications over your company's data.

LlamaCloud

- Enterprise RAG platform: Connect with your data sources and provide advanced multimodal RAG capabilities
- Rapid Deployment and Performance: Set up data pipelines for LLMs in a few clicks, while minimizing hallucination and ensuring high retrieval performance
- Security and Scalability: Enterprise grade security and scalable infrastructure for all deployment sizes

LlamaIndex (open source framework)

- Comprehensive AI Development Framework to build Advanced AI Applications
- Seamless Integrations with data sources, LLMs, and vectorDBs along with customizability
- Strong community and ecosystem

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.

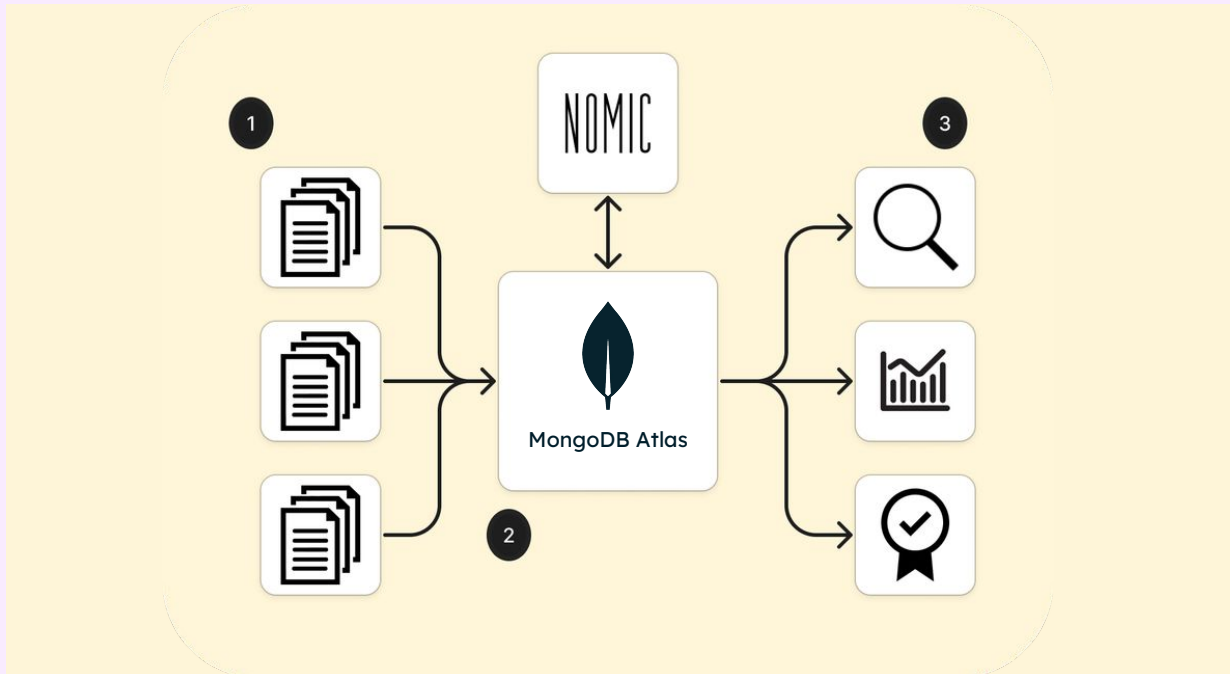


Figure 76: Use-case: PDF Search

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: Building Super-Fast GenAI Apps with Inference, Embeddings, and MongoDB Atlas



Founded in San Francisco in 2022, [Together AI](#) is the AI Acceleration Cloud for building and running generative AI (gen AI). Over 150,000 developers, and organizations like Zomato, The Washington Post and DuckDuckGo run gen AI in production using Together's Platform. The company has raised over \$200 million, counting NVIDIA, Salesforce Ventures, Kleiner Perkins, Lux, and NEA as investors.

Together Inference

Run super-fast inference at production scale for models like Llama-3 or your own custom models, with up to 4x faster performance than vLLM and 2x faster than hyperscalers. Deploy effortlessly with our developer-friendly APIs and scale easily.

Together Embeddings and Rerank APIs

Build Retrieval-Augmented Generation (RAG) applications with ease, with access to leading open-source embeddings and rerank models through Together's easy-to-use APIs, at up to 12x lower cost than proprietary solutions.

"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

By integrating [MongoDB Atlas](#) with [Together AI](#)'s inference and embedding capabilities, developers can build AI applications that deliver fast, real-time insights. This joint solution enables AI models to leverage RAG to recommend accurate, data-driven results that meet user criteria while maintaining accuracy.

- **Reduced Complexity and Cost** MongoDB Atlas and Together AI simplify RAG app development by making it easy to keep embeddings up to date. Together AI also provides high-performance inference at significantly lower costs than closed-source solutions.
- **Faster Time-to-Market** By integrating MongoDB Atlas with Together APIs, developers achieve up to 4x faster inference for models like Llama-3, accelerating development and speeding time-to-market.

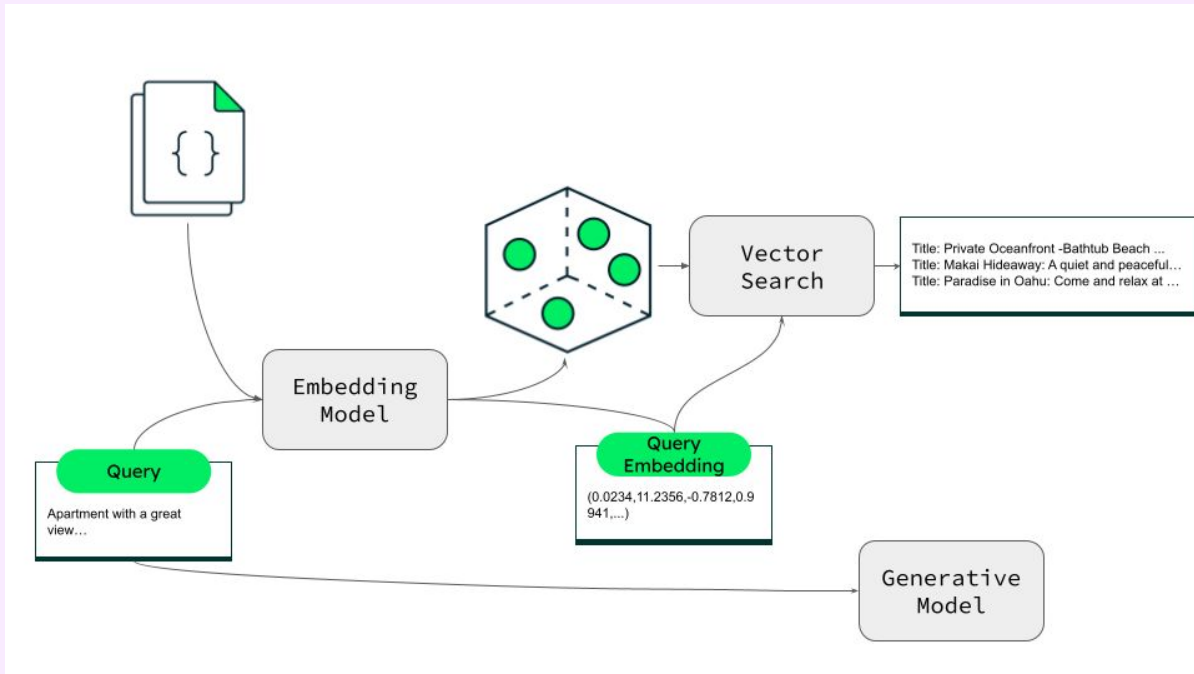


Figure 77: Together AI framework

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.

Own your GenAI with Arcee AI



[Arcee AI](#) meets you where you are on your AI journey, always giving you full ownership of your models and data. We provide companies with cutting-edge out-of-the-box models, state-of-the-art custom models, and/or our easy-to-use model training platform. You get the freedom to deploy the models to any environment—including SaaS, dedicated SaaS, on-premise, or VPC—with guaranteed data privacy and security. Our solutions cater to various use cases from customer service to software development, offering flexible pricing options and dedicated support. Our featured products include:

SuperNova & Other Top-Tier Models

High-performance general use models out-of-the-box, like [SuperNova](#), our distilled version of Llama-405B that outperforms leading models in various benchmarks.

Advanced Model Refinement

Our best-in-class post-training pipeline, incorporating synthetic dataset generation, SFT, reward modeling, sparse auto-encoding, model merging/infusion, and DPO to produce state-of-the-art models.

Optimized AI models for edge devices

Small custom models including device-optimized options like Arcee Ember (1.5B) and Arcee Pulse (3.8B) for edge devices.

Swarm: A Domain-Specific Model Network

A network of highly-specialized “expert” models (Swarm) that outperform generalist LLMs in specific domains.

Unleashing Arcee AI with MongoDB Atlas

- **Simple, Straightforward Deploymentless:** [MongoDB Atlas](#) users can access SuperNova through the AWS Marketplace by subscribing to the model. A SageMaker CloudFormation script then configures and deploys SuperNova to a SageMaker endpoint.
- **Superpowered Search:** [SuperNova](#) provides advanced insights and can answer complex, multi-step questions. This empowers Atlas users with deeper understanding of their data and makes it easier and faster for them to surface relevant information.
- **Cost Conscious:** With many customers working in a usage-based environment, Arcee AI offers both the 70B parameter SuperNova and the 8B parameter SuperNova Lite which is also extremely powerful but has the cost-efficiency advantages of being very small.

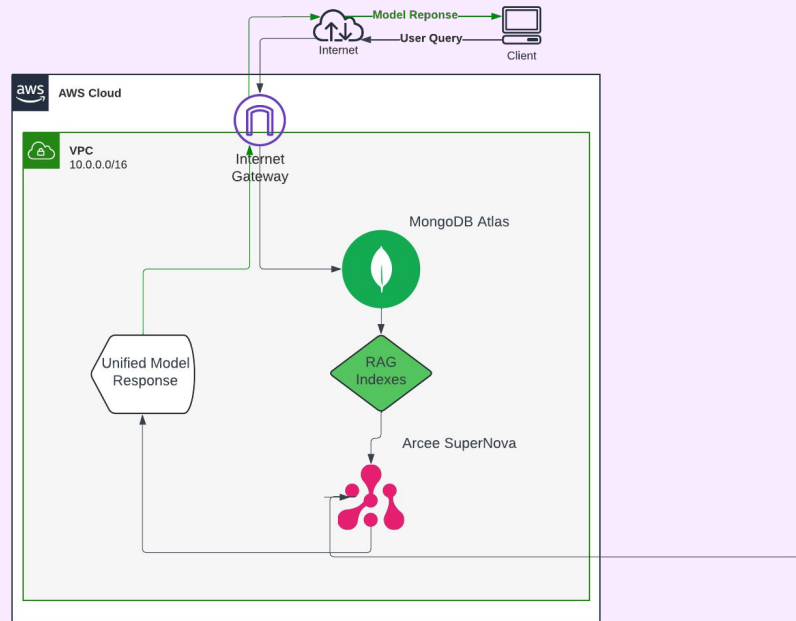


Figure 78: MongoDB Atlas and Arcee AI integrated in AWS for real-time, AI-driven responses using RAG.

The integration of MongoDB Atlas and Arcee AI, as depicted in the diagram, enables efficient data retrieval and AI-driven responses through a seamless pipeline. MongoDB's scalable cloud infrastructure ensures fast access to data using RAG indexes, allowing Arcee AI to augment the retrieval with powerful generative capabilities. This combination enhances the speed and accuracy of responses, making it ideal for real-time applications like interactive AI systems. MongoDB's reliability and Arcee AI's advanced inference capabilities complement each other, resulting in a unified model response. Overall, this architecture enables dynamic, scalable, and intelligent query handling in cloud environments.

- **Scalability:** MongoDB Atlas ensures flexible, cloud-native scaling to handle growing data needs without compromising performance.
- **AI-Augmented Retrieval:** Arcee AI enhances retrieved data with advanced generative processing for more accurate and meaningful responses.
- **Real-Time Performance:** The combined architecture delivers fast, reliable query handling suitable for interactive and time-sensitive applications.

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.

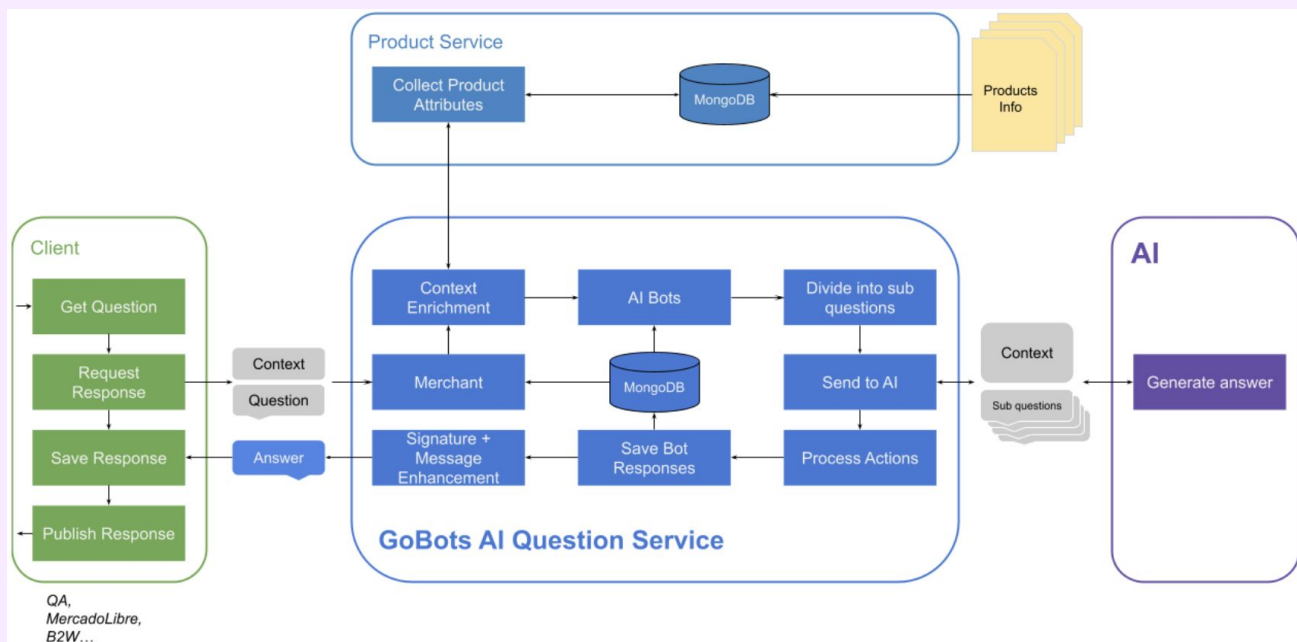


Figure 79: GoBots question processing architecture

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

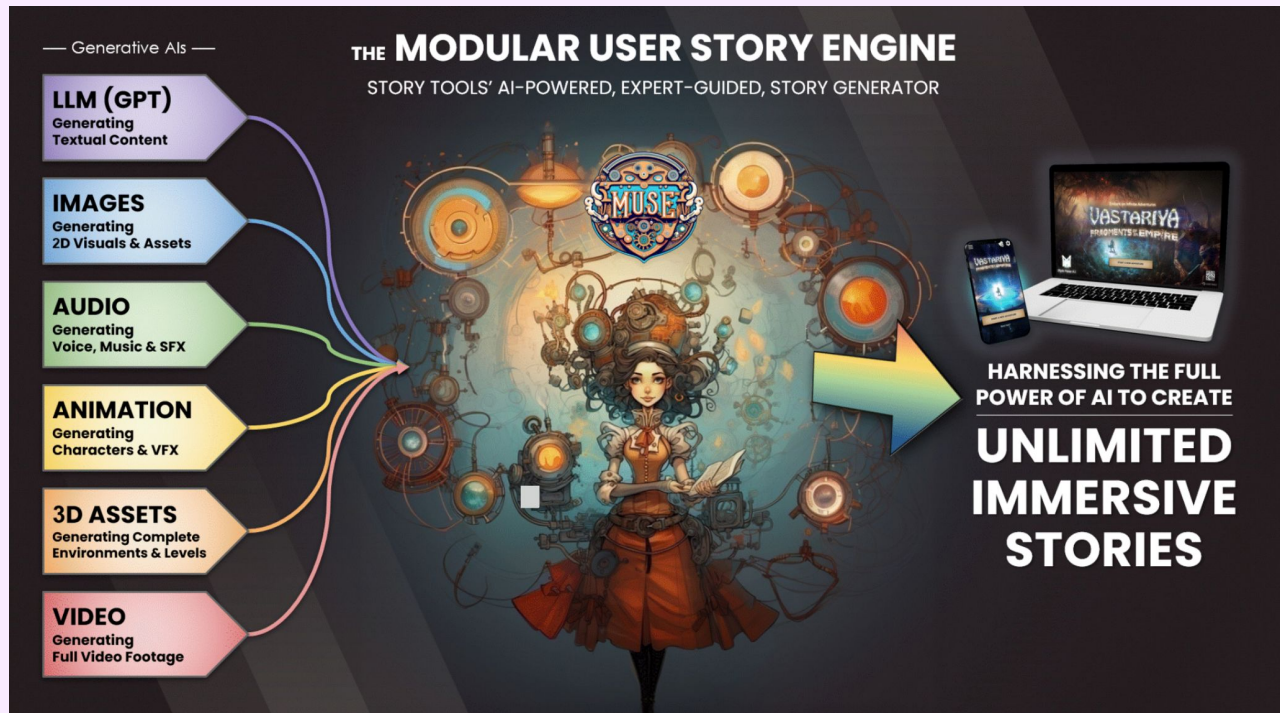


Figure 80: Story Tools modular user experience engine

MUSE orchestrates multimodel gen AI 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 at 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 completion 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](#).

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



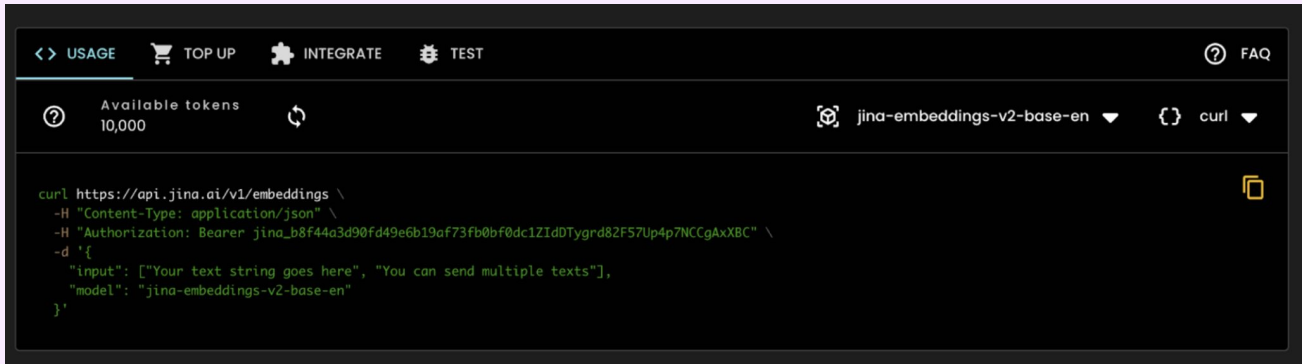
Jina AI has swiftly risen as a leader in multimodal AI, focusing on prompt engineering and embedding models. With its commitment to open-source and open research, Jina AI is bridging the gap between advanced AI theory and the real world AI-powered applications being built by developers and data scientists.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

The image shows a dark-themed web interface for the Jina AI API. At the top, there are navigation links: <> USAGE, TOP UP, INTEGRATE, and TEST. On the right, there is a FAQ link. Below the navigation bar, the status 'Available tokens 10,000' is shown on the left, and the selected model 'jina-embeddings-v2-base-en' and the tool 'curl' are on the right. The main area contains a curl command for the API endpoint https://api.jina.ai/v1/embeddings. The command includes headers for Content-Type (application/json) and Authorization (Bearer token), and a JSON body with an 'input' array containing two sample text strings and a 'model' field set to 'jina-embeddings-v2-base-en'.

```
curl https://api.jina.ai/v1/embeddings \
-H "Content-Type: application/json" \
-H "Authorization: Bearer jina_b8f44a3d90fd49e6b19af73fb0bf0dc12Id0Tygrd82F57Up4p7NCCgAxx8C" \
-d '{
  "input": ["Your text string goes here", "You can send multiple texts"],
  "model": "jina-embeddings-v2-base-en"
}'
```

Figure 81: 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.

Superduper: Build Next-Gen AI Apps on Your Existing Database



[Superduper.io](#), provides an open-source framework for implementing AI applications and workflows with their existing databases. By transforming the database into a central AI platform and consolidating all critical steps and components of AI applications within a single environment, AI teams can implement custom AI solutions without the need for data migration via pipelines or complex MLOps infrastructure. From GenAI, vector search, and RAG to classic machine learning.

Their enterprise platform is built for efficient and secure scalability, self-hosted on existing data infrastructure (whether in the cloud or on-prem), so that data does not need to be shared externally.

Superduper offers different AI app and workflow templates **ready-to-install on MongoDB** and highly configurable with minimal development effort. Because the application templates are based on their [open-source development framework](#), you can fully own the codebases and control any level of implementation detail. Current use-cases include Multi-Modal Vector Search & RAG, Document Extraction & Analysis, Anomaly Detection, Visual Object Detection, and Image and Video Search.

Find an executive summary [here](#).

“We integrate MongoDB as a core backend database for our platform, using the PyMongo driver for app connectivity and Atlas Vector Search for vector embedding storage and querying. Having MongoDB Ventures on board enables us to work even more closely with their teams, optimizing our product while engaging with MongoDB’s vast developer community.”

Duncan Blythe
Co-Founder & CTO of Superduper.io

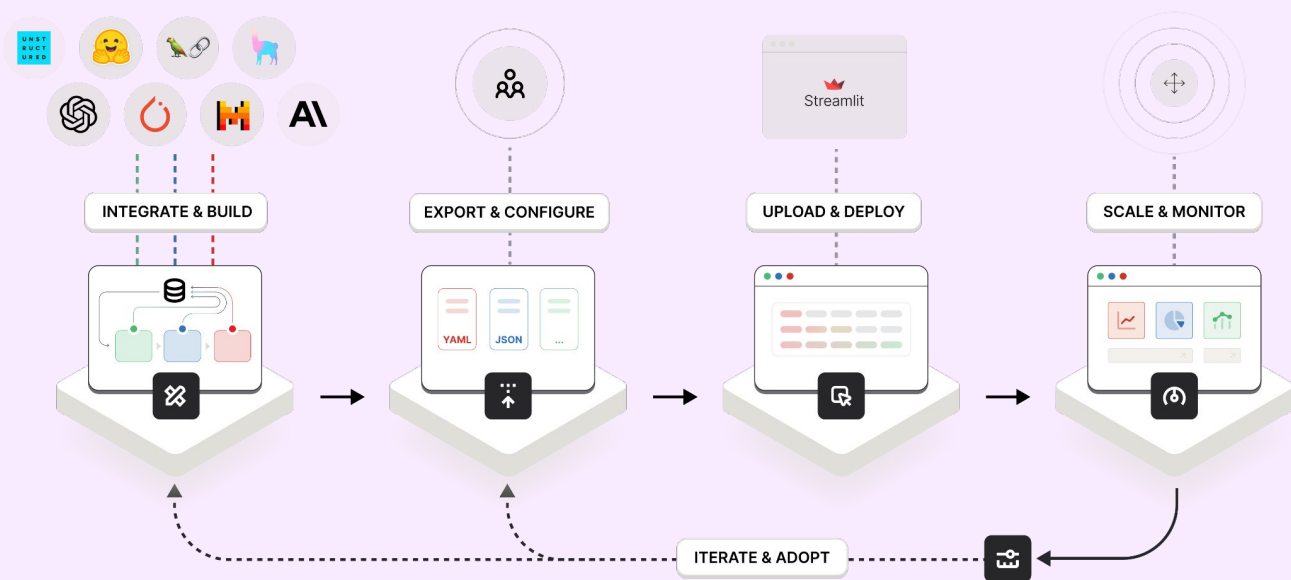


Figure 82: AI application lifecycle with Superduper.io

Development Features

- Integrations with major AI frameworks, model hubs, and APIs and compatible with any Python functionality
- Ability to switch between and combine different tools and providers, even in the same workflow
- Operating and navigating data simply with database queries (instead of building pipelines)
- A simple Python interface with reusable building blocks
- Application templates for different AI use-cases, ready-to-install on the database

Deployment Features

- Compute running where data resides, minimizing data movement on existing infrastructure, whether cloud or on-prem
- Scalable model hosting, low latency streaming inference and model training/fine-tuning
- Database transformation into an AI platform, including feature store and model hub
- Granular management and observability user interface
- Relying on existing security and rights management controls of the datastore

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

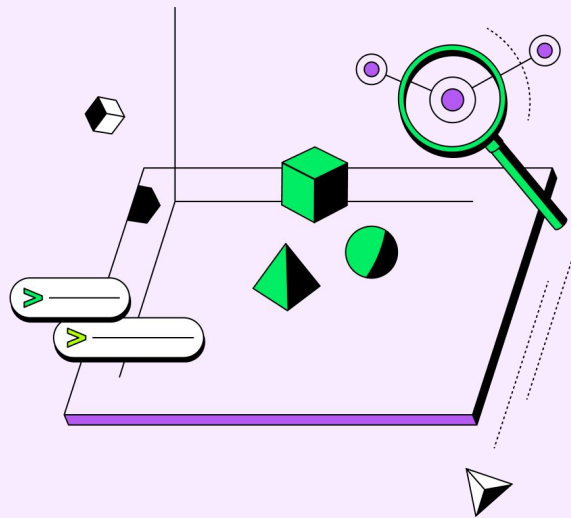
The AI agent participates in various tasks such as joining slack threads, meetings, transcribing calls, generating summaries, responding to emails, and updating issue trackers.

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.



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.



Figure 83: 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

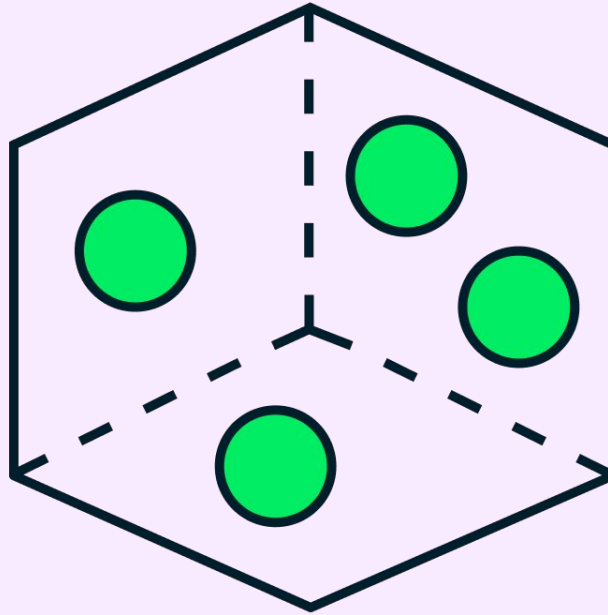
Crewmate's platform features context-aware search powered by Atlas Vector Search, delivering relevant content to users on brand community pages.

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.



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."

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.

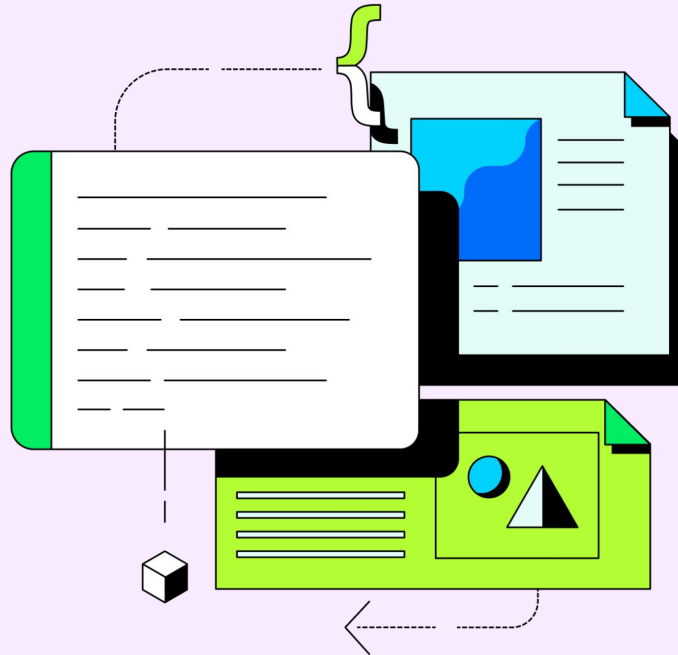
AI Models and Technologies

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



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.”

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

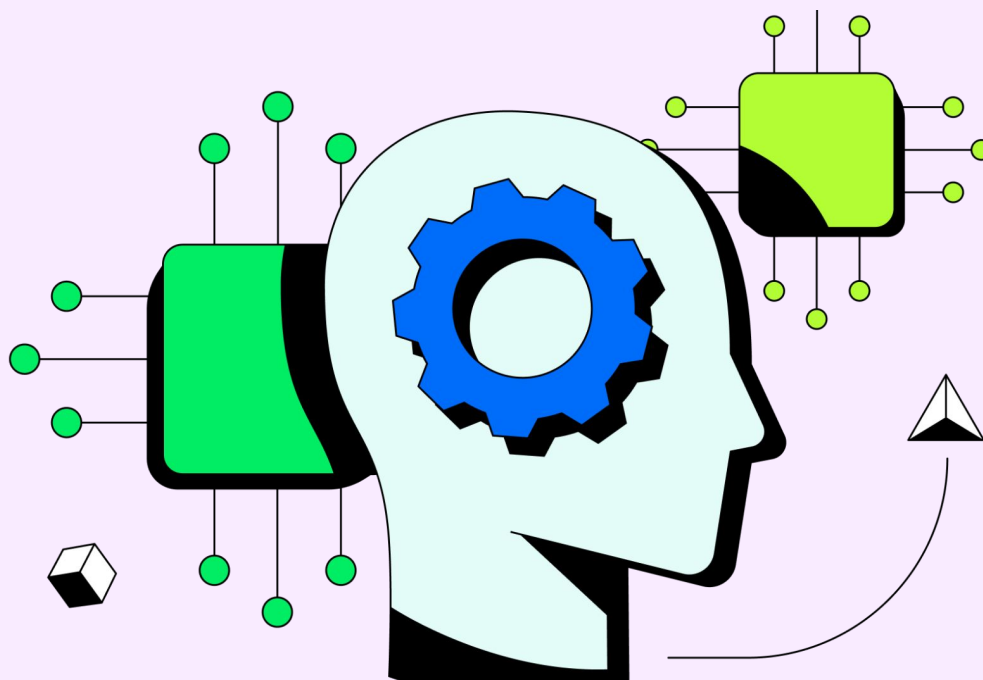
Artificial Nerds offers voicebots that automate phone calls for businesses. These voicebots understand user language and provide solutions, making business interactions more efficient and effective.

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.



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.

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.

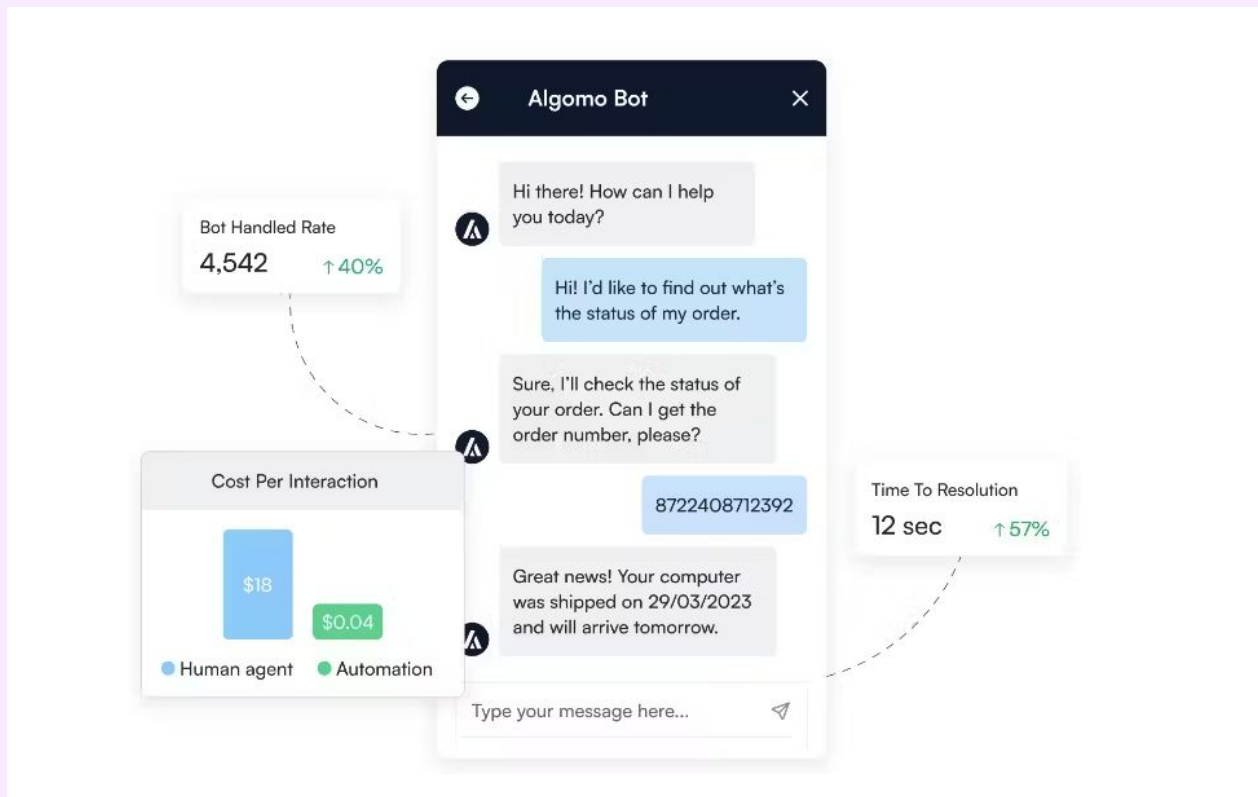


Figure 84: Algomo Bot

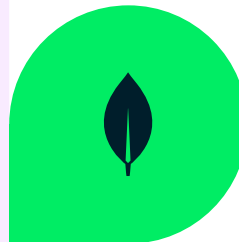
Unlocking Personalized Customer Experiences with Algomo's Conversational AI

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 at industry.solutions@mongodb.com

Innovation workshops

Learn about our 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.

